Estudio de parámetros ambientales utilizando técnicas espectroscópicas, datos meteorológicos y métodos estadísticos

TESIS DE DOCTORADO

Gustavo Enrique Ratto

Presentada ante la *Facultad de Ingeniería*
de la *Universidad Nacional de La Plata* (*UNLP*)- Argentina
para optar por el título de

Doctor en Ingeniería

Lugar donde se llevó a cabo el trabajo de tesis:

Centro de Investigaciones Opticas (CIOp), La Plata, Provincia de Buenos Aires, Argentina, dependiente de la *Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC BA)* y del *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)* de la Argentina.

Director de Tesis: - Dr. Daniel Carlos Schinca

Co- director de Tesis: - Dr. Jorge Reyna Almandos

Jurados de Tesis: - Lic. Laura E. Dawidowski
- Dr. Salvador Enrique Puliafito
- Dra. Beatriz Margarita Toselli

Fecha de la defensa oral y pública: 15 Junio de 2016

Agradecimientos

A los Dres. Christian Weber y Gustavo Torchia, a la Lic. Nelly Cap y en particular al Dr. Fabián Videla del Centro de Investigaciones Opticas (CIOp) por su colaboración en las distintas gestiones involucradas en el trabajo de tesis.

Un agradecimiento especial al Prof. Dr. Mario Gallardo (ex- director del CIOp) por su apoyo para que esta tesis fuera posible.

También al personal de la Universidad Tecnológica Nacional (Facultad Regional La Plata) que colaboró con el aporte de datos meteorológicos y de contaminantes. En particular, a los ingenieros Victor Sacchetto, Juan Carlos Ragaini y de manera muy especial al ingeniero Mario Rosato por su permanente predisposición.

Mi gran reconocimiento a los Dres. Guillermo Berri y Ricardo Maronna por la predisposición docente de ambos, por sus colaboraciones en el desarrollo de las publicaciones y en la revisión del presente escrito.

Al Prof. Ing. R. Pessacq y al Departamento de Ingeniería Química, en particular a la secretaria del departamento, Sra. Eva por su permanente predisposición y apoyo.

Y finalmente, un profundo agradecimiento a mi co-director de tesis con quien compartimos esta “aventura interdisciplinaria”, ¡muchas gracias Jorge!

G.E.R.

Resumen

(700 palabras)

La ciudad de La Plata y alrededores (ubicada en el Estuario del Río de La Plata en Sudamérica) es una de las seis urbes más pobladas de la Argentina (alrededor de 800 000 habitantes) y, como tal, posee una importante actividad económica (industrial, administrativa y de tránsito vehicular). Como la mayoría de las grandes ciudades, debe poner en consideración la incidencia de enfermedades respiratorias causadas por la contaminación local del aire al mismo tiempo que debe afrontar los desafíos del cambio climático global.

Dado que la ciudad se halla en una zona con baja capacidad de depuración atmosférica, que no posee una red oficial para el seguimiento de los contaminantes del aire y que estudios previos han indicado niveles altos de algunos contaminantes (material particulado, compuestos orgánicos volátiles e hidrocarburos aromáticos policíclicos entre otros), fue posible formular un conjunto de objetivos de estudio que permitan tanto enriquecer el conocimiento del ambiente en la zona como sugerir estrategias para la mejora. Tales objetivos forman parte de la presente tesis, que posee un fuerte carácter multidisciplinario. Los mismos pueden definirse en términos de compilación de información ambiental (actualmente escasa y difusa), capacitación en el manejo de equipamiento de monitoreo, medición de especies contaminantes, estudio de patrones de vientos y sus escalas y el análisis de la dinámica del viento como agente de transporte de los contaminantes (principalmente industriales).

El estudio de datos ambientales (principalmente dióxido de azufre y vientos) se llevó a cabo utilizando métodos estadísticos de análisis univariado y multivariado, tanto desde la perspectiva inferencial como desde la perspectiva exploratoria. Los métodos de análisis por conglomerados jerárquicos, escalamiento multidimensional, componentes principales, correlación y regresión (entre otros) se presentan y discuten en términos de las aplicaciones a temas ambientales. Estos métodos constituyeron la principal herramienta para formular conclusiones acerca de los fenómenos físicos involucrados. Se discute la importancia de los conceptos de similitud y disimilitud y el "arte" de encontrar grupos como parte del reconocimiento de patrones. El concepto de robustez estadística es un tema transversal a todo el tratamiento de datos y a la aplicación de métodos de modelado.

La presencia de dióxido de azufre (de origen preponderantemente industrial) resultó ser significativa como para sugerir el seguimiento continuo de este gas. El estudio de los vientos de la zona permitió hallar concordancia con fenómenos de mesoescala. A escala local se pudieron distinguir y caracterizar dos grupos de direcciones de viento muy importantes: el Sector 1 (NNO-N-NNE-NE) que transporta los contaminantes desde el área industrial al casco urbano es dominante al mediodía y la tarde temprana y el Sector 2 (ENE-E-ESE) que transporta a los contaminantes hacia zonas residenciales es dominante en horas del anochecer. Entre ambos suman una ocurrencia diaria promedio superior al 50%. La observación de estos sectores desde distintos puntos de monitoreo mostró que se hallan fuertemente correlacionados mostrando, un patrón generalizado en toda el área de estudio. Por otra parte, la ocurrencia de calmas promedio es al menos 11.6% independientemente de la estación del año y las velocidades de los vientos son lo suficientemente bajas (< 10 km h^{-1}) la mayor parte del tiempo como para considerarse una causa facilitadora de la acumulación de contaminantes del aire.

Los patrones de direcciones de viento representados por 24 rosetas horarias tienen un comportamiento estacional y pueden ser descriptos por un número reducido de representantes (5 a 8) que dan cuenta de la dinámica de los vientos en el ciclo diario de la capa límite planetaria. Tanto el análisis por conglomerados y el de escalamiento multidimensional como los métodos para cuantificar disimilitud entre patrones permitieron

detectar la presencia del fenómeno de brisa de mar-tierra pudiéndose observar fluctuaciones entre sitios de observación (uno lejano y otro cercano a la costa) y entre estaciones del año (predominio del verano).
Fue posible establecer las ventajas estratégicas de los sitios de observación y señalar otros puntos potenciales que servirían para el seguimiento de las concentraciones de fondo y la mejora en la detección de contaminantes de origen industrial y vehicular. Como resultado del estudio surge, sin lugar a dudas, la gran necesidad que tiene la ciudad y sus alrededores de realizar el seguimiento continuo tanto de parámetros meteorológicos como de los principales contaminantes del aire.

Palabras clave

(25 entradas)

Análisis exploratorio, análisis multivariado, análisis por conglomerados jerárquicos, brisa marina, calmas, componentes principales, contaminación del aire, correlación, Curvas de Andrews, DOAS, escalamiento multidimensional, estructura de datos, La Plata, meteorología, métodos robustos, patrones estacionales, red de monitoreo, regresión, rosetas de viento, SO_2, tendencia, transporte de contaminantes, valores atípicos, vientos.

Study of environmental parameters employing espectroscopic techniques, meteorological data and statistical methods

Abstract

La Plata City and surroundings (located on the estuary of De la Plata River in South America) is one of the six cities most populated of Argentina (around 800 000 inhabitants) and, as such, has important economic activity (industrial, administrative and vehicular traffic). Like most large cities, it has to put into consideration the incidence of respiratory diseases caused by local air pollution at the same time it faces the challenges of global climate change.
Since the city is in an area with low air purification capacity, which does not have an official network for monitoring air pollutants and previous studies have indicated high levels of some pollutants (particulate matter, volatile organic compounds and aromatic polycyclic hydrocarbons among others), it was possible to formulate a set of learning objetives that allow both to enrich the knowledge of the environment in the area and to suggest strategies for improvement. These objectives are part of this thesis, which has a strong multidisciplinary character. They can be defined in terms of compilation of environmental information (currently limited and diffuse), training in the management of monitoring equipment, measuring pollutant species, studying wind patterns and their scales and, analyzing the dynamics of wind as an agent for transporting pollutants (mainly of industrial sources).
The study of environmental data (primarily sulfur dioxide and winds) was conducted using statistical methods of univariate and multivariate analysis, employing both the inferential and the exploratory perspective. The methods of hierarchical clustering, multidimensional scaling, principal components, correlation and regression analysis (among others) are presented and discussed in terms of applications to environmental issues. These methods were the main tool for formulating conclusions about the physical phenomena involved. The importance of the concepts of similarity and dissimilarity and of the "art" of finding groups in data as part of pattern recognition is discussed. The concept of statistical robustness is an issue that cuts across the entire data processing and application of modeling methods.
The presence of sulfur dioxide (from predominantly industrial origin) proved to be significant enough to suggest the need of the continuous monitoring of this gas. The study of the winds in the area allowed to find agreement with mesoscale phenomena. At the local scale it was possible to distinguish and characterize two groups of important wind directions: Sector 1 (NW-N-NNE-NE) which transports pollutants from the industrial to the urban area is dominant at midday and early evening and Sector 2 (ENE-E-ESE) that transports pollutants towards residential areas is dominant in evening hours. Both sectors add an average daily occurrence above 50%. The observation of these sectors from various monitoring points showed that they are strongly correlated pointing out a generalized pattern throughout the study area. The average occurrence of calm is at least 11.6%, regardless of the season, and wind speeds are low enough (<10 km h^{-1}) most of the time, both parameters indicate facilitating causes for the accumulation of air pollutants.
Wind direction patterns, represented by 24 hourly wind roses, have a seasonal behavior and can be described by a small number of representatives (5- 8) that reflect the dynamics of the winds in the daily cycle of the planetary boundary layer. Cluster analysis, multidimensional scaling as well as the methods for measuring dissimilarity between patterns allowed detecting the presence of the sea- land breeze phenomena and further observ fluctuations between monitoring sites (one far and the other near the coast) and between the seasons of the year (summer was prevalent).
It was possible to establish the strategic advantages of the observation sites and identify other potential points that would be qualified to monitor background concentration levels and in this way to improve the detection of contaminants from industrial and vehicular origin. As a result of the study arises, with no doubt, the great need for the city and its surroundings to perform continuous monitoring of both the meteorological parameters and the main air pollutants.

Publicaciones vinculadas a la tesis

Revistas internacionales con referato

Rosato, M., Reyna Almandos, J., Ratto, G., Flores, A., Sacchetto, V., Rosato, V.G., Ripoli, J., Alberino, J.C. y Ragaini, J.C. (2001) Mesure de SO_2 a La Plata, Argentine, *Pollution Atmosphérique,* 169: 85–98.

Ratto, G., Videla, F., Reyna Almandos, J., Maronna, R., Schinca, D. (2006) Study of meteorological aspects and urban concentration of SO_2 in atmospheric environment of La Plata, Argentina, *Environmental Monitoring and Assessment,* 121: 327- 342.

Ratto, G., Videla, F., Maronna, R. (2009) Analyzing SO_2 concentrations and wind directions during a short monitoring campaign at a site far from the industrial pole of La Plata, Argentina, *Environmental Monitoring and Assessment,* 149: 229- 240.

Ratto, G., Videla, F., Maronna, R., Flores, A., De Pablo, F. (2010a) Air pollutant transport analysis based on hourly winds in the city of La Plata and surroundings, Argentina, *Water Air and Soil Pollution,* 208: 243- 257.

Ratto, G., Maronna, R., Berri, G. (2010b) Analysis of wind roses using hierarchical cluster and multidimensional scaling analysis at La Plata, Argentina, *Boundary Layer Meteorology*, 137: 477- 492.

Ratto, G. y Nico, A. (2012a) Preliminary wind analysis regarding different speed ranges in the city of La Plata, Argentina, *Revista Brasileira de Meteorologia*, 27(3): 281 – 290.

Ratto, G., Maronna, R., Repossi, P., Videla, F., Nico, A., Reyna Almandos, J. (2012b) Analysis of Winds Affecting Air Pollutant Transport at La Plata, Argentina, *Atmospheric and Climate Sciences*, 2: 60-75.

Ratto, G., Berri, G., Maronna, R. (2014a) On the application of hierarchical cluster analysis for synthesizing low-level wind fields obtained with a mesoscale boundary layer model, *Meteorological Applications,* 21: 708–716.

Ratto, G., Videla, F., Reyna Almandos, J. (2014b) Analysis of the Homogeneity of Wind Roses' Groups Employing Andrews' Curves, *Atmospheric and Climate Sciences*, 4: 447-456.

Congresos internacionales con referato

Ratto, G., Videla, F., Maronna, R., Reyna Almandos, J. (2012c) Calm analysis using a robust method. *Primer Congreso Internacional de Ciencia y Tecnología Ambiental* Argentina y Ambiente 2012. Mar del Plata, 28 de Mayo- 1 de Junio de 2012, Argentina.

Otras publicaciones

Se presentaron numerosos trabajos de divulgación en forma de poster y actas de congresos entre los cuales se pueden citar:

AFA- Asociación Física Argentina

Ratto, G., Videla, F., Reyna Almandos, J., Schinca, D. (2005) *Análisis preeliminar de parámetros meteorológicos y prospección para el estudio de calidad de aire en la zona del Polo Petroquímico La Plata*, Actas de la 90^{va} Reunión AFA.

LINTA- Laboratorio del Territorio y Medio Ambiente
Videla, F., Schinca, D., Ratto, G., Ragaini, J.C. (2006) *Desarrollo de equipos ópticos para medir SO_2 en chimeneas y aire ambiente. Presentación de resultados de mediciones de SO_2 y parámetros meteorológicos utilizando equipamiento comercial en el área de La Plata,* Tecnologías e instrumentos para su evaluación integral, Sección: La calidad del ambiente urbano: Tecnologías e Instrumentos para su Evaluación Integral". CIC BA (Comisión de Investigaciones Científicas de la Pcia. de Bs. As.). Poster y Libro de Actas LINTA.

EOA- Encuentro de Optica Aplicada
Ratto, G., Videla, F., Schinca, D., Reyna Amandos, J. (2007) *Medidas ópticas de contaminantes y de parámetros meteorológicos para el estudio de calidad de aire,* EOA, Fac. de Ing., UBA (Universidad de Buenos Aires), Buenos Aires y CIOp (CIC-CONICET), Gonnet. Poster.

PROIMCA- Proyecto Integrador para la Mitigación de la Contaminación Atmosférica
Reyna Almandos, J., Videla, F., Schinca, D., Ratto, G., Ragaini, J.C., Sacchetto, V., Rosato, M., Arrieta, N., Bazán, J. (2007) *Métodos ópticos aplicados al monitoreo de contaminantes atmosféricos.* Poster y Libro de Actas PROIMCA (publicado en 2009).

Indice general

Capítulo III

Fenómenos físicos

Capítulo IV

Similitud- disimilitud, regresión y tendencia

Capítulo V

Análisis por conglomerados y escalamiento multidimensional

Capítulo VI

Síntesis y conclusiones finales

"We are changing the Earth more rapidly than we are understanding it"
In: Human domination of Earth's ecosystems
Revista Science (1997)

"The growth society is based upon excess, and it is leading us into a blind alley"
Serge Latouche
Farewell to growth (2009)

"The twenty-first century will be the age of sustainable development or the age of ruin. Worldwide economic growth over the past two centuries has brought remarkable progress but also remarkable risk"
Jeffrey D. Sachs
Director of the Earth Institute at Columbia University
Special Advisor to UN Secretary General Ban Ki-Moon on the Millennium Development Goals (2011).

Capítulo I
Introducción, organización y aportaciones de la tesis

I.1 Introducción
I.1.1 Generalidades
Contexto del monitoreo de la calidad del aire en Argentina y América Latina

La información sobre la contaminación del aire en los países en desarrollo o en economías de transición es limitada y las series en el tiempo son escasas (Fenger, 1999; CAI, 2012); muy pocos programas muestran la evolución de la contaminación de largo plazo y hay indicadores de que la situación se deteriora debido a que se prioriza el desarrollo económico frente a la protección ambiental (Fenger, 2009). A partir de mediados de la década de 1990, la región de América Latina y el Caribe (ALC) manifiesta disminución de la pobreza y aumento de la clase media (Ferreira et al., 2013), pero al mismo tiempo hay una fuerte tendencia a la motorización (mayor presencia de vehículos impulsados por combustibles fósiles) y al crecimiento descontrolado de las ciudades (UN-HABITAT, 2012) haciendo que la relación justicia ambiental- desarrollo sostenible tenga un carácter inmaduro (Carrizo y Berger, 2010; UNEP, 2014a).

La Argentina ha sido, históricamente, un país con escasa tradición en el monitoreo de los contaminantes del aire. Un reporte del Banco Mundial de 1995 manifiesta que "la contaminación ambiental en la Argentina es mayor de lo que se podría esperar en un país con su nivel de desarrollo..." y que "hay una falta de estudios periódicos sistematizados ….. el análisis y el monitoreo ambiental son casi nulos en el caso de la mayoría de los contaminantes en la mayor parte del país.....''; la contaminación del aire se menciona entre los principales problemas de los ecosistemas urbanos (Wais de Badgen, 1998). En un documento sobre la situación del aire en América Latina, Kork y Sáenz (1999) ubican a la Argentina como un país con "limitada capacidad de monitoreo", esto debe comprenderse no solamente en relación a la carencia de registros sino por la calidad de los mismos (Cifuentes et al., 2005). En "La Salud en las Américas" (WHO, 1998) al hablar de la situación del aire en la Argentina se dice que "muchas de las estaciones que integran la red del Sistema Mundial de Vigilancia del Medio Ambiente no cuentan con un sistema de monitoreo continuo, lo que impide efectuar un análisis específico". Gassmann y Mazzeo (2000) señalan que, en general, en toda la Argentina, hay pocos estudios observacionales de la calidad del aire. El informe Geo- Argentina "Perspectivas del Medio Ambiente de la Argentina" (PNUMA, 2004) explicita que "con respecto a la contaminación atmosférica debe consignarse que no se han identificado fuentes de información que den cuenta de registros sistemáticos de la calidad del aire que permitan formular una caracterización general del estado del recurso a escala nacional". En una revisión sistemática de literatura acerca de los efectos sobre la salud de la contaminación del aire en ALC, no figuran publicaciones de la Argentina (OPS, 2005). Estudiando las fortalezas y debilidades de la

gestión de la calidad del aire en la Argentina, Puliafito (2009) señala "la ausencia de un sistema de gestión ambiental del recurso aire para diversos ámbitos; municipales, provinciales, nacionales e internacional". Esto revela, no solamente en que medida la carencia de registros de la calidad del aire limita la posibilidad de evaluar su impacto sobre la salud humana (Bell et al., 2006) y otros aspectos del ecosistema (García-Huidobro et al., 2001; Mölders, 2012) sino como pueden quedar enmascarados los altos costos implicados (Miranda, 2006; Sánchez-Triana et al., 2007). A esta situación se le agrega el hecho de que los valores límite para los contaminantes del aire en las regulaciones argentinas se hallan, en la actualidad y en la mayoría de los parámetros, atrasados respecto de los lineamientos internacionales de la OMS (Organización Mundial de la Salud), la EPA (Environmental Protection Agency) de EUA y la agencia ambiental de la Unión Europea (CAI, 2012); en particular la Provincia de Buenos Aires (Sosa, 2015) posee un vacío importante en cuanto a los niveles de $PM_{2,5}$ (material particulado de diámetro inferior o igual a 2.5 micrones) y los HAP (hidrocarburos aromáticos policíclicos) adsorbidos en tales partículas. En ese mismo texto, se señala que "la normativa no es eficiente respecto al control de los COVs (compuestos orgánicos volátiles)".

Contexto global, la ciudad como hábitat y la salud

A nivel mundial se han establecido agendas basadas en los índices de calidad de aire para lo cual el monitoreo de los contaminantes clave resulta fundamental (CEPAL, 2006; Gurjar et al., 2008; NU, 2013). La necesidad de un control ambiental está basada en la influencia que los ambientes biofísicos, sociales y económicos tienen sobre la salud humana (Lebel, 2005; Andrade y Scarpatti, 2008). Del conjunto amplio de categorías (cáncer, tuberculosis, inmunodeficiencias, diarrea, etc.) involucradas en el cálculo de la Carga Global de Enfermedad (CGE) -un parámetro que abarca causas de perjuicios a la salud, enfermedades y muerte- las enfermedades respiratorias agudas (incluyendo las de origen viral y bacterial) representan en los países menos desarrollados el 9.4% siendo el mayor de los porcentajes dentro de los factores de riesgo ambiental. En los países desarrollados este porcentaje es solo 1.6 (Smith et al., 1999) aunque cabe agregar que en Europa, donde existe un sostenido cumplimiento de las leyes ambientales, muchos expertos y otras partes interesadas perciben que los estándares de calidad de aire no son todavía seguros (WHO, 2013) mientras que el documento de WHO (2006) pone en evidencia la necesidad de realizar acciones para proteger la salud de los niños. Prüss-Üsütun y Corvalán (2007) muestran que la porción ambiental de la CGE es globalmente 24%. Cifuentes et al. (2005) estiman que en ALC por lo menos 100 millones de personas están expuestas a niveles de contaminación del aire por encima de los recomendados por la OMS. Un comunicado de prensa de este organismo (WHO, 2014) establece que "la contaminación atmosférica constituye en la actualidad, por sí sola, el riesgo ambiental para la salud más importante del mundo" dado que una de cada ocho del total de muertes en el mundo es debida a la exposición a la contaminación atmosférica.

Más de la mitad de la población mundial habita en áreas urbanas (Cochrane, 2008; Kruijt y Koonings, 2009), en ALC la cifra es cercana al 80% (PNUMA, 2012). Estas áreas son vulnerables a los cambios climáticos globales actuales y estos ya no son solo un tema "tradicional" de interés científico o de organizaciones ecologistas sino que, debido a su dimensión psicosocial, abarcan a otros actores: tomadores de decisiones, empresarios, medios de comunicación y organizaciones no gubernamentales (Urbina Soria y Martinez Fernandez, 2006). Son varias las causas que originan el cambio climático global (Hay et al., 2002; Mu y Mu, 2013); en muchos estudios, tales como los de Gasper et al. (2011), Krämer et al. (2011), Rosenzweig et al. (2011) y Kraas et al. (2014) se demuestra que, en todo el mundo, las ciudades (por su ubicación y actividades económicas) enfrentan grandes

desafíos en relación al aumento del nivel del mar, aumento de precipitaciones o desertificación, daños de la infraestructura, escasez de agua, etc. incluyendo la calidad del aire (Harlan y Ruddel, 2011). Jacob y Winner (2009) compilan varios estudios que dan cuenta del efecto del cambio climático global sobre la calidad del aire; en ellos se señala por ejemplo que, manteniendo las emisiones actuales constantes se prevé que en algunas zonas el O_3 troposférico aumente entre 1 y 10 ppbv en las próximas décadas solo debido al cambio climático (aumento de la temperatura). Los autores señalan la importancia de reducir las emisiones así como de generar modelos predictivos más confiables y consensuados. Es de considerarse (NU, 2013) que desde 1990 a 2010 el CO_2 (gas de efecto invernadero) va en aumento en todo el mundo.
Como señala Fenger (1999), la ciudad reúne a los más altos niveles de contaminación junto a la mayor cantidad de agentes de recepción. En "Contaminación del Aire en el Siglo XXI: Asuntos Prioritarios y Políticas" Smook (1998) recuerda que, dada la complejidad de las amenazas ambientales, la ciudad como tal es un sitio ambientalmente peligroso para vivir y señala que un posible eslogan para una buena actitud hacia el ambiente urbano podría ser: "pensar comprehensivamente y actuar localmente" aunque enseguida agrega que esto implica "planificar globalmente e implementar localmente". Este enfoque, no solo hace evidente la importancia de la conciencia ciudadana y de las políticas ambientales públicas (Godish, 2004; FARN, 2013) sino que pone en valor el aspecto "local" que es donde se apoya esta tesis.
Por lo expuesto arriba, ha de destacarse el rol que pueden tener las disciplinas científicas y tecnológicas (tales como meteorología, química, ingeniería, estadística, ecología y medicina ambiental (Ayres et al., 2010)) en el abordaje de las problemáticas implicadas y en el contexto de la amplia gama de actores sociales que intervienen (Sportisse, 2008). De esto último, así como de la dimensión ética implicada, da cuenta la abarcativa Carta Encíclica Papal (CEP) "Laudato Si" (CEP, 2015).
Martínez y Romieu (1997) señalan que la toma sistemática de datos no solo influye en la confiabilidad de los mismos sino en el desarrollo de estrategias de control. Un estimulante ejemplo de programas de reducción de contaminantes puede verse en Holland et al. (2004), sobre los beneficios que aportan el estudio de los contaminantes del aire en la salud en Krämer et al. (1999), Jaakkola et. al (1999) y Bell et al. (2011) y sobre la influencia en la reducción de costos por parte del Estado y de particulares se encuentra en Hall et al. (2010). Una perspectiva más amplia está representada en publicaciones como las de Bates (1995), Jedrychowski et al. (1999) y WHO (2005) dedicadas a la importancia de la contaminación del aire en la salud y desarrollo de los niños, los estudios de Bard et al. (2010) que ponen de relieve la estratificación social, el de Salas-Cárdenas y Sánchez-Gonzáles (2014) que da cuenta de la relación salud-ambiente que viven los adultos mayores o los de Lioy (1990; 2006) que señalan la importancia de evaluar la exposición total de los individuos a los agentes contaminantes: esto último implica considerar a todos los aportantes (suelo, agua, alimentos, aire, plantas) y todas las rutas de entrada al organismo (inhalación, ingesta, dérmica, sexual).

Contexto local: "estado del arte" de la calidad del aire en La Plata y alrededores

Desde una perspectiva académica y siguiendo a Albritton (1994) es conducente preguntarse que es lo que se conoce y lo que no se conoce de una determinada temática, que significado tiene el "estado del arte" de la misma y que es lo que debería ser encarado a futuro. Con distinto grado de profundidad, la presente tesis trata de dar algunas respuestas, en relación a los parámetros ambientales de La Plata y alrededores, desde el punto de vista aplicado, principalmente, en lo que hace a algunas características de los vientos y su relación con los contaminantes del aire.

Barros et al. (2005a), en el Capítulo 2, indican que la zona sudeste de Latinoamérica (comprendida entre las latitudes 20ºS- 50ºS y las longitudes 45ºO- 65ºO) es crecientemente vulnerable a eventos climáticos e hidrológicos extremos como consecuencia de los cambios globales que han tenido lugar a partir de la década de 1970. En Barros et al. (2005b) se muestran de manera sencilla futuros posibles escenarios. La ciudad de La Plata, fundada en 1882 en las cercanías de la desembocadura del Río de La Plata en Sudamérica, se halla ubicada en esa franja (35ºS 58ºO). La misma fue concebida urbanísticamente como un modelo de "metrópoli sana" (Cowen, 2010) y actualmente, junto con sus alrededores constituye una zona densamente poblada (aproximadamente 800 000 habitantes) en donde se desarrolla una gran actividad industrial (Polo Petroquímico (IPA, 2011), industria siderúrgica, astillero, etc.), posee una Central Térmica de generación eléctrica (560MW de capacidad) puesta en operación en 2012, un puerto naviero, un aeropuerto y un gran parque automotor. Un dato importante, tanto en el presente como para el futuro, es que los partidos de La Plata y Ensenada (que junto a Berisso conforman el Gran La Plata) son lindantes de partidos del tercer cordón poblacional de la Ciudad Autónoma de Buenos Aires (una megaciudad de aprox. 3 millones de habitantes (al 2010), cuyos alrededores se extienden decenas de kilómetros en tres cordones concéntricos dando lugar en su conjunto al Gran Buenos Aires con aproximadamente 10 millones de habitantes) con gran crecimiento poblacional y urbanístico. Petcheneshsky et al. (1998) informan que La Plata es una de las seis ciudades potencialmente más contaminadas de la Argentina. En un estudio ambiental (AAPLP, 2006), dedicado principalmente al estudio de los suelos de La Plata y alrededores, se señala que entre los riesgos antrópicos "los de la contaminación son unos de los más importantes y no solo del suelo sino del aire y del agua". Gassmann y Mazzeo (2000) realizaron un estudio regional de la contaminación potencial del aire en Argentina y llegaron a la conclusión de que La Plata está localizada en una zona con baja capacidad de autodepuración atmosférica. Díscoli y Barbero (2001) relacionaron la capacidad de absorción de CO_2 (dióxido de carbono) del medio natural con la emisión de CO_2 debido a los consumos energéticos urbanos para cuantificar el grado de equilibrio energético-ambiental del partido de La Plata. Encontraron que el intercambio de flujos era altamente desproporcionado en detrimento del medio natural, reflejando "una encrucijada difícil de abordar en el marco del patrón de crecimiento actual". El modelo de atlas urbano-ambiental de La Plata (San Juan et al., 2006) se vería enriquecido al contar con registros permanentes de la calidad del aire. En el contexto de un modelo de calidad de vida urbana para la ciudad Dicroce et al. (2010) dan cuenta de la presencia significativa que tiene la contaminación del aire tanto en el casco urbano como en las periferias. Por su parte, Blanco y Porta (2013) indican que la ciudad y los alrededores poseen características ambientales que hacen que la salud pública posea un cierto nivel de riesgo.

A pesar de estos hechos no existe en la zona una red oficial de monitoreo de los contaminantes del aire ni las estaciones meteorológicas correspondientes. Vale decir que, sumado a los desafíos que se imponen en la zona debido al cambio climático global, la ciudad debería asumir la vigilancia sistemática y continua de los principales contaminantes del aire.

Varias ciudades importantes de la Argentina tales como Buenos Aires (Mazzeo et al., 2005; Arkouli et al., 2010; Fujiwara et al., 2013), Córdoba (Olcese y Toselli, 2002; Diez et al., 2013, Achad, 2015), Mendoza (Puliafito et al., 2003; Allende et al., 2013, 2015), Santa Fe (Caminos et al., 2011) y Bahía Blanca (Puliafito et al., 2007; Arranz et al., 2015) entre otras, han realizado esfuerzos por evidenciar los problemas crecientes debidos a la contaminación del aire, algunas de ellas poseen redes de vigilancia con mayor o menor alcance y continuidad. Además, las ciudades de Buenos Aires, Córdoba (PNUMA, 2010), Rosario y Tucumán (PNUMA, 2007) han asumido compromisos ante la comunidad

internacional al suscribirse al Proyecto Geo Ciudades (PNUMA, 2012).
En La Plata algunos reportes "históricos" de la calidad del aire lo constituyen Mazzeo et al. (1971, 1972), Mazzeo y Nicolini (1974), Cattogio et al. (1989) y Cattogio (1990). En 2001, en virtud de la Ordenanza Municipal N° 8863/98 se creó el "Observatorio de Calidad de Vida del Partido de La Plata" (OCVPLP) que toma como uno de los indicadores de la calidad de vida a la calidad del aire, mediante el seguimiento de la "evolución de los niveles de contaminación ambiental ... gaseosa...". En su primer y único documento (MLP-UNLP, 2001) el OCVPLP no llega a dar cuenta de la situación del recurso aire en la zona.
Varios trabajos de investigación han contribuido a la caracterización de diferentes aspectos de la contaminación del aire en la ciudad y alrededores (Colombo et al., 1999; Ronco et al., 2001; Massolo et al., 2002; Marañon Di Leo et al., 2004; Rehwagen et al., 2005; Nitiu, 2006, Negrin et al., 2007, Massolo et al., 2010, Colman Lerner et al., 2012, 2014; Orte et al., 2015). Entre otros contaminantes destacan los elevados niveles de material particulado (MP_{10} y $MP_{2,5}$), HAPs y COVs en el casco urbano y en áreas cercanas al complejo industrial. Bilos et al. (2001) y Rehwagen et al. (2005) señalan la importancia de las variaciones estacionales de los metales y de los HAPs ligados al material particulado en el aire. Whichmann et al. (2009) revelan el efecto adverso del material particulado y los COVs provenientes de las fuentes industriales sobre la salud de los niños en diferentes áreas de la ciudad. Cabe destacar que en estos trabajos las referencias a la distribución horaria de los vientos son prácticamente nulas.
Son varios los factores que señalan la necesidad de caracterizar los patrones de viento de superficie. Por un lado la cantidad, la magnitud y la ubicación de las distintas fuentes de emisión. Por otro, la carencia de información sobre la calidad del aire (WHO, 1998; SPA, 2007; PNUMA, 2004, 2012; CAI, 2012), el carácter no específico de los datos oficiales sobre los vientos en relación los contaminantes, el carácter fragmentario y la diferente calidad de los datos meteorológicos disponibles en las distintas instituciones y la escasez de estudios sistemáticos de seguimiento de los contaminantes.

I.1.2 Meteorología y contaminación

Es de notar que los eventos de contaminación del aire más graves registrados en el mundo, ya sean de origen antrópico o natural, tales como el de Meuse Valley- Bélgica (1930), el de Donora- Pensylvania (1948), el de Londres (1952) (Jacobson, 2002), el de Nueva York (1966) (Fensterstock y Fraunkhouser, 1968), el de Londres (1991) (Anderson et al., 1995) entre otros o el evento de "humo" en La Plata (abril- mayo de 2008) originado por la quema descontrolada de pastizales en el delta del Río Paraná, ocurrieron a partir de grandes emisiones en concomitancia con condiciones meteorológicas desfavorables para la dilución de los contaminantes (fuertes inversiones, bajas velocidades de viento, etc.). Es oportuno agregar aquí que, además de la carga "local" de contaminantes debido a las distintas actividades económicas, existen aportes "regionales" que pueden ser tenidos en cuenta tales como las emisiones del volcán Puyehue- Cordón Caulle de Chile que afectaron una gran parte de la Argentina en 2011 (Otero et al., 2012) o los debidos a los productos de quema de biomasa transportados por las corrientes a chorro bajas (Ulke et al., 2007). Pero más allá de las anomalías, la meteorología se halla en el centro de la relación entre la contaminación del aire y la salud humana (McGreggor, 1999). De aquí también, la importancia de contar con registros aptos para la toma de decisiones ante emergencias así como para la elaboración de modelos de difusión en micro y mesoescala (Mattio, 2009; Blanco y Berri, 2013).
El viento fue elegido como parámetro fundamental de estudio por ser el principal agente de transporte de los contaminantes del aire y debido a que los datos crudos de varias estaciones meteorológicas poseían buena completitud y confiabilidad. Pero cabe destacar

que la escasez, ausencia o poca confiabilidad de otros datos de interés ambiental (tales como altura de capa de mezcla, turbulencia, estabilidades atmosféricas y humedad) no permitieron enriquecer los resultados de índole predictiva.
En relación a los contaminantes del aire se adoptó al SO_2 (dióxido de azufre) como especie principal debido a su importancia como gas testigo de las emisiones industriales (Smith et al., 2010) y por constituir uno de los contaminantes clave (WHO, 2000a,b, 2006, 2013). Este gas es muy reactivo (la temperatura y la humedad son variables que juegan un papel importante en el porcentaje de conversión de las reacciones del SO_2 en la atmósfera) reaccionando tanto en fase gaseosa como líquida (Seinfeld y Pandis, 2006). El dióxido de azufre produce aumento de la corrosión de varios materiales (Graedel, 1994) y afecta la vida de los organismos vivos (Godish, 2004); combinado con la humedad ambiente reduce la visibilidad (Wark et al., 1998); es el principal precursor de la lluvia ácida y en concentraciones promedio anuales de 10 ppbv (parte por billón en volumen) en presencia de material particulado tiene impacto sobre la incidencia de enfermedades respiratorias (US ASTDR, 1998). El tiempo de residencia medio del SO_2 en una atmósfera limpia varía entre de 2 y 6 días (Godish, 1997) hasta 10 días (US ASTDR, 1998) pudiendo llegar a semanas (Sigrist, 1994).

I.1.3 Estadística y ambiente

Como se señaló anteriormente, existe a nivel mundial una conciencia creciente de los problemas ambientales que enfrenta la humanidad. Estos problemas son vastos, abarcando temas tales como conservación del ambiente, evaluación y control de la contaminación, monitoreo de ecosistemas, gerenciamiento de recursos, cambio climático, efecto invernadero, agricultura, etc. (Barnett, 2004). Tanto los individuos como las organizaciones y los gobiernos se ven llamados a proteger el ambiente y esto ha generado diversas respuestas en los últimos 30 años, algunas de ellas implican el desarrollo de especialidades tales como la "estadística ambiental". Esta especialidad difiere de otras, tales como "estadística industrial" o "estadística médica", en el énfasis y en la variedad temática. El surgimiento de la estadística ambiental se fundamenta tanto en la necesidad actual que impone la temática como en la complejidad de los temas a abordar (Piegorsch y Bailer, 2005). Los matemáticos dedicados a la estadística (estadísticos) desempeñan un rol protagónico para evaluar las "incertidumbres" y las "variaciones" de los problemas ambientales. A ellos se los confronta con la necesidad de desarrollar o adaptar métodos específicos que permitan encarar y comprender mejor los temas ambientales (Barnett, 2004). Algunos textos tales como "Encyclopaedia of Environmetrics" de El-Shaarawi y Piegorsch de 2002, en sus varios volúmenes, dan cuenta de las vastedades temáticas y metodológicas involucradas en la estadística ambiental.

I.1.4 Análisis inferencial y exploratorio

Dadas las características inherentes de las bases de datos ambientales (grandes cantidades de datos y muchas variables involucradas) y el objetivo de elaborar conclusiones con base temporal horaria se recurrió al análisis de datos. Según Tukey (Tukey, 1977) este abordaje (que es una aspecto de la estadística matemática) ha dado lugar a la distinción entre *análisis inferencial* (o confirmatorio) y análisis *exploratorio*.
El análisis inferencial refiere a la estadística clásica en cuanto a que supone que los datos siguen un modelo, sobre cuyos parámetros se trata de obtener conclusiones en la forma de estimadores, intervalos de confianza y tests.
El análisis exploratorio refiere a la identificación de patrones (regularidades) en los datos (Behrens, 1997; Mirkin, 2011) a través de un conjunto de métodos. Desde esta perspectiva no es fundamental conocer el origen de los datos, se considera que los datos disponibles

reflejan las propiedades del fenómeno que se quiere estudiar. Este tipo de enfoque provee de un entendimiento básico de los datos y de las relaciones entre las variables puestas en juego (Figueras y Gargallo, 2003) y según Mirkin (2005) puede ser utilizado cuando hay ausencia de conocimientos teóricos o conceptuales claros y/o se desconocen las regularidades subyacentes. Es decir, puede no contarse con un conocimiento *a priori* sobre la naturaleza de las relaciones entre los objetos o las variables (Marques de Sá, 2007). Un término que suele utilizarse casi como sinónimo de análisis exploratorio de datos es el de minería de datos, sin embargo Hand et al. (2001) señalan que no refieren estrictamente a lo mismo; minería designa trabajar con grandes masas de datos en donde pueden requerirse estrategias especiales de abordaje (Holmes y Jain, 2012) mientras que una exploración puede realizarse con un grupo más o menos pequeño de datos. Ambas tienen en común que emplean herramientas estadísticas para el cómputo y la visualización (Gorunescu, 2011) y que tratan de minimizar las suposiciones (Velleman y Hoaglin, 2004) sobre la naturaleza de los datos a tratar.

Lo que en la perspectiva inferencial es muy importante, como por ejemplo, la consistencia de un coeficiente empleado, desde la perspectiva del análisis exploratorio puede no ser necesario (Mirkin, 2005) porque el énfasis está puesto en "una descripción rica de los datos" (Behrens, 1997); comparativamente el enfoque exploratorio puede proveer más información que la que aporta la simple constatación de un test. Ejemplo: si no se conoce o no se puede suponer el cumplimiento de un modelo estadístico en los datos será poco relevante estimar la "bondad" del método que se emplee mediante un test, puesto que no se podrá verificar su "error". Algunos autores (Tukey, 1977; Behrens, 1997) señalan que aún teniendo buena información de partida es bueno realizar un análisis exploratorio para luego realizar el análisis inferencial (carácter complementario).

El análisis exploratorio le presenta al investigador un conjunto de métodos para realizar búsquedas efectivas en los datos e intenta realizar descripciones simples y fáciles de interpretar. Constituye un verdadero (otro) punto de vista que pone de relieve aspectos no esperables desde el punto de vista inferencial (Tukey, 1977) o, como dicen Velleman y Hoaglin (2004), el análisis exploratorio ha agregado "una nueva dimensión en la forma en que la gente puede acercase a los datos". Es común en este enfoque adoptar criterios heurísticos (prácticos, informales) lo cual implica que las suposiciones no se hallan explicitadas (Everitt et al., 2011). Estos criterios van asociados a la toma de decisiones rápidas y "frugales"; conceptualmente abandonan la idea de certidumbre (Gigerenzer et al., 1999). En contraste, los métodos empleados en la perspectiva inferencial están diseñados para ser "los mejores" posibles siempre que se cumplan ciertas suposiciones establecidas. Pero cuando la situación práctica se aleja de tales suposiciones estos métodos suelen comportarse incorrectamente.

En las distintas aplicaciones de la tesis fueron empleadas ambas perspectivas. Esto se plasmó a través de una variedad de métodos de cálculo y gráficos. Para los casos univariado y bivariado (una o dos variables puestas en juego) se utilizó tanto el enfoque inferencial como el exploratorio mientras que para el caso multivariado se utilizó preferentemente el análisis exploratorio.

I.1.5 Estadística clásica y robusta

Otra elección realizada de forma simultánea a las antedichas la constituye el hecho de trabajar con el concepto de robustez estadística.

La estadística robusta tiene puntos de referencia en el siglo XIX pero la mayor parte de los desarrollos fueron llevados a cabo en la segunda mitad del siglo XX con los aportes fundamentales de John Tukey, Peter Huber y Frank Hampel (Maronna et al., 2006). Ortega Dato hace una interesante reseña de la evolución de los métodos robustos (Ortega Dato,

2001). En los últimos 50 años el campo de investigación en estadística robusta se incrementó sustancialmente y, recientemente, los distintos paquetes de software fueron incorporando elementos de cálculo aunque, como señala Maronna (Maronna, CP), no se hallan plenamente difundidos en la medida de los beneficios que proporcionan.

En el enfoque clásico de la Estadística se supone que los datos siguen un modelo que se cumple *exactamente.* Pero los procedimientos que se deducen de esta suposición pueden fallar en el caso más realista en que el modelo se cumpla sólo *aproximadamente*. Esto ocurre en particular cuando el modelo supone una distribución normal.

Sin embargo, los datos pueden tener intrínsecamente otra distribución, o puede haber algunos que se alejen del grueso de las observaciones (llamados valores atípicos o simplemente atípicos (Peña, 2002)). Por ejemplo, al estimar una medida de tendencia central o los parámetros de una regresión lineal con errores normalmente distribuidos es posible que las suposiciones en torno a la normalidad solo se cumplan parcialmente, debido a la presencia de observaciones que siguen otro patrón o sencillamente ninguno. En estos casos habrá un alejamiento de la normalidad. Estos comportamientos son frecuentes en el análisis de datos y en el modelado estadístico (Maronna et al., 2006).

Los valores atípicos pueden deberse a eventos excepcionales, a errores, o bien pertenecer a otra población (Rousseeuw y Hubert, 2011). Cuando el análisis de datos se realiza de manera clásica (por ejemplo al estimar la media aritmética), la importancia de la presencia de valores atípicos reside en que con solo uno de ellos se puede producir una gran distorsión en el valor de los estimadores. Si el o los valores atípicos producen colas largas ("pesadas" o "aplanadas") en la función de densidad de distribución esto repercutirá en una varianza innecesariamente grande y si los valores atípicos producen asimetría en una de las colas esto repercutirá en un sesgo importante (Maronna et al., 2006); ambas situaciones producen el alejamiento de la "normalidad".

Uno de los enfoques para tratar con los valores atípicos es el de los *diagnósticos estadístico*s; existen textos con capítulos dedicados a esto (Cook y Weisberg, 1999) y otros dedicados enteramente, tal como el de Belsley et al. (2004) (para regresión), en el que se utilizan métodos específicos, tanto gráficos como numéricos, con el fin de poner en evidencia la existencia de desvíos respecto de un modelo. Los métodos diagnósticos pueden dar una descripción amplia de los datos (lo cual puede ser uno de los objetivos de la investigación), sin embargo, presentan dos desventajas: a) no son siempre confiables, y b) una vez que se detecta el valor atípico, la decisión de dejarlo o descartarlo queda a criterio del investigador (decisión subjetiva) (Maronna et al., 2006). Si bien, es mejor realizar un diagnóstico que no hacer nada, el enfoque robusto puede ser más confiable y más abarcativo. Confiable en el sentido de que la estimación de los parámetros no se halla fuertemente influenciada por la presencia de los valores atípicos. Abarcativo en el sentido de que funciona bien tanto para cuando los datos siguen, por ejemplo, una distribución normal sin valores atípicos como para cuando la distribución se aparta algo de la normal debido a la presencia de los mismos.

Los estimadores robustos se han diseñado específicamente para el caso en que la desviación de un modelo (por ejemplo, distribución normal) se deba a la presencia de valores atípicos (se habla de muestra contaminada). De modo más formal, si se asume que el grueso de los datos provienen de una distribución G, el estimador robusto para tales datos se diseña para que se comporte satisfactoriamente para una distribución $G_\varepsilon = (1-\varepsilon)\ G + \varepsilon\ H$ donde H es otra distribución y $\varepsilon(0,1)$ es el término de "error" que está distribuido normalmente con media cero y varianza unitaria. Por ejemplo, si G es la distribución normal, H es la distribución de donde provienen los atípicos y ε representa el grado de contaminación de la muestra, los estimadores robustos que se obtengan a partir de este diseño funcionarán bien cuando haya colas "pesadas" cercanas a la normal, tales como se

da el caso de la distribución *t* de Student (Filzmoser et al., 2009). Es importante señalar que el enfoque robusto es tradicionalmente conocido por brindar estimadores "resistentes" a los potenciales valores atípicos pero comprende también la estimación de intervalos de confianza y tests robustos.

La estadística robusta propone ajustar los datos de tal manera que sea similar al ajuste dado por el enfoque clásico cuando no hay valores atípicos (Rousseeuw y Hubert, 2011). Ejemplo: se tiene una muestra de cinco datos y se quiere estimar la tendencia central de los mismos:

6.27 6.34 6.25 6.31 6.28

la media (estimador clásico de posición) es $\bar{x}$ =6.29. Supongamos que el cuarto dato se registra erróneamente asignándose el número 63.1. En este caso la media (contaminada) $\bar{x}_C$ =17.65 se halla lejos del valor 6.29. Si se estima la mediana (otro estadístico de tendencia central) la misma será 6.28 que es un valor razonable (aún con la presencia de un valor atípico). Se dice entonces que la mediana es un estimador robusto de la tendencia central de los datos y que la media es muy sensible (poco robusta) a los valores atípicos.

Un criterio para la detección de valores atípicos que suele aparecer en los textos es el de la "regla de los dos (o los tres) sigmas" que establece por ejemplo, que si $|x-\bar{x}|>2s$ siendo *s* el desvío estándar de la muestra, entonces *x* puede considerarse como atípico. Dada la muestra:

2 3 4 5 6 7 8 9 10 50

se estima la media $\bar{x}=10.4$ y el desvío estándar (estimador de escala) $s=14.15$.

Puesto que $|50-10.4|=39.6>2\ x\ 14.15=28.3$ se concluye que el valor 50 es atípico. Si reemplazamos el anteúltimo dato (10) por el de 50 obtendríamos $\bar{x}_C$ =14.4 (media contaminada) y $s_C=18.9$ (desvío estándar contaminado). Ahora $|x-\bar{x}_C|=35.6$ que es menor que 2x18.9= 37.8 por lo que el valor 50 ya no se puede considerar un atípico. Este "problema" surge de poner en juego, en la detección de atípicos, estimadores que son influenciables por los mismos.

Para evitar situaciones de este tipo es necesario trabajar con estimadores más robustos. Para estimar la escala (desvío o dispersión) de los datos de manera más robusta una opción la constituye la MAD (desvío absoluto de la mediana) que se calcula obteniendo el valor absoluto de la resta entre cada dato y la mediana y luego calculando la mediana. Entonces cabe preguntarse ¿porque no abandonar el uso de los estimadores clásicos y calcular todo con mediana y MAD? Una respuesta informal a esto lo constituye el hecho de que en algunos casos en que no hay atípicos, los estimadores robustos suelen ser menos eficientes (Maronna et al., 2006). La eficiencia de un estimador mide la relación entre su varianza y la varianza del "mejor" estimador para el modelo. Por ejemplo, el método de regresión lineal por cuadrados mínimos $y=\beta x+\varepsilon$ da estimadores con mínima varianza cuando los "errores" ε están normalmente distribuidos. Cualquier otro método tendrá una varianza algo mayor, en particular los métodos robustos. Por lo tanto, al trabajar con un método robusto será importante tener en cuenta no solo su robustez sino también su eficiencia (Filzmoser et al., 2009). Por ejemplo, la media tiene una máxima eficiencia en poblaciones normales pero es muy poco robusta; en cambio la mediana es robusta pero su eficiencia en el caso normal es del 67% (Maronna et al., 2006). Estas dos características deben ser sopesadas según el objetivo de la investigación. Los mejores métodos robustos combinan alta resistencia a valores atípicos (robustez) con alta eficiencia (Maronna y Yohai, 2014).

Muchos métodos robustos pueden ser descriptos en términos de los clásicos a los que se les

ha asignado una función de pesos diferenciales según el dato. O sea, la mayor parte de los datos recibirán un peso muy similar pero algunos (los potenciales atípicos) recibirán menos peso. Esto puede traducirse en que la robustez esté asociada al "ajuste de la mayoría" (Filzmoser et al., 2009), lo cual impide que se trunquen algunos datos (con la consecuente pérdida de información), haciendo simplemente que estos tengan menos importancia. Por otra parte, una vez que se han obtenido los estimadores robustos es posible identificar los potenciales valores atípicos presentes en los datos y evaluarlos a la luz del tema a tratar.
Tukey (1977) recomienda aplicar el método clásico *y* un método robusto, y comparar los resultados. Si éstos son "parecidos", quedarse con el clásico, y si no, analizar los datos para encontrar el origen de la discrepancia. Desde un punto de vista práctico, dadas la variedad de métodos robustos y las diferencias entre sus performances, conviene que la elección de los métodos esté orientada por un especialista en robustez.

I.1.6 Mutidisciplina e interdisciplina

La presentación realizada hasta aquí induce a pensar en la interrelación de disciplinas (tales como ingeniería ambiental, meteorología y estadística) con base en la física, la química y las matemáticas (estas últimas fundamentales en el campo de las ingenierías).
Tradicionalmente un trabajo multidisciplinar implica una yuxtaposición de disciplinas, cada una aportando su punto de vista con mayor o menor grado de integración. El trabajo interdisciplinario implica, en distinto grado de desarrollo, una síntesis que involucre a más de una disciplina. Finalmente, el trabajo transdisciplinar es más holístico, implica alcanzar una alto grado de integración del conocimiento (Palmer, 2001).
El mundo real de los trabajos de investigación se ve, en general, confrontado a buscar soluciones más allá de categorías disciplinares. Pero el proceso de investigación en la interfaz de dos o más disciplinas no ocurre automáticamente, más bien constituye un desafío y un estímulo a la creatividad (Brewer, 1999; Lyall et al., 2011).
Sin la cohesión con la que trabaja un equipo interdisciplinario en un único proyecto, el proceso de esta tesis promovió el intercambio multidisciplinar e interdisciplinar con áreas de la física en general, la óptica, la espectroscopia, la química ambiental, la meteorología y el análisis estadístico de datos. En algunos casos, el resultado de esa interacción se plasmó en la aplicación de métodos poco conocidos fuera del ámbito de la estadística y en particular dentro de las disciplinas ambientales. De esta integración quedaron descriptos fenómenos meteorológicos que involucraban muchas variables con un alto grado de síntesis de información, quedando el planteo inicial resuelto de forma no convencional (respecto de la tradición disciplinar). O sea, como fruto de la sinergia interdisciplinaria un estudio de fenómenos preponderantemente locales adquirió, por su capacidad para generalizarse, un carácter más universal desde el punto de vista académico tal como se reporta en Ratto et al. (2010b).

I.2 Organización de la tesis

En base a lo expuesto hasta aquí, situándose en los contextos local, nacional y global y, considerando a la ciudad como ámbito posibilitador del desarrollo humano, este trabajo busca proveer un panorama del "estado del arte" del recurso natural aire en La Plata y alrededores desde el punto de vista de la contaminación potencial y observada. Otro foco de interés está constituido por el análisis de los patrones horarios de viento y sus dinámicas como agente de transporte de los contaminantes fisicoquímicos del aire que, hasta el momento, habían sido muy poco estudiados. Teniendo en cuenta observaciones de parámetros ambientales y el modelado obtenido con distintos métodos estadísticos (los cuales se aplican con sentido crítico) la tesis tiene por objeto proporcionar evidencia fundamentada sobre las necesidades actuales del recurso aire y, sus resultados, podrán ser utilizados como antecedentes para la

eventual instalación de una red de monitoreo continuo de la calidad del aire en el marco de un programa ambiental de mejora de la calidad de vida.

Los contenidos y sus contextos respectivos se desarrollan a lo largo de seis capítulos.

Este primer capítulo **(Capítulo I**: **Introducción, organización y aportaciones de la tesis)** tiene como objetivos presentar al lector el tema de tesis a partir de una introducción general que la contextualiza y fundamenta, proveer de una descripción de la organización del trabajo realizado y ofrecer una síntesis de las principales aportaciones de la misma.

En el **Capítulo II (Región de estudio, datos y equipamiento de trabajo y entrenamiento en técnicas espectroscópicas**) se describen las características de los datos de trabajo, el equipamiento utilizado para realizar mediciones y las fuentes de información y provisión de datos. Se describen además las áreas de estudio (La Plata y alrededores y una zona amplia del Río de La Plata), sus características geográficas, demográficas y meteorológicas. Por otra parte se presentan dispositivos diseñados en el CIOp (Centro de Investigaciones Opticas) que fueron caracterizados y optimizados durante un período de entrenamiento en manejo de equipos de medición ambiental (ópticos, electroquímicos, etc).

El **Capítulo III (Fenómenos físicos)** está destinado a la descripción de los fenómenos físicos más importantes (principalmente atmosféricos) a los que se hace referencia en los capítulos IV y V.

El **Capítulo IV (Similitud- disimilitud, regresión y tendencia)** reúne recursos gráficos y analíticos de estadística inferencial y análisis exploratorio (principalmente para los casos uni y bivariados) que asisten a la discusión de las observaciones de SO_2 (dióxido de azufre gaseoso) y a los fenómenos meteorológicos involucrados (vientos, calmas, brisa marina).

El **Capítulo V (Análisis por conglomerados y escalamiento multidimensional)** está dedicado a discutir algunos métodos de análisis multivariado (principalmente análisis por conglomerados jerárquicos y escalamiento multidimensional no métrico) que han sido utilizados como pilares para la descripción y caracterización de los vientos en las zonas de estudio y como herramientas para la interpretación de algunos fenómenos físicos de la capa límite planetaria.
Otros métodos tales como Curvas de Andrews, diagrama de Siluetas, Componentes Principales, *k*-medias y un procedimiento para realizar agrupamiento jerárquico con restricciones, se presentan con distinto grado de profundidad para asistir en la discusión de aspectos particulares o como enfoques alternativos.

El **Capítulo VI (Síntesis y conclusiones finales)** resume los principales temas de tesis y se elaboran conclusiones sobre la temática ambiental y sobre el empleo de los métodos estadísticos aplicados. Sobre esa base se destacan las necesidades actuales del área de estudio y se realizan sugerencias.

I.3 Principales aportaciones de la tesis

1) La presente tesis reúne datos y genera información de interés ambiental de la zona de La Plata. El estudio realizado puso en evidencia la escasez de información en relación a los contaminantes del aire, tanto a nivel de bases de datos disponibles como de reportes y trabajos científicos publicados. Se evidenció además la ausencia de un organismo oficial

que compile y ponga a disposición de los interesados la información ambiental tanto a nivel del seguimiento de los contaminantes del aire como de los datos meteorológicos correspondientes. Las mediciones realizadas de SO_2, a la luz de los nuevos lineamientos de organismos internacionales tales como la Organización Mundial de la Salud, dan cuenta de la necesidad de llevar a cabo mediciones sistemáticas y de realizar estudios epidemiológicos de largo plazo. Por otro lado, la tesis constituye un estudio original que revela, analiza y produce conclusiones sobre las características horarias de los vientos de superficie (principalmente direcciones, velocidades y calmas) que son de interés en cuanto a agente de transporte de los contaminantes. El carácter local del estudio aporta una base que se deberá fortalecer en el futuro inmediato para poder afrontar compromisos más globales tales como el de Geo Ciudades (Sección I.1).

2) La tesis provee fundamentos para la instalación de una red de vigilancia de los contaminantes del aire. Tomando como punto de partida las características de la zona de estudio (cantidad y densidad de habitantes, escasez de información ambiental, actividad económica, fuentes emisoras y aspectos meteorológicos) y los hallazgos de otros investigadores en relación a la presencia de contaminantes del aire, concomitantemente con los desafíos que propone el cambio climático global en la zona y considerando los criterios de países más avanzados en materia ambiental, es posible concluir que la necesidad de establecer una red de monitoreo continuo de los contaminantes del aire de origen industrial y vehicular es imperativa.

3) Dados los fundamentos enunciados en la introducción (Sección I.1), que engrosan el trabajo de otros investigadores del campo, el producido de la presente tesis puede constituir una referencia técnica adicional que justifique y estimule la modificación y/o creación de leyes que aseguren la instalación y el mantenimiento de redes fijas de vigilancia de la calidad del aire en ciudades argentinas con problemáticas similares.

4) Dentro del campo de la meteorología de capa límite y la contaminación del aire, la tesis utiliza métodos de análisis exploratorio no convencionales para el estudio de los vientos. Estos métodos, tales como el análisis de conglomerados, el escalamiento multidimensional y las Curvas de Andrews permiten realizar una exploración rica de los datos y facilitan la síntesis de información de grandes cantidades de datos que poseen muchas variables.

5) La tesis tiene en cuenta el concepto de robustez estadística desde un punto de vista aplicado. Dadas las ventajas de este enfoque en el análisis de datos, tanto el presente texto como las publicaciones involucradas, tienden a difundir este enfoque mediante aplicaciones en el campo de las ciencias ambientales que, dados los desarrollos teóricos y de software comercial, deberían estar más difundidos.

6) La tesis posee una faceta interdisciplinaria que no solo se plasmó en resultados (publicaciones) sino que contribuyó a enriquecer la perspectiva disciplinar del doctorando con disciplinas relacionadas (tanto en conocimientos científicos como en el desarrollo de un lenguaje en común) promoviendo las relaciones humanas. También enriqueció al doctorando en aspectos operativos y de caracterización de equipos ópticos diseñados en el CIOp, constatando la posibilidad tecnológica de producir equipos nacionales.

"Numbers have an important story to tell. They rely on you to give them a voice"
Stephen Few

"I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts"
Sir Arthur Conan Doyle

Capítulo II
Región de estudio, datos y equipamiento de trabajo y entrenamiento en técnicas espectroscópicas

II.1 Características climáticas de la región
II.1.1 Generalidades

El análisis de datos meteorológicos y agentes contaminantes realizado en la tesis refiere a dos zonas: la ciudad de La Plata y alrededores y el estuario del Río de La Plata. La primera zona es de escala local y se halla inscripta en la segunda que pertenece a una escala sinóptica (Capítulo III).

La región es de clima subtropical húmedo (Arhens, 2009) según la clasificación de climas de Köppen modificada (Figura II.1). Esto implica que la temperatura media del mes más frío se halla debajo de los 18°C y por encima de los 3°C, hay presencia de humedad en todas las estaciones del año y la temperatura media del mes más cálido está por encima de los 22°C; además, existen al menos cuatro meses al año en que la temperatura media mensual es mayor a 10°C.

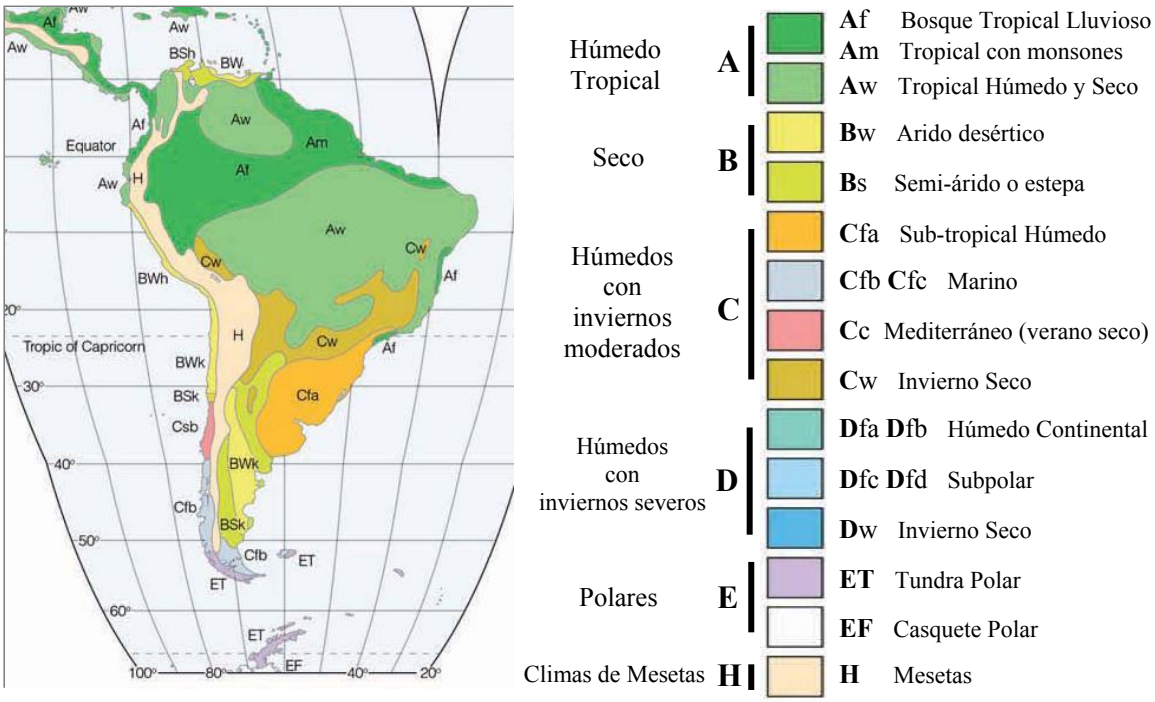


Figura II.1: Mapa parcial de clasificación mundial de regiones climáticas según Köppen modificado (Arhens, 2009). Las clases están designadas con las letras mayúsculas, las subclases poseen siglas específicas y un código de color.

II.1.2 Localización de los sitios de referencia y vientos de escala sinóptica y local

La Figura II.2a muestra un mapa del estuario del Río de La Plata con las estaciones meteorológicas de la red del SMN (Servicio Meteorológico Nacional) y estaciones de

Uruguay cuyos datos, correspondientes a distintos períodos, participaron en el análisis realizado en los capítulos IV y V.

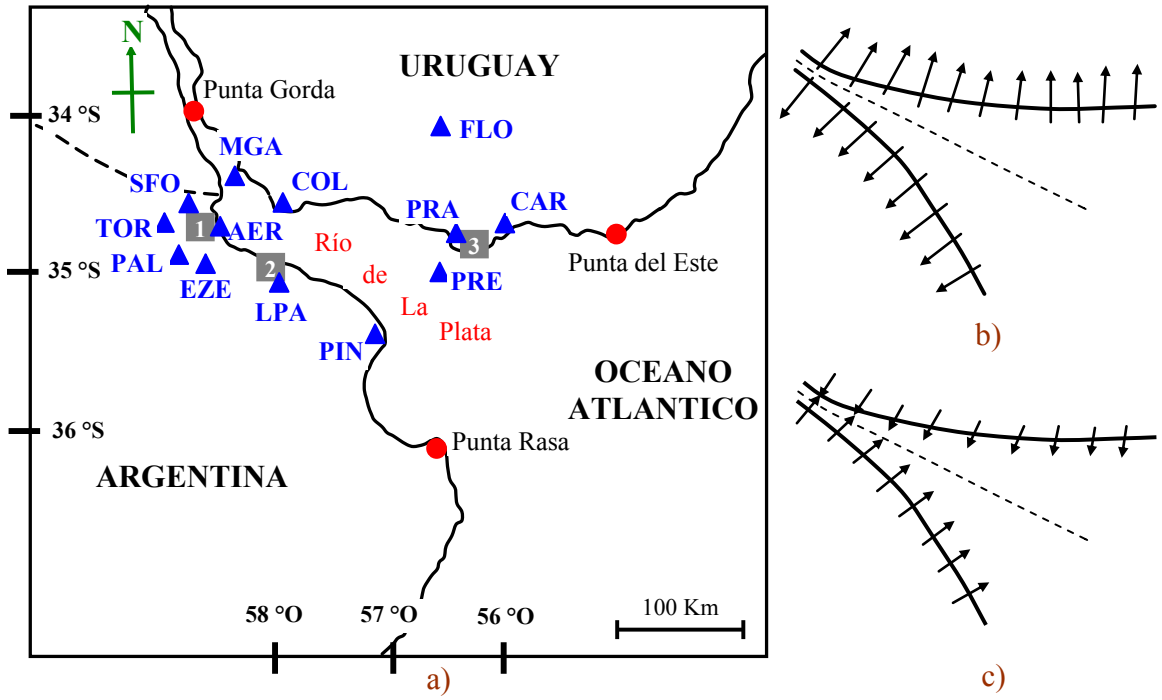


Figura II.2: a) Mapa del Estuario del Río de La Plata. La Ciudad de Buenos Aires está indicada con el número 1, La Plata con el número 2 y Montevideo con el número 3. Punta Gorda indica el nacimiento del Río de La Plata con un ancho aproximado de 1.4 km. La línea que une Punta Rasa con Punta del Este (cubre 219 km) se considera el límite del río.
Las estaciones meteorológicas de la región en orden alfabético son: Aeroparque (AER), Carrasco (CAR), Colonia (COL), Don Torcuato (TOR), El Palomar (PAL), Ezeiza (EZE), Florida (FLO), La Plata Aero (LPA) también llamada Punto K, Martín García (MGA), Punta Indio (PIN), Pontón Recalada (PRE), Prado (PRA) y San Fernando (SFO).
b) y c) son representaciones simplificadas de las costas del río, siendo la línea de rayas la zona media del río donde pueden tener lugar los fenómenos de convergencia y divergencia
b) se muestra mediante flechas la dirección hacia donde se dirigen los vientos debidos a la brisa de mar
c) se muestra mediante flechas la dirección hacia donde se dirigen los vientos debidos a la brisa de tierra (esta última con menor intensidad que la brisa de mar).

Celemín (1984) muestra para las 8 direcciones principales de la brújula las distintas configuraciones de centros ciclónicos (bajas presiones- captadores de vientos) y anticiclónicos (altas presiones- emisores de vientos) característicos del Río de La Plata. En este contexto las calmas tienen lugar cuando estos centros se hallan muy alejados entre sí respecto del río o cuando se establece un centro ciclónico sobre el río y el anticiclónico más cercano se halla muy alejado (a miles de km) del mismo.
La circulación de vientos sobre el Río de la Plata y el océano adyacente depende fuertemente del anticiclón subtropical del Atlántico sur, especialmente de su borde oriental. La ubicación de este sistema (que se combina con distintos centros de baja presión) varía durante el transcurso del año produciendo variaciones en las direcciones de los vientos sobre toda la región de influencia.

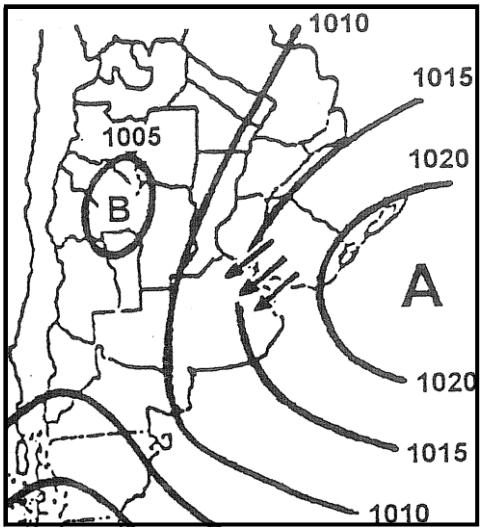


Figura II.3: Vientos característicos emitidos desde el centro Anticiclónico del Atlántico Sur (la "A" indica zona de "alta" (presión) y refiere a dicho centro; la "B" es una zona de "baja"). Aquí el centro anticiclónico "A" se halla ubicado a más de 500 km al este de Punta del Este (Uruguay) (Celemín, 1984).

En verano (estación de mayores intensidades medias de vientos en el estuario) el anticiclón del Atlántico Sur se halla en los 35°S 45°O haciendo que la dirección media del viento sea ENE (este-noreste).

En otoño la dirección media del viento es NE (Figura II.3).

En invierno la posición media del anticiclón del se ha desplazado a 30°S 45°O, la dirección media del viento sobre la mayor parte del estuario del río es NO (noroeste).

Las direcciones medias del viento observadas durante la primavera son del ENE y E, muy similares a las del verano.

En síntesis, el efecto del cambio en la posición del anticiclón es el de producir una rotación del viento desde el ENE y E (verano- primavera) hacia el NO en invierno (Barros et al, 2005).

La longitud del Río de La Plata es de aproximadamente 300 km y su cuenca cubre un área de aproximadamente 3.2×10^6 km^2 siendo una de las más extensas del planeta. Las grandes extensiones de agua y de tierra que quedan comprendidas permiten la generación de un gran contraste de temperatura permitiendo el desarrollo de circulaciones de superficie con las características de brisa de mar y tierra (Capítulo III). Debido a este fenómeno, durante la fase diurna del ciclo diario, se incrementan las componentes sur de los vientos de superficie sobre la costa norte del río (lado uruguayo) mientras que sobre la costa sur (lado argentino) se incrementan las componentes norte (Figura II.2b). El ciclo diario de contrastes de temperatura entre el agua y la tierra produce cambios significativos en las direcciones de los vientos predominantes de la zona (Berri et al. 2010).

Con el objeto de ilustrar el carácter homogéneo de los vientos de la región se muestra la Figura II.4, las partes a) y b) fueron tomadas de Ratto et al. (2010b).

En la Figura II.4a se muestra una comparación entre las direcciones de viento observadas durante el verano en LPA durante la década 1991- 2000 y el promedio de observaciones de los veranos en cinco puntos (AER, EZE, MGA, PIN, PRE- Figura II.2) de la REM (Red de Estaciones Meteorológicas) del SMN (Servicio Meteorológico Nacional) durante el período 1959- 1984 tomadas como referencia.

En la Figura II.4b se comparan las observaciones mencionadas de LPA con observaciones obtenidas en otros sitios de la ciudad de La Plata durante el período 1998- 2003: puntos A y J (Figura II.6).

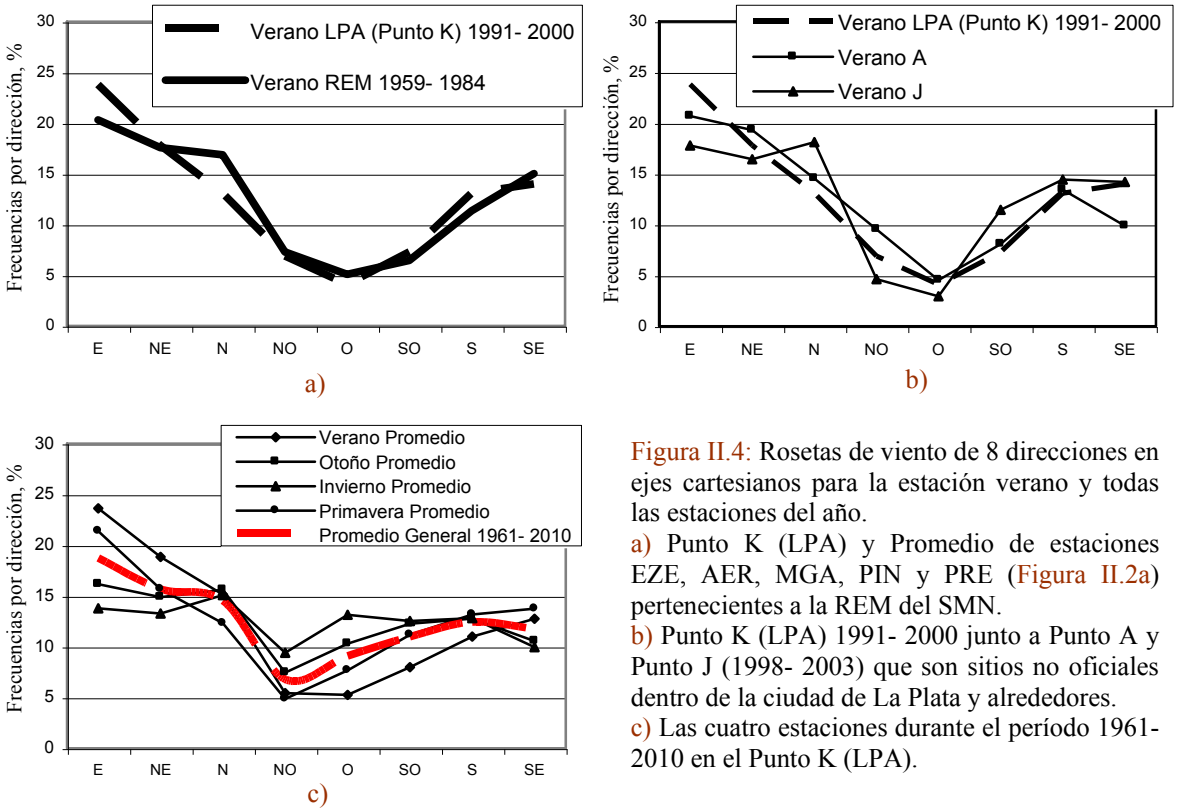


Figura II.4: Rosetas de viento de 8 direcciones en ejes cartesianos para la estación verano y todas las estaciones del año.
a) Punto K (LPA) y Promedio de estaciones EZE, AER, MGA, PIN y PRE (Figura II.2a) pertenecientes a la REM del SMN.
b) Punto K (LPA) 1991- 2000 junto a Punto A y Punto J (1998- 2003) que son sitios no oficiales dentro de la ciudad de La Plata y alrededores.
c) Las cuatro estaciones durante el período 1961- 2010 en el Punto K (LPA).

La Figura II.4c muestra las rosetas de viento promedio observadas en el Punto K durante 5 décadas (1961- 2010) (SMN, 1971, 1981, 1992, 2001, 2011). Como puede apreciarse los vientos dominantes en orden de mayor a menor presencia provienen del E, NE y N. Dependiendo de la década que se analice puede diferir el orden. Las velocidades promedio para las cinco décadas son: verano 19.8 km h^{-1}, otoño 17.7 km h^{-1}, invierno 18.5 km h^{-1} y primavera 20.5 km h^{-1}. Estas velocidades se hallan entre "brisa leve" y "brisa moderada" en la Escala de Beaufort (Sección III.4- Capítulo III).

II.2 Características de La Plata y alrededores, principales fuentes de emisión y sitios locales de referencia

En el presente trabajo de tesis se habla de La Plata y alrededores de modo genérico, pero es posible establecer algunas precisiones que dan un panorama más rico de la población expuesta a los contaminantes del aire.

La ciudad de La Plata (35°S 58°O), que suele encontrarse en los distintas fuentes de información citada también como Casco Urbano de La Plata, Casco Fundacional o Casco Urbano Fundacional es la Capital de la Provincia de Buenos Aires. Esta ciudad es la cabecera del Partido de La Plata, este último tiene una extensión aproximada de 942 km^2 (ELP, 2011) siendo uno de los 135 partidos de la Provincia de Bs. As. El partido de La Plata contiene al Casco Urbano (aprox. 25 km^2) y un conjunto de 17 Centros Comunales (AAPLP, 2006): Abasto, Arturo Segui, City Bell, Etcheverry, El Peligro, Gonnet, Gorina, Hernandez, Lisandro Olmos, Los Hornos, Melchor Romero, Ringuelet, San Carlos, San Lorenzo, Tolosa, Villa Elisa y Villa Elvira (Figura II.5).

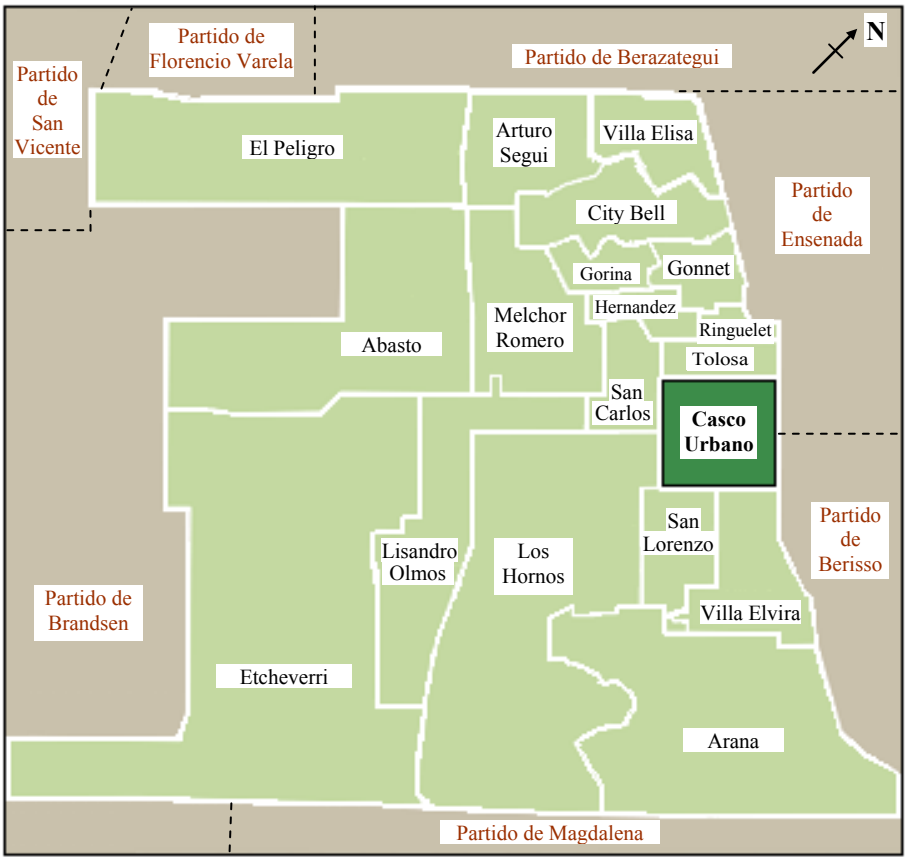


Figura II.5: Plano de La Plata (Casco Urbano) y los Centros Comunales que forman el Partido de La Plata (942 km^2) con los partidos limítrofes.

Según el Censo Nacional de 2010 el Partido de La Plata tiene 654 324 habitantes (CN, 2010) de los cuales (según estimaciones hechas a partir de las proporciones dadas por el Censo Nacional de 2001 (CN, 2001) debido a que el de 2010 no es tan completo) 636 003 habitantes residen en áreas urbanas del Partido (aproximadamente 159 km^2) y el resto en zonas rurales o de transición.
La población del Casco Urbano (utilizando el porcentaje dado en el Censo 2001) constaba en 2010 de aprox. 209 384 habitantes. Por lo tanto, es posible establecer que:
a) la densidad de habitantes total del partido (654 324 hab./942 km^2) es de 695 hab./km^2
b) la densidad urbana del partido (636 003 hab./159 km^2) es de 4000 hab./km^2 y que
c) la densidad del Casco Urbano (209 384 hab./25 km^2) es de 8375 hab./km^2.

En distinta bibliografía y fuentes de información aparece el Gran La Plata definido como la suma de las poblaciones urbanas del Partido de La Plata (636 003 habitantes), del Partido de Berisso (87 231 habitantes) y del Partido de Ensenada (56 593 habitantes). La proporción de habitantes urbanos respecto del total para Berisso y Ensenada se calculó de forma análoga a lo hecho para el Partido de La Plata, o sea, siguiendo las proporciones del Censo Nacional 2001. Por lo tanto, es posible establecer que:
a) la cantidad de habitantes del Gran La Plata al 2010 era de 779 827 y
b) que la suma total de habitantes de los tres partidos al 2010 era de 799 523 hab., correspondiendo la diferencia con el ítem anterior a la población rural.

Por otra parte, y dado que no se encontraron valores de la superficie urbana de los partidos de Berisso y Ensenada y que la cantidad de población rural era muy baja en 2001, se ha

calculado directamente la densidad de población de los respectivos partidos. Por lo tanto, es posible establecer que en 2010:
a) el Partido de Berisso tenía una densidad de (88 470 hab./135 km^2) 655 hab./km^2 y que
b) el Partido de Ensenada tenía una densidad de (56 729 hab./101 km^2) 536 hab./km^2.

La ciudad y sus alrededores se hallan ubicados geográficamente sobre una planicie típica de la "pampa húmeda" en las cercanías del Río de La Plata distando aproximadamente 56 kilómetros de la Ciudad de Buenos Aires (una megaciudad). Los alrededores poseen algunas zonas pantanosas con depósito de sedimentos en la franja costera. Un sistema de arroyos drenan las aguas en dirección perpendicular a la costa del Río de La Plata; por lo general, las zonas urbanas han invadido las planicies de inundación de los mismos, lo que genera frecuentes inconvenientes al recibir precipitaciones o cuando se produce el fenómeno conocido como "sudestada" (período usualmente de dos o tres días de persistencia de vientos fuertes del SE acompañado de lluvias y crecientes del Río de La Plata (Celemín, 1984)). La integración de las diferentes cuencas de estos arroyos mediante distintas obras hidráulicas públicas ha generado una situación en donde las inundaciones y la contaminación por vertidos en aguas superficiales adquieren una gran significación en toda la región (AAPLP, 2006).
La Figura II.6 muestra un mapa de la ciudad y sus alrededores destacándose los sitios de medición y fuentes de datos así como otros sitios de referencia.
Un complejo industrial importante (rectángulo de líneas de trazo en la figura), perteneciente al Partido de Ensenada, ubicado aproximadamente a 8 km del centro de la ciudad de La Plata, contiene a la refinería de petróleo más grande del país (con una capacidad de procesamiento de 38 000 m^3/día- (Blanco y Porta, 2013)) junto a plantas petroquímicas adyacentes de producción de hidrocarburos aromáticos (benceno, tolueno y xileno), solventes alifáticos (n- pentano, n-hexano y n-heptano), polipropileno, polibuteno, anhídrido maleico, ciclohexano, metanol, metil-terbutil eter y carbón de petróleo como principales productos.
Esta zona industrial cuenta además con industria siderúrgica (Wikipedia, 2011), un astillero y un puerto con gran movimiento naviero (Blanco y Porta, 2013). El Puerto de La Plata (fundado en 1890) se halla localizado en el Partido de Ensenada (PLP, 2015), tiene una jurisdicción total de 2249 hectáreas (22,49 km^2) y una zona exclusiva de operación de aprox. 460 hectáreas (4,6 km^2). Es un puerto fluvial (a solo 37 km fluviales de Bs. As.) dedicado preponderantemente al transporte de cargas con potencial importancia para el Mercosur. Parte de este puerto está compuesto por el Astillero Río Santiago (fundado en 1953), uno de los más grandes de América Latina (ARS, 2015), dedicado a la construcción de barcos mercantes y de guerra así como de material ferroviario. Ocupa un predio de aprox. 230 hectáreas (2,3 km^2) y contaba al 2014 con 3600 empleados.
En zonas aledañas al complejo industrial de Ensenada se halla la central termoeléctrica "Central Térmica Ensenada de Barragan" (Punto L de la Figura II.6), que dista aproximadamente 10 km de la ciudad. Tiene una capacidad de generación de 560 MW (megawatts) (ampliable a 840 MW) y es una de las "grandes" centrales de generación de energía eléctrica del país. Puede operar con gas natural y con gasoil y sus principales efluentes gaseosos son el óxido de azufre, óxidos de nitrógeno y material particulado.
Al hablar de industrias, en la tesis, se hace referencia a aquellas que se hallan localizadas en la zona del complejo industrial de Ensenada pero, es de notarse, que existe un gran número de pequeñas industrias que tienen ubicación dentro del casco urbano y en los alrededores del mismo (MLP- UNLP, 2001), existiendo además dos Parques Industriales dentro del Partido de La Plata. Uno de ellos (58 hectáreas= 0,58 km^2) se halla en pleno funcionamiento (con pequeñas y medianas industrias de tipo manufacturero, textiles, de

plásticos, autopartes, productos medicinales, veterinarios, etc.) y está ubicado en las intersección de la Ruta Nacional N° 2 y la Ruta Provincial N° 13. El otro, ubicado en el km 50 de la Ruta Nacional N° 2, se halla en desarrollo y consta de 93 hectáreas (PILP, 2015). También cabe mencionar que en el Partido de Berisso existe un parque industrial de 9 hectáreas con un gran número de pequeñas industrias con distinto grado de impacto ambiental.

Siendo la capital de la Provincia de Buenos Aires, La Plata tiene alto tránsito vehicular (más de 300 000 vehículos registrados en su partido según Whichmann et al. (2009)) que constituye otra de las fuentes importantes de los contaminantes antropogénicos del aire. En 2010 la tasa de mortalidad infantil era de 12.7 (Ministerio de Salud, 2012) y los indicadores mostraban que había 5.8% de indigencia y un 13.0% de pobreza en personas en los alrededores de La Plata (ELP, 2011). En 2010 (CN, 2010) la cantidad de habitantes entre cero y 9 años y de habitantes mayores a 65 años era de 171 711 en el Partido de La Plata. Todas estas cifras se consideran relevantes tanto desde el punto de vista social como en lo que hace a la vulnerabilidad de la salud por factores ambientales.
Siguiendo a distintos autores (Colombo et al, 1999; Rehwagen et al., 2005; Nitiu, 2006 Negrin et al., 2007) es posible señalar áreas de interés para el monitoreo de los contaminantes del aire tales como las que comprenden al puerto, el complejo industrial, el casco urbano y los barrios residenciales (Gonnet, City Bell, Villa Elisa y otros) además de zonas costeras, semirurales y rurales.

II.3 Datos de trabajo y equipamiento

II.3.1 Datos de concentración de SO_2

Los valores observados de SO_2 en aire ambiente fueron medidos a aproximadamente 2 metros de altura con la unidad analizadora Lear Siegler® que se describe en la Sección II.3.3. Estas mediciones pertenecen a dos conjuntos de datos:
a) mediciones llevadas a cabo en el Punto A (proporcionadas por la Universidad Tecnológica Nacional- Facultad Regional La Plata en Berisso) cubriendo el período 1996-2000 (Rosato et al., 2001; Ratto et al., 2006, 2010a) cuyos registros no siguen un protocolo de muestreo.
b) mediciones realizadas en el Punto D durante una campaña de 92 días (Septiembre-Diciembre de 2005) con registros tomados cada un minuto (Ratto et al., 2009).

II.3.2 Datos meteorológicos

El análisis de los vientos en relación al transporte de los contaminantes del aire se llevó a cabo a partir de varios conjuntos de datos cuyos registros fueron tomados en el área del estuario del Río de La Plata (Figura II.2) y en la Ciudad de La Plata y alrededores (Figura II.6).
a) Registros pertenecientes a la red de estaciones meteorológicas del Servicio Meteorológico Nacional- SMN (se incluyen algunas estaciones de Uruguay): son observaciones provenientes de los sitios indicados en la Figura II.2a durante 1959- 1984 y 1994- 2008. Estas estaciones proveyeron datos mensuales medidos a 10 m de altura sobre el nivel del suelo y corresponden a rosas de viento de 8 direcciones.
b) Registros provenientes de la aplicación de un modelo climatológico de mesoescala (Berri et al., 2010) que predice vientos de superficie en una zona del estuario del Río de La Plata.

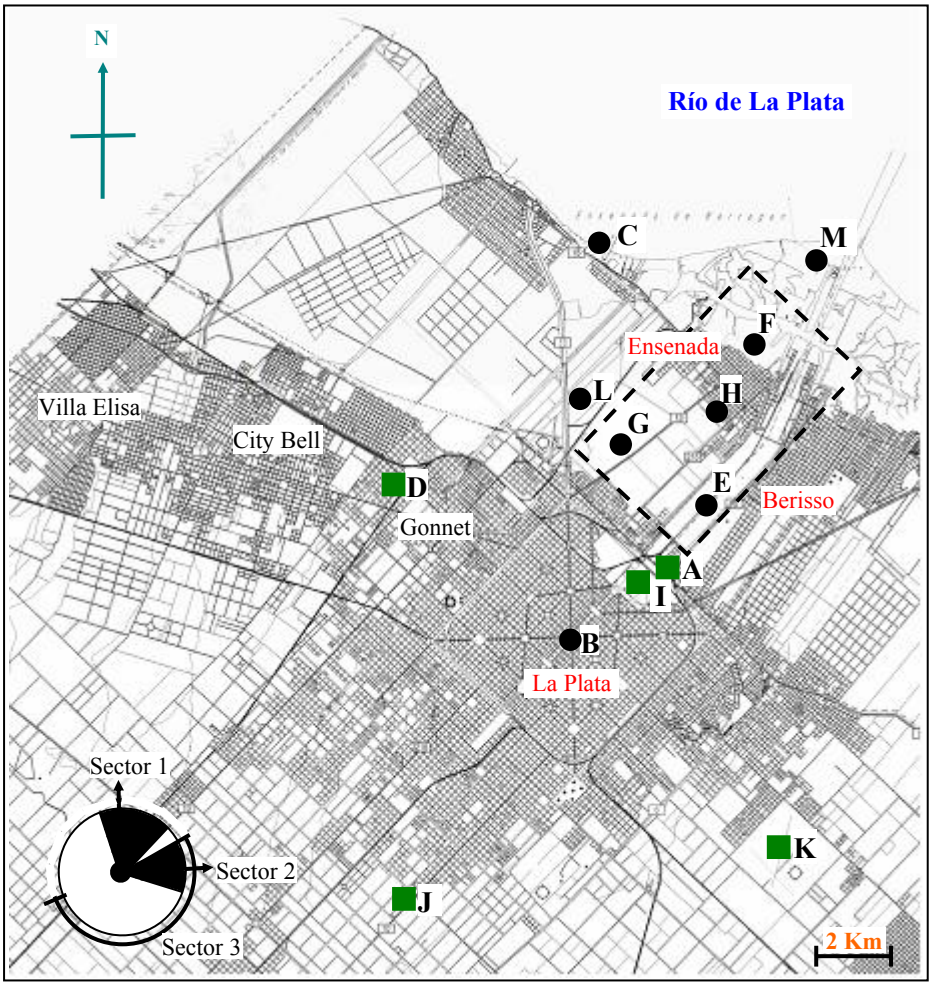


Figura II.6: Mapa de La Plata y Alrededores. Los puntos de medición (vientos y/o dióxido de azufre) se hallan indicados con un cuadrado. Los otros puntos de referencia con un círculo. Punto A: Universidad Tecnológica Nacional- Facultad Regional La Plata. Punto B: centro de la ciudad. Punto C: costa del río. Punto D: CIOp (Centro de Investigaciones Opticas- Gonnet) Punto E: Refinería de Petróleo. Punto F: Astillero. Punto G: Plantas de procesamiento de acero. Punto H: centro del rectángulo indicativo de un área de alta actividad industrial. Punto I: Observatorio de la Facultad de Ciencias Astronómicas y Geofísicas de la Universidad Nacional de La Plata (Paseo del Bosque). Punto J: Estación Agrometeorológica Julio Hirschhorn de la Universidad Nacional de La Plata. Punto K: Aeropuerto de La Plata (designado como LPA en la Figura II.2). Punto L: Central Termoeléctrica. Punto M: Puerto de La Plata. La distancias directas de B a D es aprox. 6.5 km, de D a E aprox. 8.5 km, de B a E aprox. 5 km, de B a J aprox. 8 km y de B a K aprox. 7 km.
El diagrama ubicado en la parte inferior izquierda de la figura indica grupos de direcciones de viento que fueron de particular interés en la tesis a) nornoroeste-norte-nornoreste-noreste (Sector 1) (la flecha indica la dirección del viento proveniente del norte) b) estenoreste-este-estesudeste (Sector 2) (la flecha indica la dirección del viento del este). El Sector 3 cubre de este-noreste a oeste-noroeste en dirección horaria.

c) Registros pertenecientes a La Plata y alrededores (puntos A, D, I, J y K en la Figura II.6). A excepción del Punto K (LPA), el resto de las estaciones meteorológicas no corresponden a sitios oficiales (aquellos pertenecientes o asociados al SMN) y por lo tanto no dan cumplimiento a protocolos tales como los de WMO (1983, 2008) de la OMM

(Organización Meteorológica Mundial). Cabe aclarar que el Punto K no cumple con requisitos tales como los de EPA (2008) que establece requerimientos para la medición de parámetros meteorológicos en relación a la medición de contaminantes. Estas estaciones fueron instaladas respectivamente con distintos objetivos institucionales y debido a esto poseen diferencias en la toma de datos (calidad).

c1) Punto A durante 1997- 2003. Durante todo este período los datos fueron provistos en promedios cada 15 minutos. Las mediciones correspondientes a este sitio estuvieron, para la mayor parte de los registros de los primeros años, afectadas por una deficiencia que afectó las observaciones de la dirección NNE (nor-nor-este) que quedaron algo subestimadas (Ratto et al., 2010a).

c2) Punto D durante 2006- 2007. Los datos fueron registrados en promedios cada 15 minutos.

c3) Punto I durante 1967- 1994: promedios mensuales del período (28 años).

c4) Punto J durante 1997- 2009. Los registros del invierno de 2000 no fueron suficientes debido a desperfectos técnicos. Los datos de esta estación meteorológica corresponden a promedios horarios.

c5) Punto K (Aeropuerto de La Plata o LPA). Los datos se hallan agrupados en dos conjuntos. Uno corresponde a promedios mensuales que cubren 5 décadas (1961- 1970; 1971- 1980; 1981- 1990; 1991- 2000; 2001- 2010); estos grupos fueron tomados de las estadísticas meteorológicas respectivas (SMN, 1971, 1981, 1992, 2001, 2011). Los valores de vientos (direcciones y velocidades) están dados para rosetas de 8 direcciones y fueron medidos a 10 m de altura sobre el nivel del suelo.
El otro conjunto de datos son promedios horarios entre 1995 y 2005 proporcionados en base a un pedido especial al SMN. Los registros corresponden a rosetas de vientos de 16 direcciones y fueron medidos a 10 m de altura sobre el nivel del suelo.

A lo largo de esta tesis los promedios horarios hacen referencia a bloques de horas, por ejemplo, “velocidades de la Hora 0” equivale al promedio de los valores de velocidad registrados durante ese período entre las 00:00 y las 00:59 Hora Local.

II.3.3 Estaciones meteorológicas y unidad analizadora de SO_2

El Punto A (Figura II.6) se halla localizado en un área urbana en Berisso (municipio del Gran La Plata- ver en la figura) y pertenece a la Universidad Tecnológica Nacional. Este sitio operó una estación meteorológica Davis (Davis Instruments, CA) modelo Weather Monitor II Euro Version® (Figura II.7).
El Punto J se halla ubicado en una zona semirural que pertenece a la Estación Agrometeorológica Julio Hirschhorn de la Universidad Nacional de La Plata. Este sitio operó una estación Davis modelo GroWeather Industry®.
Ambos equipos operaron realizando observaciones cada 22.5 grados (obteniéndose rosas de viento de 16 direcciones) con una exactitud de ±7°.El límite de detección de las velocidades es de 1.6 km h^{-1} (velocidades inferiores son contadas como calmas) y la resolución es de 1.6 km h^{-1} en ambos casos.

La estación del Punto A se hallaba instalada a una altura de 12 m mientras que la del Punto J a 5 m sobre el nivel del suelo.

El Punto D contó con la estación meteorológica del Punto A durante el período 2006- 2007 que fue instalada a 12 m de altura.

El Punto I se halla en la Facultad de Ciencias Astronómicas y Geofísicas de la UNLP (Universidad Nacional de La Plata) ubicada en una zona arbolada llamada "Paseo del Bosque". Los datos de trabajo utilizados en esta tesis son los de una estación meteorológica instalada a 40 m de altura sobre el nivel del suelo.

Figura II.7: Fotografía que muestra la estación meteorológica del Punto A (Universidad Tecnológica Nacional). A la izquierda se observa el recinto donde se halla el medidor de humedad y el sensor de temperatura. A la derecha el anemómetro y la veleta de direcciones.

Figura II.8: Unidad Analizadora Lear Siegler ML 9850 utilizada para realizar mediciones de SO_2 en el Punto A y en el Punto D.

La unidad analizadora de SO_2 (dióxido de azufre) es un equipo comercial Lear Siegler® modelo ML 9850 (Figura II.8) que operó en el Punto A y en el Punto D en distintos períodos. Este equipo basa su funcionamiento en la espectroscopia óptica de emisión y método no dispersivo y está diseñada para el monitoreo continuo de SO_2 en aire ambiente. El SO_2 absorbe fuertemente entre los 200 y 240 nm (nm = 1 nanometro = 10^{-9} m). La absorción de fotones en ese rango da lugar a la emisión de fluorescencia en la longitud de onda entre 300 y 400 nm, siendo la cantidad de fluorescencia emitida proporcional a la concentración de SO_2 existente. Como fuente de radiación se utiliza una lámpara de descarga de Zn cuyo haz lumínico pasa por un filtro centrado en 213.9 nm. Esa radiación es enfocada en la celda de fluorescencia donde se produce la interacción entre el SO_2 presente en la celda y la luz incidente. La fluorescencia resultante se colecta y dirige a un fotomultiplicador (sistema de detección), pasando previamente por un filtro centrado en 350 nm. Otro detector (referencia) monitorea la emisión de la lámpara y es utilizado para corregir fluctuaciones temporales en la misma.

La Figura II.9 muestra un esquema simplificado de la unidad analizadora de gases:

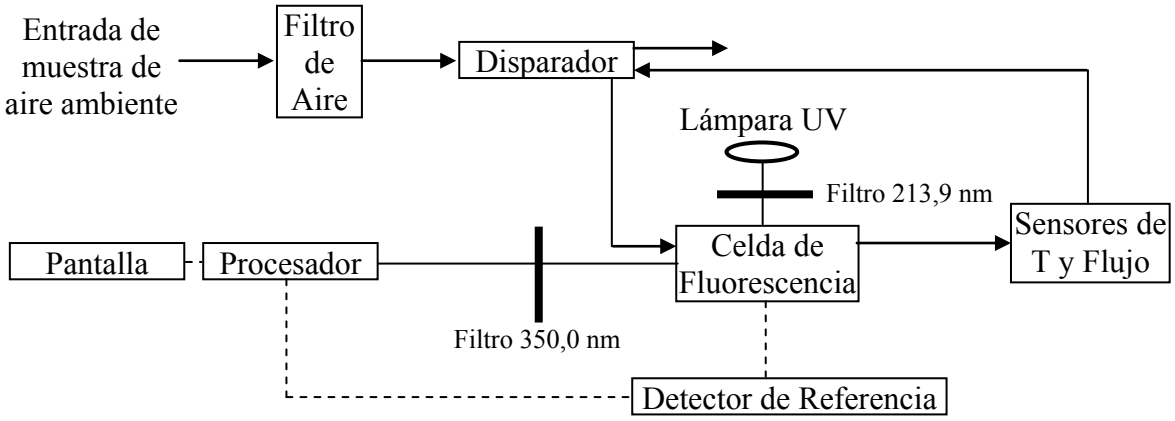


Figura II.9: Esquema simplificado del equipo comercial de monitoreo de SO_2 en valores de calidad de aire. Las líneas que terminan en flecha indican el circuito de la muestra de aire en estudio, las líneas llenas indican el circuito óptico y las líneas a rayas el circuito eléctrico.

La precisión del equipo en la medida de SO_2 es de 0.5 ppbv (partes por billón en volumen= 1 x 10^{-3} ppmv) o 1% de la escala completa. Este equipo opera bajo norma US EPA entre 15ºC y 35ºC de temperatura ambiente pero puede desempeñarse en el rango entre 5ºC y 40ºC.

II.4 Entrenamiento en técnicas espectroscópicas

II.4.1 Generalidades

El diagrama de flujo de la Figura II.10 muestra un sistema de manipuleo de gases diseñado específicamente para operar la cámara de ensayos (CE) que permite montar dispositivos ópticos, sensores y la sonda de muestreo de un equipo electroquímico (Sección II.4.2) para medir concentraciones de gases de emisión característicos de chimeneas.
Las distintas tuberías permiten llenar la CE con gases mezcla a partir de gases puros que son mezclados con N_2 (de tubo) o aire (del compresor) o bien, gases ya mezclados (por ejemplo, tubo con SO_2 y NO_2). De esta manera la instalación hace posible preparar una gran variedad de concentraciones de los gases de interés, que permiten simular situaciones de chimeneas y establecer las curvas de calibración y ajuste del cero del instrumento que se quiere diseñar o chequear.

Los brazos F- D de la CE (ver Figura II.10) indican el montaje de la fuente de luz (F) y el sistema de detección (D). El equipo Testo® 360 (Sección II.4.2) fue utilizado como referencia. Una bomba de vacío permite evacuar la cámara haciendo pasar los gases por un sistema de filtros (Filtro) con capacidad para retener los gases contaminantes. Un manómetro de estado sólido permite conocer el grado de vacío o sobrepresión (procesos de descarga y enjuague) y la presión deseada de operación (medición de gases). La cámara puede ser calefaccionada hasta 350° C permitiendo evaluar la performance del equipo en diseño frente a cambios de temperatura.

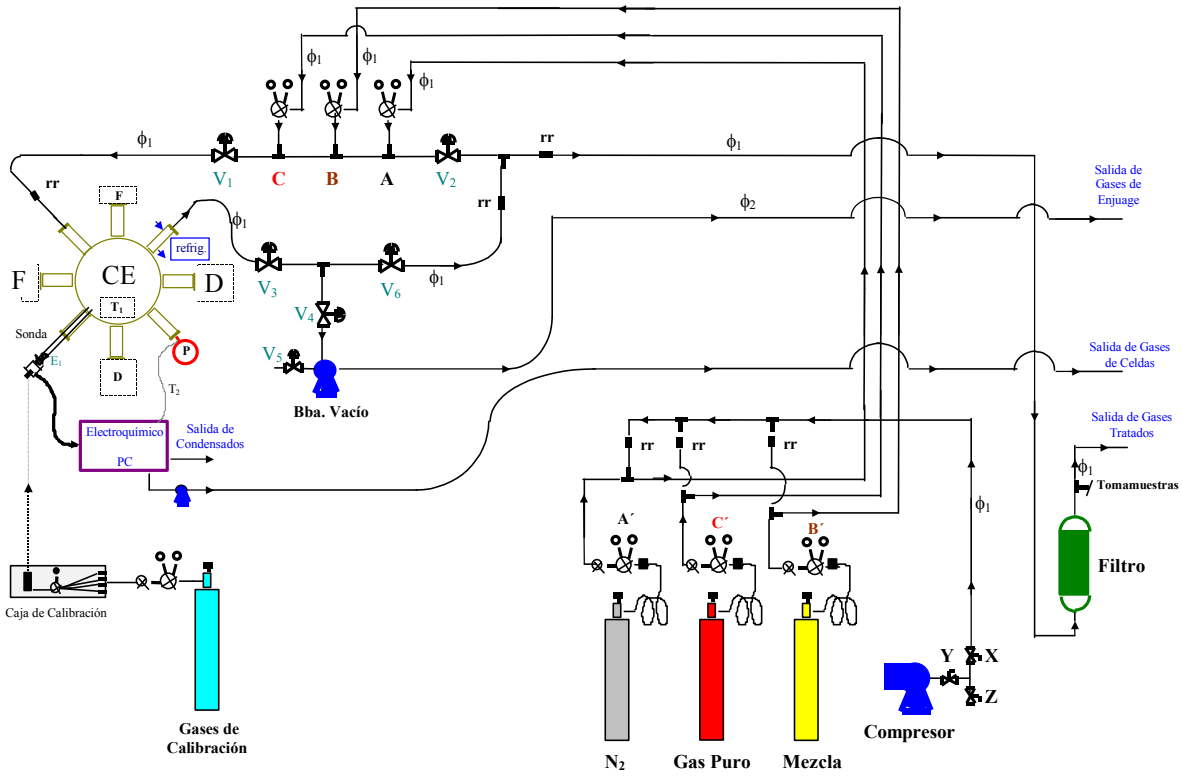


Figura II.10: Esquema simplificado del circuito de gases y cámara de ensayos (CE) en el laboratorio de ensayos del CIOp (Centro de Investigaciones Opticas).

Otras referencias de la Figura II.10:
X,Y,Z válvulas control de aire de compresor
rr : válvulas antiretorno
ϕ_1 : ¼ “ acero inoxidable
ϕ_2 : manguera de ¾”
T_2: Termocupla del medidor electroquímico.
F: Fuente
D: Sistema Detector
CE: Cámara de Ensayos
P: medidor de presión de estado sólido

Válvulas Reguladoras
Todas de cuerpo y diafragma de acero inoxidable 316, excepto la de N_2.
A, B, C : válvulas reguladoras de segunda etapa (300—2-40 PSI)
A´, B´, C´ : válvulas reguladoras de una o dos etapas

Válvulas a Diafragma
V_i: válvulas a diafragma multivuelta con cuerpo de acero inoxidable 316 aptas para regular flujo.

La Figura II.11 muestra el laboratorio donde se hallaba la CE y la mayor parte de los artefactos e instrumentos de la Figura II.10.

Figura II.11: Fotografía del laboratorio de ensayos de contaminantes del CIOp (Centro de Investigaciones Opticas- CIC- CONICET en Gonnet partido de La Plata, Pcia. de Buenos Aires, Argentina).
La cámara de ensayos (color amarillo) se halla en el centro algo hacia la izquierda debajo de la campana extractora de gases ambiente.
Abajo de la mesa, hacia la derecha, puede apreciarse una vista del equipo electroquímico utilizado como referencia.

II.4.2 Equipo de referencia

El equipo comercial Testo® 360 (Figura II.12) de procedencia alemana (certificado por el TÜV- Technical Surveying Institute) es un modelo portable (~ 20 Kg) diseñado para realizar mediciones en rangos de emisión (principalmente chimeneas). El modelo que se operó permite medir O_2, NO, NO_2, CO y SO_2 mediante celdas electroquímicas y CO_2 mediante un sensor infrarrojo. La Tabla II.1 indica el rango de operación y la exactitud de medición del equipo para cada parámetro.

Figura II.12: Equipo electroquímico Testo 360.

Tabla II.1

Gas	Rango	Exactitud en % de final de escala
O_2	0 - 21 %	≤ 1.2 %
NO	0 - 3000 ppmv	≤ 2.8
NO_2	0 - 500 ppmv	≤ 1.0
CO	0 - 5000 ppmv	≤ 2.5
SO_2	0 - 10000 ppmv	≤ 2.0
CO_2	0 - 25 %	≤ 1.5
T	0 - 600 °C	≤ 0.5

Tabla II.1: Rangos operativos y exactitud de la unidad portable Testo 360.

El software de operación del equipo permite ver en pantalla los valores que están siendo analizados (tiempo real). La calibración del mismo se realiza con gases patrones US EPA

(United States Environmental Protection Agency) para cada gas con una unidad de calibración provista por el fabricante.

II.4.3 Equipo diseñado en el CIOp

Se trata de un sistema óptico para monitorear de manera continua y en tiempo real gases de interés industrial (principalmente SO_2 y NO_2) en valores de emisión. Estos gases presentan (entre otras zonas del espectro) absorción en el rango UV (ultravioleta).
Los sistemas no dispersivos son llamados así porque no cuentan con dispositivos que dispersen la luz tal como lo hacen las redes de difracción o los prismas. Los componentes básicos de un sistema no dispersivo son: fuente de radiación, filtros ópticos que permiten seleccionar rangos de longitudes de onda y detectores que recogen la energía lumínica afectada, según el caso, por la presencia de especies gaseosas en el aire. Cuanto mayor sea la presencia de gases contaminantes a medir que se interpongan en el haz de luz entre la fuente y el detector, menor luz alcanzará a este último. O sea, la luz que alcanza al detector guarda una relación con la presencia y concentración del gas en estudio. El sistema de detección incluye la conversión de la señal lumínica en señal eléctrica la que a su vez se convierte en valores de concentración. La Figura II.13 muestra un esquema típico de un equipo no dispersivo utilizado para realizar mediciones continuas en chimeneas.

La transmisión de luz monocromática a través de un gas está caracterizada por la ley de Lambert- Beer:

$$I(\lambda) = I_0(\lambda)\ e^{(-c\ \sigma(\lambda)\ L\ )} \qquad \text{ec. II.1}$$

donde

$I(\lambda)$ irradiancia incidente en el detector [Watt cm^{-2}]

$I_0(\lambda)$ irradiancia emitida por la fuente de luz

c es la concentración del gas que absorbe luz en un determinado rango de longitudes de onda [moléculas cm^{-3}]

L es la distancia que recorre la luz (camino óptico) [cm]

$\sigma(\lambda)$ es la sección eficaz del gas que se quiere medir [cm^2/moléculas]

Si al haz de luz se le interpone un filtro óptico (medio que por absorción o interferencia retiene algunas longitudes de onda y permite transmitir otras) la ecuación II.1 queda:

$$I(\lambda) = I_0(\lambda)\ F(\lambda)\ e^{(-c\ \sigma(\lambda)\ L\ )}$$

donde $F(\lambda)$ es el factor que tiene en cuenta la transmitancia del sistema de filtros.

Para detectar el gas en estudio se debe seleccionar una zona del espectro de absorción en el que dicho gas tenga mucha absorción (esto producirá señales fuertes en el detector). Además, deberá buscarse, dentro de lo posible, que no haya otros gases que absorban a las mismas longitudes de onda. Los filtros pasabanda e interferenciales cumplen la función de acotar las longitudes de onda que llegan al detector para hacer que la única señal que llegue sea debida al gas en estudio.

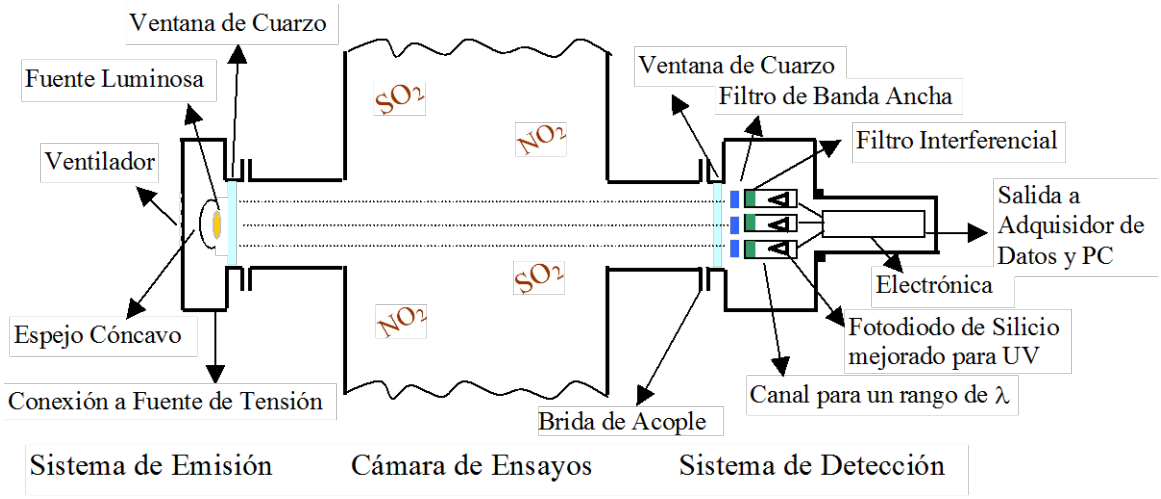


Figura II.13: Esquema de un equipo no dispersivo típico. Este equipo fue montado a la cámara de ensayos de la Figura II.11 para evaluar su performance con distintas concentraciones y mezcla de gases.

La señal lumínica que llega al detector se verá afectada por la eficiencia del mismo que depende de la longitud de onda, $\eta_{(\lambda)}$ [Amperes/Watt]; el fotodetector dará una señal de diferencia de potencial (por ejemplo, en milivoltios) dada por:

$$V_i = \int \Delta_i \ \eta(\lambda) \ F(\lambda) \ I_0(\lambda) \ e^{(-c \ \sigma(\lambda) \ L)} d\lambda$$

donde Δ_i es el ancho del filtro pasabanda (i es un código para identificar un filtro en particular entre los varios posibles).

Para compensar las variaciones de la intensidad lumínica de la fuente o posibles reducciones en la transmisión del ensamblaje óptico (ensuciamiento, etc.) es usual realizar un cociente de señales. O sea, se elige una zona del espectro de absorción que sea cercana a la longitud de onda del pico de absorción del gas en estudio pero cuya absorción sea muy baja. De esta manera se minimiza el cambio de intensidad lumínica de la fuente debido a la las distintas longitudes de onda (pico y referencia).

La Figura II.14 muestra las señales observadas en los fotodetectores (Figura II.13) cuando la cámara de ensayos está llena de gas N_2 (1 atm.). A 300 nm corresponde el pico de absorción del SO_2, a 380 nm corresponde el pico de absorción del NO_2 mientras que 320 nm es la longitud de onda de referencia (ambos gases presentan muy poca absorción a 320 nm).

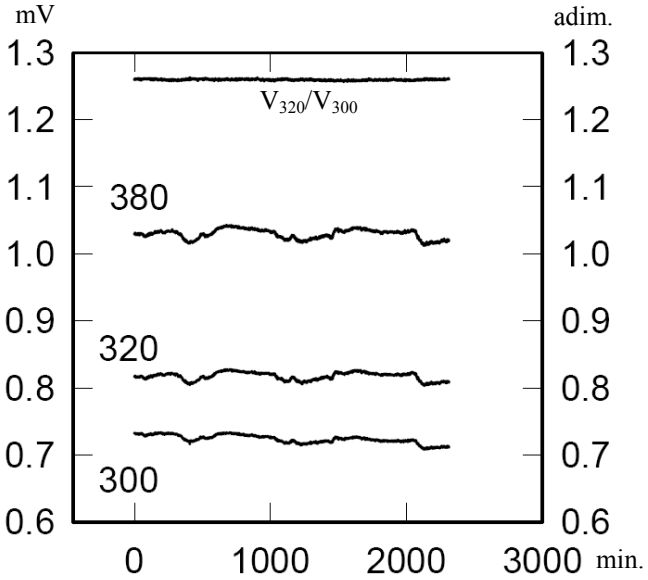


Figura II.14

Figura II.14: Curvas de las señales que producen los tres canales de detección (300 nm, 320 nm y 380 nm) cuando la cámara de ensayos de la Figura II.11 se halla en presencia de N_2 (gas que no absorbe en el rango de trabajo). La curva superior similar a una recta horizontal es el cociente de señales V_{320}/V_{300} que muestra el efecto de atenuación de fluctuaciones respecto de cada canal independiente. El eje de las *X* es el tiempo en minutos. El eje de las *Y* a la izquierda está dado en milivoltios (mV) y el de la derecha es el cociente de señales por lo cual es adimensional.

La Figura II.15 muestra los valores de los cocientes de las señales para distintas concentraciones de los gases de estudio. Estas curvas pueden considerarse como curvas preeliminares de calibración, dado que el equipo en diseño es evaluado con un equipo comercial calibrado según los requerimientos del fabricante (Sección II.4.2).

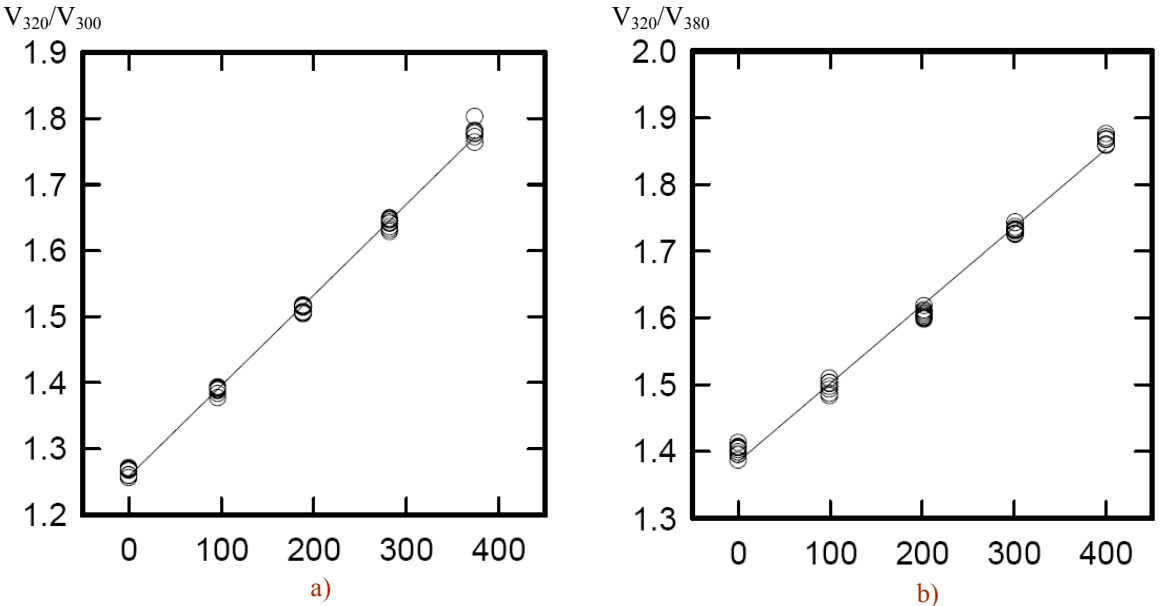


Figura II.15: Cociente de señales en el fotodetector (eje *Y*) versus concentraciones medidas con el equipo Testo 360 en la cámara de ensayos. a) SO_2 en ausencia de NO_2 y b) NO_2 en ausencia de SO_2. La presencia de varias circunferencias para cada concentración (con un paso de 100 ppmv) se debe a que para cada concentración de referencia se realizaron replicados.

Puesto que el canal de 320 nm es una referencia para la medición de ambos gases, es necesario considerar como es la performance del equipo cuando ambos gases se hallan presentes simultáneamente. Este tema se halla delineado en Videla et al. (2006) y en Ratto et al. (2007).

II.4.4 DOAS (Diferential Optical Absorption Spectroscopy)

A diferencia de los no dispersivos los equipos que trabajan con el método DOAS (espectroscopia óptica de absorción diferencial) están basados en el análisis de los espectros de luz (descomposición espectral). Este tipo de equipo es apto para medir gases ambientales en bajas concentraciones (Platz et al., 1979; Edner et al., 1993; Sigrist, 1994; Platz y Stutz, 2008); un prototipo se hallaba diseñado y construido en el CIOp en etapa experimental (Rosato y Reyna Almandos, 1996). Una variante del mismo se utilizó para realizar mediciones de prueba. Un esquema sencillo se presenta en la Figura II.16 (Reyna Almandos et al., 2007).

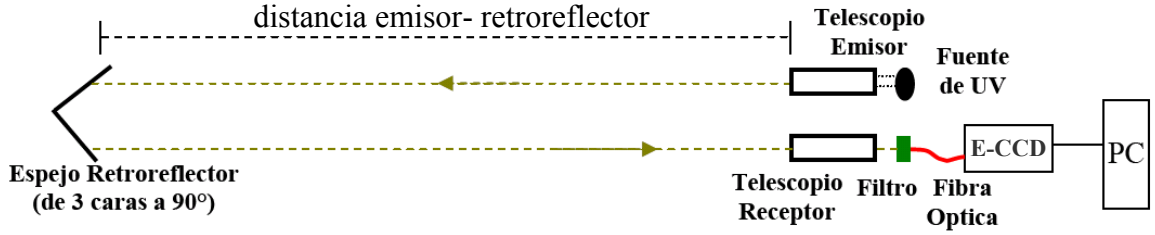


Figura II.16: Esquema alternativo de montaje de DOAS para detectar contaminantes del aire ambiente. E-CCD designa: espectrógrafo acoplado con un detector CCD ("coupled capacitor device").

La distancia entre el emisor y el retroreflector involucra una columna abierta de aire ambiente en donde se hallan las especies que se desean medir. A mayor longitud de esta columna mayor será la sensibilidad de detección posibilitando medir concentraciones bajas. La fuente de luz ultravioleta es una lámpara de arco de xenón de amplio espectro que como óptica de colimación lleva un telescopio. Para las pruebas de NO_2 se utilizó una lámpara halógena en la zona del espectro visible. Un espectrógrafo con fotodetector incorporado permitió analizar el espectro de luz que recorrió la columna de aire.

En la Figura II.17 se muestran por separado algunos de los componentes del montaje de la Figura II.16.

a) b) c)

Figura II.17: a) Espejo retroreflector (tipo "ojo de gato") b) conjunto de espectrógrafo y cámara CCD c) Telescopio emisor (grande) y telescopio receptor (pequeño).

Las pruebas con este montaje se realizaron con el haz de luz atravesando el camino Centenario (Figura II.18) a la altura del Punto D (Figura II.6). Las mismas sirvieron para evaluar las necesidades de puesta a punto del equipo las cuales incluyen: mejorar el sistema de alineación del telescopio emisor, el espejo retrorreflector y el telescopio emisor; operar a mayor distancia para bajar el límite de detección de los gases de interés (principalmente del NO_2); instalar fuentes lumínicas de más intensidad en el rango de

longitudes de onda de trabajo y realizar una reprogramación del software de adquisición de datos para que el sistema automático de registros sea más confiable.

a) b)

Figura II.18: a) Línea amarilla que indica la trayectoria de la luz desde el dispositivo de emisión a la derecha hasta el espejo retroreflector ubicado en el otro extremo (izquierda) y cubre aprox. 340 m. La zona sin edificación pertenece al predio donde se halla ubicado el Centro de Investigaciones Opticas en Gonnet. b) Vista del haz de luz hacia el espejo retroreflector y proveniente del mismo durante la noche.

"El verdadero éxito del descubrimiento no reside en encontrar nuevos territorios, sino en verlos con ojos nuevos"
Marcel Proust

"If you can't explain it simply, you don't understand it well enough"
Albert Einstein

Capítulo III
Fenómenos físicos

En este capítulo se agrupan y describen los principales conceptos y fenómenos físicos que sirven de contexto a la discusión presentada en los capítulos IV y V.

III.1. Atmósfera

La palabra atmósfera ("atmos": vapor; "spaire": esfera, globo) hace referencia a la envoltura gaseosa que recubre la superficie terrestre. Junto a la geosfera (esfera de la tierra sólida), la hidrosfera (esfera de agua) y la biosfera (esfera de la vida) es una de las cuatro "esferas" en que se divide el planeta tierra para su estudio (Lutgens y Tarbuck, 2013). Entre ellas existe un alto grado de interacción, tal como lo manifiestan La Oscilación Sur y El Niño y La Niña en relación al acoplamiento mar- atmósfera.
La composición y estructura global de la atmósfera "actual" data de aproximadamente 400 millones de años (Jacobson, 2002). Esta envoltura, que permanece cerca de la superficie debido a la fuerza de la gravedad, participa de los movimientos de la Tierra los cuales le confieren un espesor mayor en el ecuador que en los polos (Lazaridis, 2011). No es posible definir estrictamente su espesor (o altura) porque las densidades de las porciones más altas (aquellas más alejadas de la superficie terrestre) son muy bajas y es difícil distinguir entre atmósfera y espacio exterior (interestelar). Comparada con el radio de la tierra el espesor de la atmósfera es muy pequeño (como el espesor de una hoja de papel frente al radio de una pelota de tenis). Dependiendo de los fenómenos observados, la altura de la atmósfera medida desde el suelo puede llegar hasta 80- 250 km y, en algunos casos, suele considerársela de hasta varios miles de kilómetros (Lazaridis, 2011). Sin embargo, el 99% de las especies químicas de la atmósfera se hallan en los primeros 30 km (Arhens, 2009).
En términos generales la atmósfera suele dividirse en dos regiones: atmósfera baja y atmósfera alta. La atmósfera baja comienza a nivel del suelo; su límite superior suele asignarse, según los fenómenos que se tengan en cuenta para el estudio, al tope de la tropopausa (alrededor de los 15- 20 km de altura) (Finlayson- Pitts y Finlayson- Pitts, 2000) o bien en el tope de la estratosfera (alrededor de los 50 km de altura) (Seinfeld y Pandis, 2006). Por su parte, la atmósfera alta puede involucrar todas las capas sucesivas o solo alguna de ellas (principalmente la estratosfera).

La Figura III.1 (modificada de Lazaridis (2011)) muestra las distintas capas en que queda dividida la atmósfera según el perfil de temperatura (curva en verde) llegando hasta algo más de 100 km desde el suelo. Según este criterio la homosfera (capa de composición homogénea) comprende a la troposfera, a la estratosfera y a la mesosfera mientras que la heterosfera (llamada así por contener distintas especies estratificadas según el peso molecular) que parte aprox. de los 90 km, comprende a la termosfera (aprox. desde los 90 km hasta aprox. los 400 km) y a la exosfera (que parte de los 400- 500 km) que tiene sus confines en donde los átomos pueden escapar del campo gravitatorio terrestre pasando a formar parte del espacio exterior. Las zonas intermedias se denominan "pausas" y hacen

referencia a valores relativamente constantes de temperatura respecto de la capa que le da su nombre.

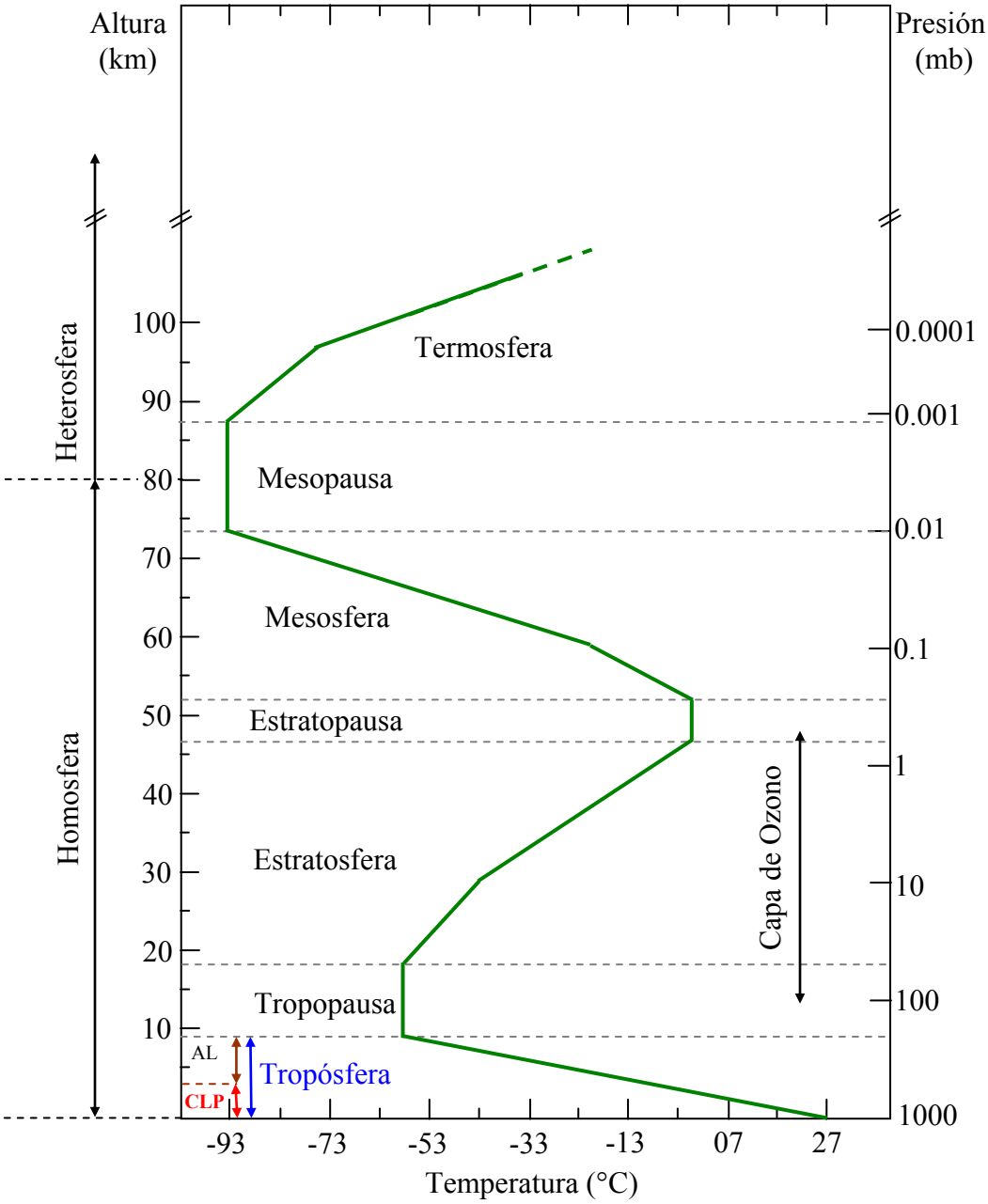


Figura III.1: Estructura vertical de la atmósfera basada principalmente en el perfil de temperatura (curva verde). Dentro de la troposfera están indicadas la Capa Límite Planetaria (CLP) y la Atmósfera Libre (AL).

Superpuestas a las zonas descriptas se hallan otras zonas de interés para su estudio: la ionosfera (que tiene su límite inferior en los 60 km y el superior en los 300 km) posee partículas cargadas debido a la radiación solar mientras que la magnetosfera (va desde aprox. los 1000 km hasta unas diez veces el radio terrestre) es la zona en donde los iones son influenciados por el campo magnético terrestre. La Figura III.1 muestra también en el eje *Y* de la derecha los valores de la presión (1 atm. = 1013,25 mb (milibares), 1mb = 1 HPa (hectopascal)) en escala logarítmica, parámetro muy ligado a los cambios de densidad característicos de la atmósfera. El descenso de la presión con la altura establece un

gradiente que se opone a la atracción gravitatoria evitando el colapso de las moléculas sobre la superficie de la tierra. Cada una de las capas así definidas son de interés para su estudio debido a los distintos procesos físicos y químicos que tienen lugar en ellas. Al mismo tiempo debe considerarse que todas ellas se hallan interconectadas (Seinfeld y Pandis, 2006). Estas capas, que solamente se citan a modo de referencia, por los temas que se tratan en otras secciones, constituyen el objeto de estudio de varias disciplinas (navegación aérea, comunicaciones, etc.). En el contexto de esta tesis es la troposfera, principalmente en su porción más cercana a la superficie de la tierra, la de mayor interés. La misma tiene como límite inferior a la superficie terrestre (tierra o agua) que si bien es muy definida presenta grandes variaciones espaciales; su altura (o espesor) llega en promedio a los 12 ± 4 km (siendo en los polos de alrededor de 8 km y en el ecuador de alrededor de 16 km) y contiene cerca del 80% de la masa total de la atmósfera. Es la más importante desde el punto de vista de los fenómenos meteorológicos (Tiempo y Clima- Sección III.2).

Para comprender las dinámicas básicas de la troposfera es necesario considerar el sistema sol- atmósfera- tierra (Arya, 2001). El sol es la fuente de energía que atraviesa todas las capas atmosféricas (filtros) llegando a la superficie. Puesto que la capacidad del aire para absorber calor es muy baja, la radiación es absorbida en su mayor parte por la superficie que se calentará y emitirá calor al aire circundante. Este efecto se aminora a medida que nos alejamos de la superficie lo cual da en promedio (observacional) un descenso de temperatura de aprox. 6.5 °C por cada km de ascenso por la troposfera (primera porción de la curva de la Figura III.1).

III.2 Meteorología y climatología

La tierra y la atmósfera constituyen un sistema dinámico en constante cambio. Los cambios en la superficie de la tierra solo son observables en grandes escalas de tiempo mientras que, en la atmósfera, los cambios pueden llevarse a cabo en pocos minutos.

La meteorología (del griego “meteoros” = suspendido en el aire) es el estudio de la atmósfera y sus fenómenos (Arhens, 2009). Constituye una de las ciencias de la tierra (entre otras tales como la geología, la oceanografía, etc.) y está en relación con ellas, a veces, superponiéndose. Tiempo (estado de la atmósfera en un tiempo y lugar dados) y Clima (la acumulación de eventos -promedios y extremos- del Tiempo durante largos períodos tales que permiten describir una región) son dos palabras muy ligadas a esta disciplina. En ambos casos son de interés la detección de patrones en los procesos que tienen lugar en la zona cercana a la superficie terrestre.

III.3 Circulaciones atmosféricas

Las circulaciones atmosféricas que cubren el planeta se hallan organizadas según una jerarquía con alto grado de integración: las mismas van desde simples rachas (ráfagas de viento) hasta tormentas que cubren miles de kilómetros. Es común definir escalas espaciales de movimiento para facilitar el estudio de fenómenos específicos; asociadas a estas se hallan las escalas temporales. La Figura III.2 (tomada de Arhens (2009)) muestra un posible esquema de fenómenos que se hallan espacial y temporalmente asociados. Cuando en el Capítulo V se haga referencia a un “modelo de mesoescala” el lector podrá observar que se trata de distancias de cientos de kilómetros, esto se halla en consonancia con la Figura III.2 dado que la misma provee “promedios” indicativos. Más detalles sobre escalas y sus fenómenos asociados se describen en Necco (1980).

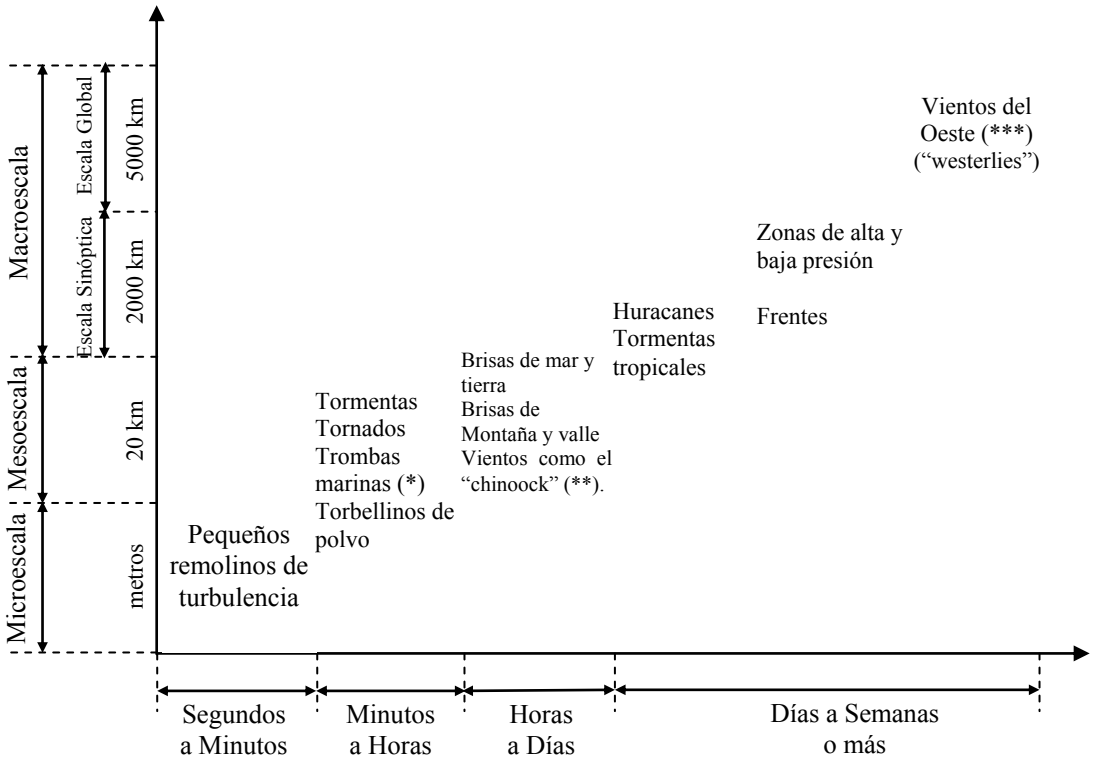


Figura III.2: Escala idealizada de movimientos de la atmósfera. El eje de las *X* indica la duración del fenómeno (que se ha colocado a manera de ejemplo). El eje de las *Y* indica la extensión probable que alcance el fenómeno atmosférico (las magnitudes son solo indicativas).
(*) Las trombas marinas (llamadas también mangas de agua) consisten en un intenso vórtice o torbellino que ocurre sobre un cuerpo de agua, usualmente conectado a una nube cumuliforme.
(**) Este viento que se da en las Rocallosas durante los meses de invierno, es un fenómeno único que puede aumentar las temperaturas más de 20 grados centígrados en un día.
(***) Las "westerlies" son circulaciones de viento en altura que ocurren en las latitudes medias de oeste a este en el hemisferio norte.

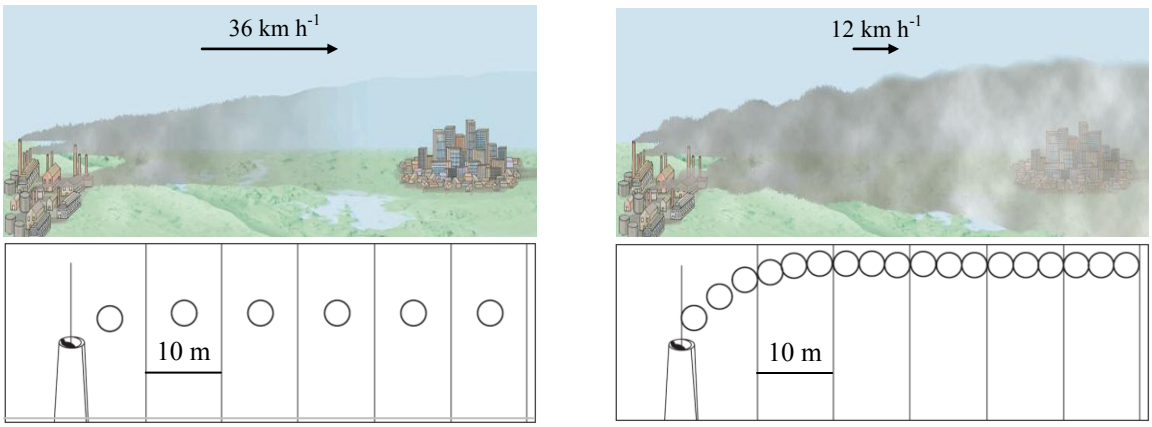
III.4 Viento

La atmósfera actual (observada a nivel del mar) contiene globalmente una mezcla de especies químicas abundantes (principalmente N_2 (≈78%) y O_2 (≈21%) en base seca, además Ar y vapor de agua) y un conjunto complejo y variable de gases traza (usualmente en concentraciones menores a 1 ppmv -parte por millón en volumen-) cuya cantidad total es menor al 1% (Seinfeld y Pandis, 2006)). Ese conjunto es denominado aire aún cuando, como en el caso de las ciudades, participen una gran variedad de especies distintas a las mencionadas arriba.

El viento es aire en movimiento horizontal (Arhens, 2009) respecto de la superficie terrestre que se produce en virtud de la diferencia de presión existente entre dos zonas. En la presente tesis el viento de superficie es el de mayor interés. Existen además otros tipos de desplazamientos del aire tales como los remolinos y las ondas que en general ocurren simultáneamente (Stull, 1988). El movimiento del aire es afectado además por otras fuerzas tales como la de Coriolis (Lazaridis, 2011). Después de las precipitaciones (en particular las lluvias), el viento es el parámetro climatológico que más puede afectar el medio físico, ya que favorece la pérdida de suelos y el arrastre de partículas. El viento, la lluvia y el

nivel de estratificación atmosférica característicos de un lugar son los parámetros meteorológicos más significativos en el destino de los contaminantes (Lesniok, 2011).
Los contaminantes son transportados horizontalmente por un flujo medio mientras que son dispersados lateral y verticalmente por perturbaciones (turbulencia).

Una manera en que la velocidad horizontal del viento influencia la concentración observada de los contaminantes se ejemplifica en la Figura III.3.



a) b)

Figura III.3: Efecto de la velocidad horizontal del viento en la dilución de los contaminantes. Las partes superiores correspondientes fueron tomadas de Lutgen y Tarbuck (2013) mientras que las inferiores de Vallero (2008). Ambas representaciones permiten comparar el efecto de dilución cuando la velocidad se triplica. Por ejemplo, el viento en a) es de 36 km h^{-1} mientras que en b) es de 12 km h^{-1}.
Las "esferas" mostradas en la parte de abajo de cada figura muestran las "unidades de masa" de aire contaminado en la unidad de longitud que se desplazan según la velocidad del viento. Es apreciable como una velocidad relativa más baja (del orden de tres veces tal como lo muestra la parte b) de la figura) induce mayor acumulación de contaminantes con la consecuente reducción de la visibilidad.

Supongamos que la emisión de contaminantes de las chimeneas puede ser discretizada en nubes o "esferas" de emisión. Cuando la velocidad del viento es de 36 km h^{-1} (10 m/s) se observarán esferas saliendo a razón de una por segundo y estarán separadas entre sí 10 m (el panorama se puede observar en la Figura III.3a). Si la velocidad disminuye a 12 km h^{-1} en el transcurso de 1 segundo saldrán tres esferas (Figura III.3b). En este caso la densidad de contaminantes será mayor. Por otra parte, es importante considerar que cuanto mayor sea la velocidad del viento habrá más probabilidad de que se establezca flujo turbulento y, por lo tanto, habrá mayor dispersión de los contaminantes.

El viento es invisible pero ofrece evidencias muy variadas de su presencia dependiendo de su velocidad. La escala Beaufort (Tabla III.1) permite identificar velocidades de manera cualitativa. Una versión de la misma se presenta a continuación con el solo hecho de hacer posible asociar observaciones realizadas en las zonas de estudio con la influencia potencial del viento.

Tabla III.1: Escala Beaufort (tierra) tomada de Arhens (2009).

Número de Beaufort	Descripción	Velocidad (km h^{-1})	Observaciones
0	Calma	0- 2	Elevación vertical del humo
1	Ventolina	2- 6	El humo se dispersa pero las veletas no giran
2	Brisa suave	7- 11	El viento se siente en la cara, las hojas crujen, las veletas se mueven, las banderas se agitan
3	Brisa leve	12- 19	Las hojas y pequeñas ramas se mueven, el viento extiende (aplana) a una bandera liviana
4	Brisa moderada	20- 29	El viento levanta polvo y tira los papeles, pequeñas ramas se mueven, las banderas hacen ondas
5	Brisa fresca	30- 39	Los árboles con pequeñas hojas comienzan a balancearse, las banderas hacen ondas
6	Brisa fuerte	40- 50	Las ramas de los grandes árboles comienzan a mecerse
7	Viento muy fuerte (alto)	51- 61	Los árboles enteros se mueven, dificultad para caminar en contra del viento, se extienden las banderas
8	Temporal	62- 74	El viento rompe la rama de los árboles, se dificulta la marcha.
9	Temporal fuerte	75- 87	Daños estructurales ligeros en las construcciones edilicias (vuelan las antenas, algunos letreros, etc.)
10	Temporal completo	88- 101	Árboles desenraizados, ocurren daños considerables
11	Tormenta	102- 119	Los vientos producen daños esparcidos
12	Huracán	$\geq$ 120	Los vientos producen daños masivos

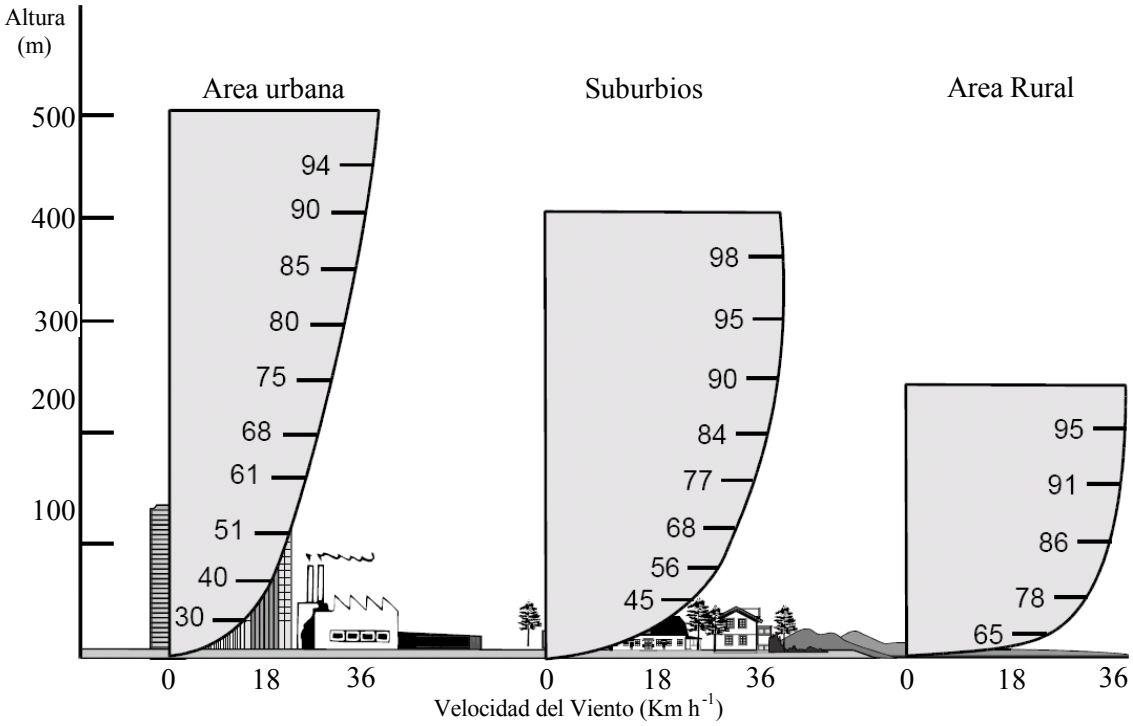
Tabla III.1: Esta escala creada por el almirante irlandés Sir F. Beaufort hacia 1805. Luego fue modificándose según las aplicaciones y tecnologías. La versión que se presenta en esta tabla contiene las observaciones específicas para ser utilizada sobre tierra pero existen las observaciones para cuando es utilizada a nivel del mar. En la actualidad se han agregado más números de Beaufort llegando hasta el 17.

III.5 Fricción y turbulencia

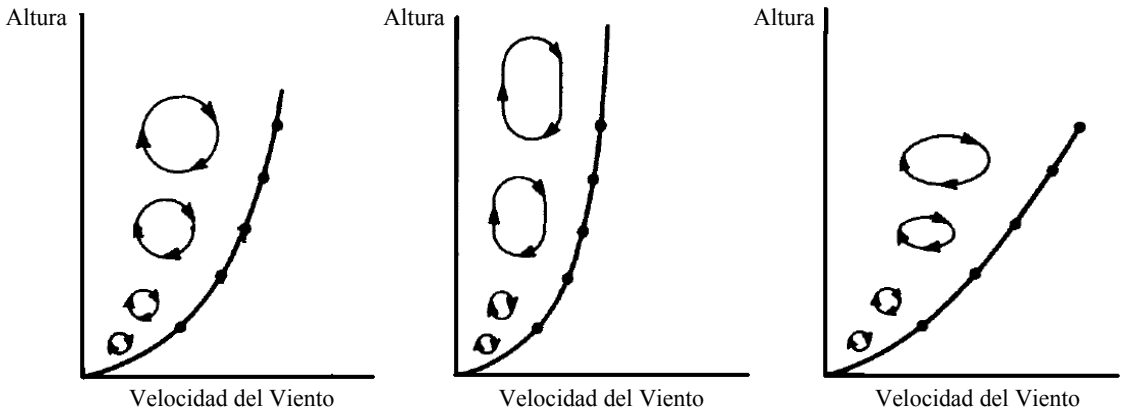
Los remolinos (movimientos erráticos de aire) característicos de la CLP -capa límite planetaria- (indicada en la Figura III.1 y descripta en la Sección III.9) crean fricción con otras porciones del aire que se hallan a mayor velocidad produciendo una disminución de la velocidad media. A este tipo de fricción se la llama fricción viscosa y refiere al "roce" que tiene lugar a nivel molecular. La turbulencia inherente se la suele llamar *turbulencia viscosa* o molecular. Cuando el aire circula en presencia de obstáculos físicos (topografía del terreno, animales, construcciones, etc.) se pueden producir un conjunto variado de remolinos (en tamaño y energía) con velocidades y direcciones que cambian rápidamente dando lugar a ráfagas (o rachas) de viento. Este tipo de turbulencia, que puede tener un desarrollo vertical de varios cientos de metros se conoce como *turbulencia mecánica* y produce, por fricción sobre la superficie del terreno, un arrastre que va decreciendo a medida que aumenta la altura desde el suelo permitiendo que los vientos incrementen su velocidad (se genera un perfil- Figura III.4). El calentamiento de la superficie terrestre da lugar a la creación de corrientes térmicas que producen celdas de convección que pueden entenderse como remolinos creados por diferencias de temperatura. La turbulencia asociada se llama *turbulencia térmica* y tiene su mayor desarrollo en la CLP cuando hay inestabilidad atmósférica (Sección III.7) puesto que esta le permite desarrollarse. La turbulencia mecánica y la térmica son tan importantes que participan de la definición de CLP de una manera más o menos explícita según los distintos autores (Sección III.9). En frecuente hallar en la naturaleza la ocurrencia simultánea de los tres tipos de turbulencia citados.

III.6 Rugosidad

La presencia de obstáculos a nivel de la superficie del terreno puede conceptualizarse como rugosidad. La ubicación, el tamaño y la densidad de los obstáculos dan lugar a distintos gradientes de velocidad de viento, tal como lo muestra la Figura III.4a tomada de EPA (2014). En esta figura el espesor de la capa atmosférica, donde influye la rugosidad del terreno, pasa de 500 a 240 m, esto se produce en virtud de las distintas características de la superficie.



a)



b1) b2) b3)

b)

Figura III.4: a) Perfiles de velocidad horizontal de viento según la rugosidad del terreno. La velocidad máxima se corresponde para cada caso con el viento gradiente (un viento de velocidad constante que sopla paralelo a isobaras curvas) que tiene lugar en el límite de la CLP. Las escalas sobre los perfiles representan porcentajes de velocidad respecto del viento gradiente. En el eje de las *X* se ha puesto con fines comparativos un límite de 36 km h^{-1} como tope.
b) Perfiles de viento con la altura según tres casos característicos de estabilidades atmosféricas (adaptada de Oke (1987)) b1) Neutra b2) Inestable y b3) Estable.

Cada uno de los perfiles de la Figura III.4a puede desarrollarse en distintas condiciones de estabilidad atmosférica (ver definiciones en la Sección III.7) y tienen asociados efectos sobre la forma de los remolinos (turbulencia) según se muestra en la Figura III.4b.

La relación de la velocidad del viento con la altura es compleja (Wark et al. 1998). Es importante contar con un modelo que permita realizar correcciones por altura de tal manera de hacer comparables datos de distintas fuentes. Existen aproximaciones empíricas de gran vigencia tales como la "ley logarítmica" y la "ley de la potencia"; una discusión detallada de ambas leyes se da en Emeis (2012). Siguiendo a Wark et al. (1998) y a Vallero (2008) y dada su simplicidad se optó por aplicar la ley de la potencia. Dicha ley, válida para alturas de pocos cientos de metros, se puede expresar así:

$$u_{(z)} = u_{(h_r)} \left(\frac{z}{h_r} \right)^p \qquad \text{ec. III.1}$$

donde

$u_{(z)}$ es la velocidad del viento "corregida" a la altura z.

El exponente p está dado según la rugosidad del terreno y la estabilidad atmosférica dominante (Sección III.7).

$u_{(h_r)}$ es la velocidad del viento observada a una altura h_r.

z es la altura a la que se desea obtener la velocidad corregida.

h_r es la altura a la que se midió la velocidad observada.

El exponente p aumenta a medida que la rugosidad de la superficie aumenta y el grado de estabilidad aumenta. Los valores de p suelen estar entre 0.07 y 0.60. Se reproduce a continuación un tabla del Capítulo 3 de Wark et al. (1998) para ser utilizada junto a la ec. III.1.

Tabla III.2

Categoría de Estabilidad	Zona Rural	Zona Urbana
A: extremadamente inestable	0.07	0.15
B: moderadamente inestable	0.07	0.15
C: ligeramente inestable	0.10	0.20
D: neutra	0.15	0.25
E: ligeramente estable	0.55	0.30
F: moderadamente estable	0.55	0.30

Tabla III.2: Valores de p para la ec. III.1. La categoría de la estabilidad atmosférica (dada por una letra mayúscula) y la zona permiten elegir un exponente para la ecuación de corrección de velocidad de viento por altura. La Tabla III.3 en la Sección III.7 contribuye a complementar información para la aplicación de la ec. III.1.

Existen recomendaciones (WMO, 2008) para la instalación de instrumentos de medición de velocidad del viento. La altura estándar recomendada para el anemómetro en terreno abierto es de 10 m. Se debe considerar que las mediciones de velocidad y de direcciones (estas últimas varían mucho menos con la altura que las velocidades) deberían ser representativas de varios kilómetros a la redonda, las mismas deben realizarse de tal manera que haya que hacer la menor cantidad de correcciones posibles.

III.7 Estabilidad atmosférica y tipos de inversión

Se ha visto, al principio de este capítulo, que la temperatura de la troposfera disminuye con la altura. Se define una curva idealizada llamada adiabática seca (adiabática porque no intercambia calor y seca porque no tiene en cuenta a la humedad atmosférica) que posee una pendiente (negativa) de alrededor de 9.8 °C km^{-1} (Seinfeld y Pandis, 2006) -notar la diferencia con la de 6.5 °C km^{-1} de la curva observacional que tiene en cuenta a la humedad (Sección III.1).

Para determinar el grado de estabilidad de la atmósfera se recurre a comparar la temperatura de una porción de aire (parcela hipotética) con el aire circundante.

El perfil de temperatura dado por la adiabática seca define la condición de atmósfera neutra y suele designarse con la letra griega ***Γ*** (gamma mayúscula) (ver Figura III.5a).

En condiciones de neutralidad una pequeña perturbación desplazará a la parcela de aire (representada por un balón en la Figura III.5a) a través de sucesivos estados de equilibrio con el medio hasta finalmente detenerse. La temperatura de la parcela de aire es en todo momento la misma que la del aire circundante.

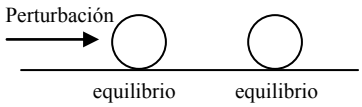


Figura III.5a: Analogía que representa la atmósfera neutra en correspondencia con la Figura III.6a).

Este estado idealizado no es muy frecuente pero se da en presencia de cielos cubiertos (inhibición de calentamiento y enfriamiento radiativo) y vientos moderados a altos (que amortiguan las desviaciones del perfil adiabático).

En la Figura III.5b, la perturbación produce el desplazamiento del balón desde un estado de equilibrio hasta un nuevo estado de equilibrio en donde vuelve a darse la condición de igual temperatura entre la parcela y el aire circundante.

Cuando el perfil de temperatura es más pronunciado que el dado por la adiabática seca se habla de atmósfera inestable (Figura III.5b): una parcela de aire más caliente que el aire que la circunda tenderá a subir (pues tendrá menor densidad) hasta un nivel en que se igualen las temperaturas.

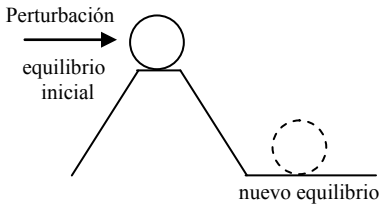


Figura III.5b: Analogía que representa la atmósfera inestable, en correspondencia con la Figura III.6b).

Algo análogo ocurrirá con una parcela de aire más frío que el circundante (más denso) que tenderá a bajar hasta que se alcance el nuevo equilibrio. Es característico de este tipo de atmósfera la presencia de corrientes de aire verticales fuertes.

Si el perfil de temperatura es menos pronunciado que el de la adiabática seca se habla de atmósfera estable. La parcela de aire no tenderá ni a subir ni a bajar y ante una perturbación pequeña volverá a su estado de equilibrio como la esfera de la Figura III.5c (ver también Figura III.6c).

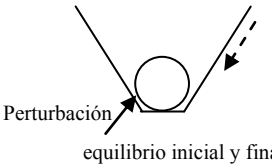


Figura III.5c: Analogía que representa la atmósfera estable en correspondencia con la Figura III.6c).

Si bien lo descripto admite mucho más nivel de detalle (Boeker y Grondelle, 1995; Arhens, 2009) posibilita conceptualizar a la estabilidad atmosférica como a aquella propiedad que adquiere la atmósfera que define la tendencia de una porción de aire a mantenerse en una posición o a moverse verticalmente (Lutgens y Tarbuck, 2013). La estabilidad es una de las características más importantes de la atmósfera en relación al transporte y dispersión de los contaminantes del aire.

El conjunto de gráficos de la Figura III.6 (tomado del Capítulo 3 de Wark et al. (1998)) ilustra de manera sencilla los perfiles de temperatura según los tipos de estabilidad atmosférica.

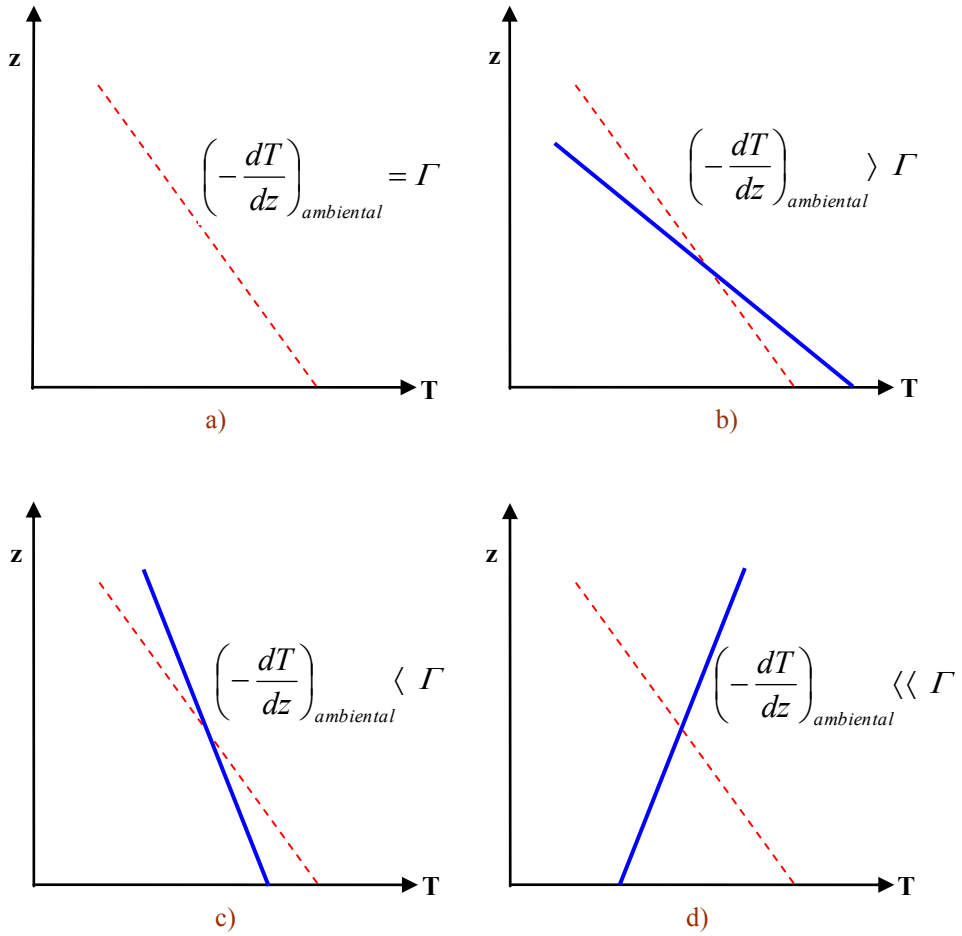


Figura III.6 Perfiles atmosféricos de temperatura.
a) Atmósfera Neutra b) Atmósfera Inestable c) Atmósfera Estable débil y d) Atmósfera Estable fuerte. La curva a rayas en rojo representa la adiabática seca mientras que la curva en azul representa los distintos casos que puede tener el perfil de temperatura real del ambiente.

A lo expuesto antes se le agrega el caso de la Figura III.6d en que la pendiente de la curva ambiental se hace positiva indicando una estabilidad fuerte. Esta inversión térmica resulta ser de gran interés en relación al destino de los contaminantes que se liberan al aire por las distintas fuentes; dado que reduce la dispersión vertical tendiendo a mantener las concentraciones en valores altos.

La clasificación dada de estabilidades atmosféricas puede enriquecerse al considerar la presencia de contaminantes. Pasquill estableció un sistema basado en seis categorías, cada una con un potencial distinto para la dispersión de los contaminantes (McCormick, 1968). La clasificación dada por Pasquill modificada por Turner es de uso práctico y se muestra en la Tabla III.3 (notar que las letras mayúsculas de la A a la D conservan las descripciones dadas en la Tabla III.2).
Para situaciones en que el cielo se halla cubierto durante la noche o el día, la estabilidad se considera neutra independientemente de la velocidad.
Una descripción más detallada del uso de estas tablas se halla en el Capítulo 3 de Wark et al. (1998).

Tabla III.3: Claves para la determinación de la Estabilidad Atmosférica según Turner.

Viento de superficie a 10 m (km h^{-1})	Día			Noche	
	Radiación solar entrante (*)			Fracción de cobertura por nubes	
	Fuerte	Moderada	Débil	cubierto o $\geq$ 50%	despejado o $\leq$3/8
< 7.2	A	A- B	B	-	-
7.2- 10.8	A- B	B	C	E	F
10.8- 18.0	B	B- C	C	D	E
18.0- 21.6	C	C- D	D	D	D
> 21.6	C	D	D	D	D

(*) Fuerte equivale a más de 700 W m^{-2}. Débil equivale a menos de 350 W m^{-2}.
A: extremadamente inestable; B: moderadamente inestable; C: ligeramente inestable; D: neutra; E: ligeramente estable; F: moderadamente estable.

Existen varios tipos de inversión térmica en la CLP. Las dos más importantes son la que se producen por enfriamiento radiativo y la que se produce por subsidencia.

Inversión térmica por enfriamiento radiativo
Durante el día el calentamiento de la tierra hace que las capas de aire por encima de ella se vayan calentando por conducción, convección y radiación desarrollándose un perfil negativo de temperatura. Por la noche, con cielo despejado y poco viento, el enfriamiento de la tierra tenderá a producirse rápidamente (debido a su baja capacidad calorífica) enfriando a las capas de aire que están por encima de ellas, produciéndose así un perfil de temperatura que crece con la altura. A este tipo de inversión térmica se la llama también inversión nocturna y alcanza su máximo gradiente durante la madrugada. Se desarrolla desde el nivel del suelo y puede alcanzar una altura de hasta 500 m lo cual "envuelve" a las emisiones, aún de fuentes altas, tendiendo a acumular los contaminantes emitidos. Cabe agregar que, dado que las inversiones nocturnas se producen con preponderancia de cielos despejados, no se esperará en esas condiciones la presencia de lluvias con su poder de "enjuague" característico.

Inversión térmica por subsidencia
Este tipo de inversión se da lejos de la superficie terrestre. Una masa de aire a determinada altura de la CLP ejerce presión sobre las capas de más abajo produciendo compresión adiabática y consecuentemente calentando el aire mientras baja (subsidencia refiere a descenso, hundimiento). Se forma una capa de inversión que hace de tapa a una capa que se encuentra por debajo. Este tipo de inversión se da en las cercanías de centros de alta presión y a alturas que están muy por encima de las fuentes de emisión por lo que no contribuye en gran medida a la contaminación de corto plazo (aunque puede ser peligrosa cuando las condiciones se mantienen por varios días).

Otros tipos de inversión térmica se dan cuando:
a) tiene lugar el fenómeno de brisa de mar y tierra (Sección III.10).
b) cuando un frente cálido pasa por encima de una gran masa de agua (que se halla más fría).

La Figura III.7 (adaptada de Wark et al. (1998)) ilustra los dos tipos más importantes de inversión térmica.

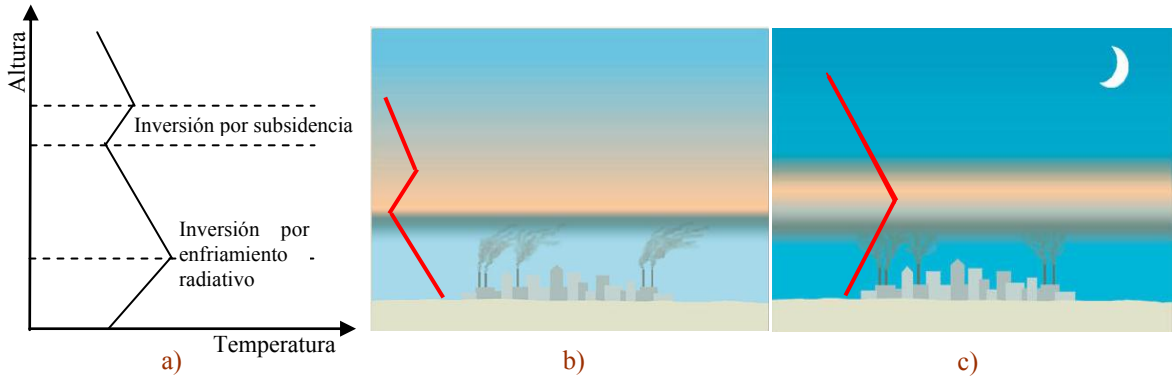


Figura III.7: a) Perfil de temperatura con dos tipos de inversiones b) Subsidencia c) Inversión nocturna. La zona celeste opaco en la parte b) indica la presencia de agentes contaminantes (zona gris) acumulados en las cercanías de la base de la capa de inversión. La parte c) muestra una atmósfera con acumulación de contaminantes (en proporción mayor que en la figura anterior) hasta llegar a la base de la capa de inversión. Las fotografías fueron tomadas de Lutgen y Tarbuck (2013).

III.8 Estabilidad atmosférica y contaminación

Además de lo mencionado en relación al viento como agente de transporte de los contaminantes, se debe destacar que existe una relación estrecha entre la estabilidad atmosférica y la calidad del aire. La atmósfera es el cuerpo receptor de los contaminantes del aire (no solamente especies químicas sino también ruidos, radiación electromagnética, etc.) emitidos por las distintas fuentes; dichos contaminantes son los agentes de degradación de la calidad del aire. En relación a las especies químicas y al material particulado (inerte, con especies depositadas o agentes bióticos) son los procesos meteorológicos los que definen el impacto de los contaminantes sobre la salud humana, la fauna, la flora y el paisaje. También definen la duración y el área de mayor impacto del evento. Se ha observado (Seinfeld y Pandis, 2006) que si las emisiones diarias son constantes o tienen variaciones por un factor menor a 2 la calidad de aire puede registrar variaciones diarias de hasta un factor de 10. Esta variabilidad en las concentraciones dada por las condiciones meteorológicas es lo que puede designar a una atmósfera como "limpia" o "contaminada".

Mientras que la velocidad horizontal del viento tiene una gran influencia sobre el grado de mezclado inicial de los contaminantes emitidos, la estabilidad atmosférica determina el grado en que el aire, que se ha contaminado, se mezcla con aire limpio de capas superiores (Lutgens y Tarbuck, 2013).

III.9 Capa límite planetaria

El límite inferior de la troposfera (superficie de la tierra) tiene una gran influencia sobre los procesos de transporte (tales como el de arrastre por fricción, de turbulencia, de evaporación, calor, etc.) hasta el rango de 100- 3000 m de altura constituyéndose una capa límite a partir de la cual dichos procesos dejan de ser dominantes. El resto de la troposfera es denominada atmósfera libre (ver Figura III.1) debido a que está libre de la influencia de la superficie y el viento es casi geostrófico (aquel que ocurre paralelo a las isobaras rectas, en general, a más de 600 m desde el nivel del suelo).

La capa límite planetaria (CLP) -también llamada capa límite atmosférica (CLA)- toma el nombre de un trabajo de Ludwig Prandtl sobre aerodinámica publicado en 1905. La CLP puede definirse como la parte de la troposfera que es directamente influenciada por la superficie terrestre y que responde a los procesos de transporte en intervalos de tiempo no mayores a una hora (Stull, 1988) o de pocas horas (Sportisse, 2008) y hasta alrededor de un día (Garratt, 1992).

Sobre los océanos la altura de la CLP varía muy poco en el tiempo y es bastante constante en grandes distancias (cubriendo cientos de kilómetros horizontales). Esto se debe a que la superficie del mar intercambia calor con capas de agua que están por debajo con mucha eficiencia y a que la capacidad calorífica del agua (comparada con la de la tierra) es alta (absorbe calor con poca modificación de la temperatura). Los cambios de la CLP sobre los océanos se deben a otras causas (fenómenos de mesoescala y escala sinóptica (ver Sección III.3)). Sobre las superficies de tierra (aunque dependiendo mucho de los materiales presentes) se produce una gran variación diurna de la temperatura en la CLP (cosa que no es apreciable en la atmósfera libre). Tanto sobre los océanos como sobre la tierra la CLP se hace más delgada en presencia de centros anticiclónicos (alta presión) que ejercen presión sobre las zonas cercanas a la superficie induciendo el transporte horizontal del aire hacia áreas de baja presión. En las zonas de baja presión convergen vientos provenientes de los centros anticiclónicos que hacen que la CLP se eleve y se haga más difícil definir su límite, en estos casos suele tomarse como referencia la altura de las nubes (aunque estas pueden hallarse bastante por debajo del tope de la CLP). En latitudes medias (estrictamente entre 30 y 60 grados de latitud sur o norte) es en donde la CLP se ha estudiado más, dado que es donde se halla la mayor cantidad de población mundial. En estas latitudes, en superficie terrestre planas y en zonas más bien de alta presión y con advección (movimiento horizontal convectivo, por ejemplo, viento suave) la CLP tiene una estructura bien desarrollada que evoluciona durante el ciclo diario de calentamiento y enfriamiento de la superficie terrestre. La Figura III.8 presenta esquemáticamente este ciclo.

Si consideramos un día que se inicia libre de nubes, la Capa de Mezcla (CM) va creciendo según ocurre el calentamiento de la superficie terrestre. Su formación comienza aproximadamente media hora después de la salida del sol que es cuando el flujo turbulento empieza a desarrollarse y va ganando altura. La CM se caracteriza por tener un alto grado de mezclado (de calor, humedad y cantidad de movimiento en el sentido vertical) e inestabilidad atmosférica generando masas de aire ascendentes (corrientes térmicas). La CM es también llamada capa convectiva por estar dominada por el transporte de calor tanto desde la superficie calentada por el sol como del enfriamiento radiativo de las nubes). Esta capa suele alcanzar su máxima altura durante el atardecer, su crecimiento se debe a la incorporación de masas de aire desde arriba (Zona de Arrastre- ZA también llamada capa interfacial (Garrat, 1992)). La ZA es una capa estable en el límite superior de la CM que hace de tope (en ZA hay inversión de temperatura por subsidencia- Figura III.7b) de las corrientes térmicas restringiendo así el dominio de la turbulencia. La ZA limita arriba con la atmósfera libre. Cuando hay suficiente humedad hay un límite para el ascenso de las masas de aire caliente teniendo lugar la formación de los denominados cúmulos de buen tiempo. Si la cobertura de nubes se hace más importante la radiación que llega a la tierra disminuye y, por lo tanto, la emisión de la tierra afectando así el desarrollo en altura de la CM.

Alrededor de media hora antes de la puesta del sol las corrientes térmicas van cesando (debe haber ausencia de vientos fríos), o sea, hay decaimiento de la turbulencia (y por lo tanto el mezclado ya no es intenso): la capa con estas características es llamada Capa Residual (CR) y es en general atmosféricamente neutra (Figura III.6a).

A medida que progresa la noche la porción baja de la CR es transformada debido al contacto con la tierra. Se forma una capa estable (con inversión de temperatura desde el nivel del suelo hasta donde comienza la CR). Esta capa es llamada Capa Estable Nocturna (CEN) y presenta niveles muy bajos y esporádicos de turbulencia con vientos suaves y grandes probabilidades de calmas cerca de la superficie. A una altura de 200 m suelen desarrollarse corrientes a chorro que pueden involucrar altas velocidades de viento (alrededor de 40 km h^{-1}). Unos metros por encima del chorro los vientos tienen velocidades

más bajas. El límite superior de la CEN es difuso, puede decirse que se va mezclando suavemente con la CR que se halla arriba; este límite suele definirse en términos de la altura a la que la turbulencia representa un pequeño porcentaje de la turbulencia que se halla a nivel de la superficie.

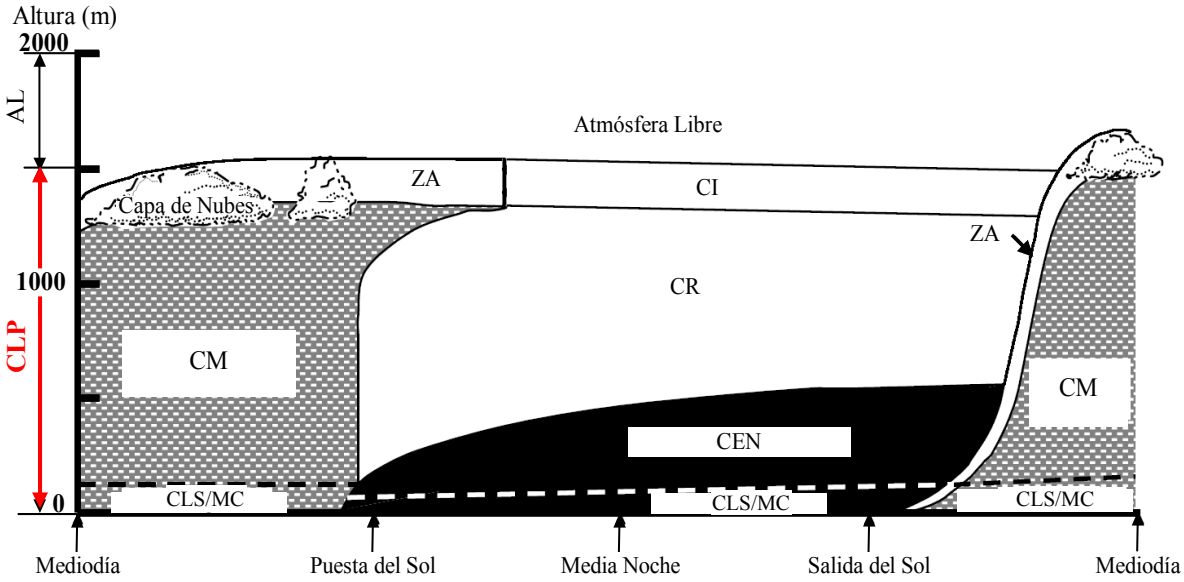


Figura III.8: Ciclo diario de la CLP (tomado de Stull (1988)). Sobre el eje *Y* se han indicado la CLP (capa límite planetaria) y la AL (atmósfera libre) en correspondencia con la Figura III.1. CLS: Capa Límite Superficial, CM: Capa de Mezcla, CR: Capa Residual, MC: Microcapa (a pesar de ser de solo unos centímetros se halla graficada ampliada para mejor visibilidad), CI: Capa de Inversión, ZA: Zona de Arrastre o capa interfacial, CEN: Capa Estable Nocturna.

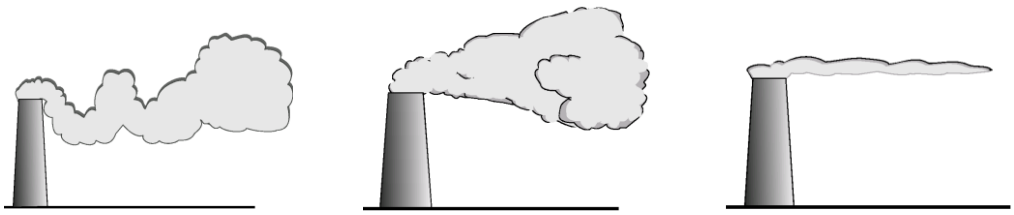
En una primera etapa del ciclo diario (luego de la salida del sol), el desarrollo de la CM va produciendo la destrucción de la CR y de la CEN mientras que en una segunda etapa (luego de la puesta del sol) es la CM la que destruyéndose como tal da lugar a la formación de la CR y a la CEN.

El ciclo "arquetípico" (Bretherton, 2002) descripto arriba se ve afectado con la presencia de altas rugosidades del terreno o topografías complejas, cielos cubiertos, el pasaje de ciclones a macroescala o fenómenos locales como el de brisa marina cuando tiene mucho desarrollo, entre otros, o sea que, dependiendo de la influencia de estos fenómenos, el ciclo descripto puede distorsionarse en distinto grado.

Existe una Capa Límite Superficial (CLS) delgada (que puede estimarse en un 10% del espesor de la CM) que se halla en la parte más baja de la CLP que se caracteriza por tener variaciones de turbulencia térmica y mecánica menores al 10%. Dentro de esta capa se halla una finísima capa (de solo algunos centímetros) donde hay dominio de la difusión molecular (por encima del transporte turbulento) llamada capa superficial o Microcapa (MC) (ver parte inferior de la Figura III.8).

En el ciclo diario es posible encuadrar varios de los fenómenos que guardan relación con los contaminantes del aire.

En la CM el mezclado de los contaminantes emitidos es alto por lo que la concentración tenderá a ser homogénea en esta capa (debido a la turbulencia) pero, debido a la inestabilidad atmosférica y en presencia de fuentes de emisión de tipo chimeneas, el flujo de los contaminantes adopta una forma ondeada con remolinos ("looping") (Figura III.9a) estando "atrapado" en la CM, principalmente, cuando hay un clara inversión en la ZA y se trata de zonas de alta presión.



a) b) c)

Figura III.9: Algunas formas que adquieren las plumas de chimeneas según los distintos tipos de estabilidades atmosféricas a) Forma de remolino (predominio de turbulencia vertical- Figura III.4b2) b) Forma de cono (equilibrio entre turbulencia vertical y horizontal- Figura III.4b1) y c) Forma de tubo (predominio de turbulencia horizontal- Figura III.4b3).

En la CR y debido a la neutralidad atmosférica, la turbulencia tiene similar intensidad en todas las direcciones (horizontal y vertical) dando lugar a un flujo de tipo cónico "coning" (Figura III.9b). Según las características de los contaminantes, los mismos tenderán a flotar o a precipitarse y en caso de ser reactivos darán lugar a nuevas especies (contaminantes secundarios). En el amanecer, cuando todavía quedan restos de esta capa, el contacto con la radiación solar puede producir reacciones fotoquímicas. La humedad que puede ir acumulándose durante varios días en la CM irá quedando retenida en la CR volviendo a pasar a la CM y en algún momento dará lugar a la formación de nubes. En la CEN los contaminantes se dispersan muy poco verticalmente predominando la dispersión horizontal. Las emisiones de una chimenea darán lugar a un flujo de tipo tubo ("fanning") (Figura III.9c). Si las velocidades horizontales de viento son bajas el flujo resultante puede oscilar en distintas direcciones. El lector podrá recurrir a la Figura III.4b para visualizar mejor la relación entre el tipo de estabilidad atmosférica y el tipo de plumas posibles. Descripciones detalladas del efecto de la estabilidad en las plumas se hallan en Stull (1988) y Arhens (2009).

Es posible definir el **potencial de contaminación** de una dada atmósfera como la capacidad que tiene la misma para diluir los contaminantes que se emitirán al aire. Para cuantificar este potencial puede calcularse el **índice de máxima ventilación** de una zona que tiene en cuenta el transporte y la dispersión de los contaminantes. El mismo se obtiene multiplicando la altura máxima de la CM por la velocidad del viento transporte (es un tipo de viento promedio que tiene en cuenta el perfil vertical de velocidades hasta el tope de la CM) (Gassmann, 1998). El índice de máxima ventilación permite tipificar el grado de autodepuración que posee la atmósfera característica de un lugar permitiendo establecer un mapa regional (Mazzeo y Venegas, 1999). Un valor crítico, por debajo del cual se considera que la atmósfera tiene baja capacidad de depuración es el de 6000 m^2 s^{-1}. A modo de ejemplo se puede considerar la zona donde se ubica el cordón industrial Rosario - La Plata, que tiene frecuencias significativas de ocurrencias con condiciones de baja ventilación (Gassmann, 1998; Gassmann y Mazzeo, 2000).

III.10 Brisas de mar y tierra

Bajo este nombre puede resumirse un fenómeno de circulación de vientos en una zona en donde se hallan colindando una superficie de tierra (plana) y un cuerpo de agua, ambos significativos por su extensión. Otros nombres comunes son brisa de lago y tierra, brisa de agua y tierra, etc.

La brisa de mar- tierra (o marina para simplificar) es un tipo de circulación térmica (entre otros tales como el de la isla de calor (Oke, 1987)); es un fenómeno que se puede dar en escala local y en mesoescala (Arhens, 2009). Puesto que la mayor parte de la población mundial vive dentro de los 200 km de una costa, resulta muy verosímil que este fenómeno haya sido observado desde la época de los antiguos griegos (Simpson, 1994).

Realizar una descripción aislada de la brisa marina implica prescindir de otros fenómenos locales y de escala sinóptica (Simpson, 1994) que en el caso real estarán superpuestos y podrán reforzar o inhibir algunos de sus características (Vallero, 2008).
En la Figura III.10 se consideran dos áreas de terrenos que se hallan a temperaturas distintas ($T_1 < T_2$). La parte a) de la figura muestra un conjunto de superficies de presión homogénea separadas una cierta distancia entre sí. Si la superficie del terreno se eleva a la temperatura T_2 (Figura III.10b) se observará un distanciamiento de las superficies de igual presión debido a que el aire se vuelve menos denso a medida que se asciende verticalmente. En la figura se ha establecido el plano de igual presión **P** como referencia. Si las dos bases representadas en la Figura III.10 fueran contiguas considerando una altura distinta de la del plano a **P** es posible considerar que se generará una dinámica de circulación horizontal de fluido.

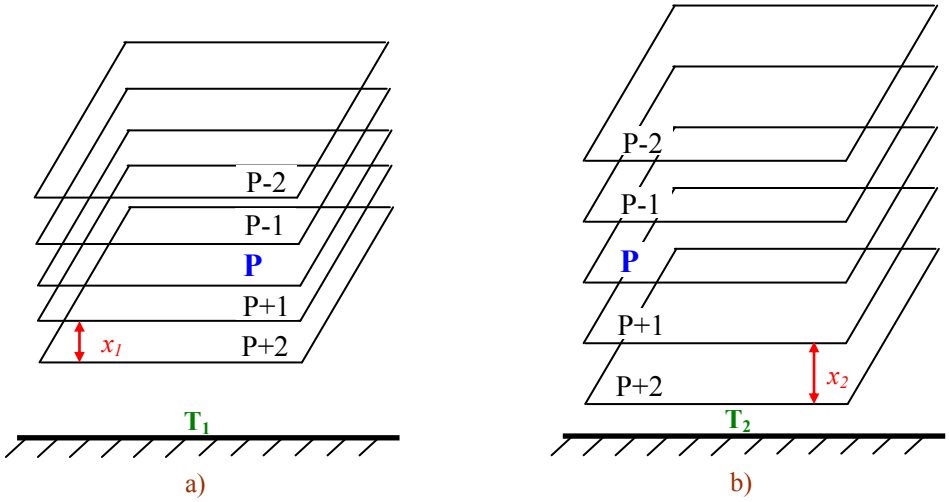


Figura III.10: Gradientes de presión en dos áreas que se hallan a temperaturas distintas (un gradiente típico cercano a la superficie terrestre es de 1 hPa/8.6m). La superficie de presión homogénea **P** ha sido tomada como referencia y se halla a la misma altura en los dos casos. **P+1** indica una unidad arbitraria por encima de **P**, podría ser por ejemplo, 1 hPa (hecto Pascal).
a) base a $\mathbf{T_1}$ tiene las superficies de igual presión separadas una cierta distancia x_1
b) base a $\mathbf{T_2}$>$\mathbf{T_1}$ muestra como la disminución de densidad del aire por la elevación de la temperatura de la base produce una mayor separación (x_2) entre las superficies de igual presión.

La brisa marina se produce por el calentamiento diferencial que experimentan el agua y la tierra cuando incide la radiación solar. Según Lutgens y Tarbuck (2013) existen cuatro factores que operan de manera simultánea produciendo el calentamiento (o enfriamiento) del agua más lentamente que el de la tierra: a) la radiación solar penetra varios metros dentro del cuerpo de agua (puede llegar hasta 6 m) entonces hay mucho volumen puesto en juego (la tierra en cambio es opaca y toda la energía se emplea en calentar apenas unos pocos centímetros de profundidad); b) el agua es muy móvil y la convección hace que se intercambie calor entre las distintas zonas haciendo que las zonas más calientes se atemperen. La tierra no tiene transporte convectivo y su capacidad conductiva es muy baja por lo que eleva más fácilmente su temperatura; c) la capacidad calorífica del agua es más de tres veces más alta que la de la tierra, por lo que para elevar 1 °C a un gramo de masa se requiere más energía; d) parte de la energía que recibe el agua la utiliza en vaporizarse (hay moléculas que alcanzan la suficiente energía como para pasar del estado líquido al de vapor) quedando solamente el resto para elevar la temperatura de la masa involucrada.
En días despejados de verano y en presencia de vientos moderados el calentamiento diferencial adquiere su máximo potencial (Vallero, 2008). Como se dijo en la Sección III.9

el fenómeno de brisa de mar y tierra modifica la CLP. La Figura III.11 muestra el fenómeno completo de circulación de la brisa marina.
Como se mostró en la Figura III.10 el calentamiento de la tierra produce un calentamiento del aire por encima de ella haciendo decrecer su densidad. En la Figura III.11 se puede apreciar la forma que adquieren las isobaras cuando se establece el fenómeno en estudio. Esto último hace que a una misma altura $P_A < P_{A'}$ mientras que $P_{B'} < P_B$.

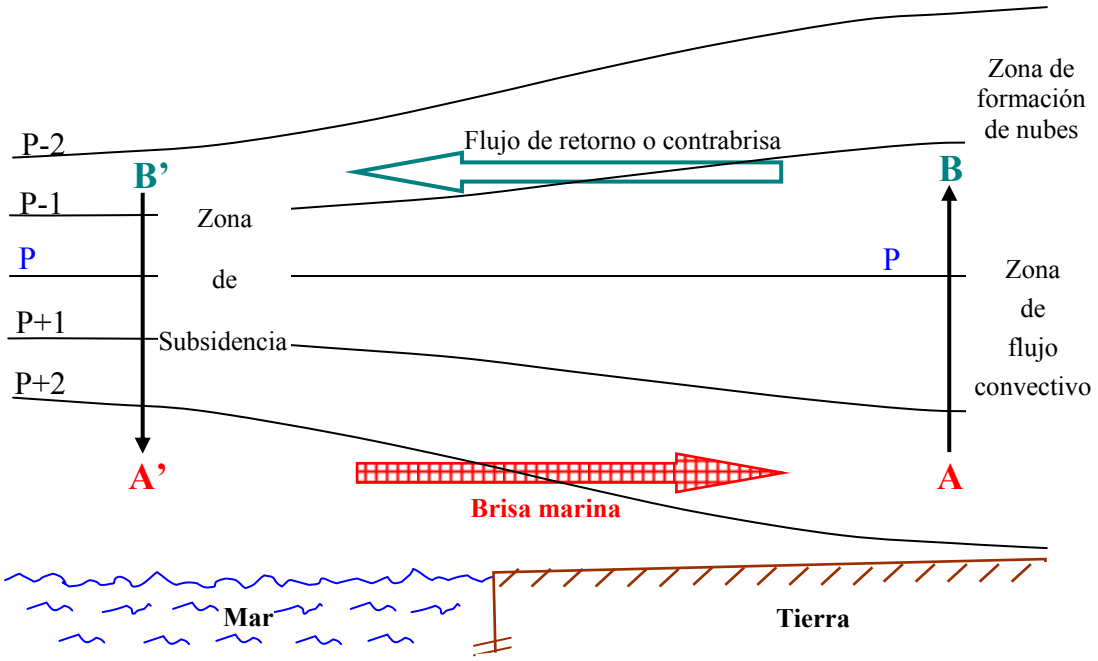


Figura III.11: Celda de circulación de la brisa marina. La denominación del fenómeno se debe al viento que sopla en la parte baja de la celda desde el mar hacia la tierra.

El aire cercano a la tierra se eleva produciendo un aumento de presión hacia las capas más altas (de tal forma que para una altura dada habrá más presión en el aire sobre la tierra que en el aire sobre el mar, o sea, $P_B > P_{B'}$). Este aire, que ha subido sobre la tierra y que ha generado baja presión cerca de la superficie, es reemplazado por aire frío proveniente del mar que se halla a mayor presión, $P_{A'} > P_A$.
El aire en la zona del punto B fluye hacia el mar por arriba en virtud de que hay un gradiente de presión favorable, $P_B > P_{B'}$ (el mar arriba tiene menos presión que la tierra arriba a la misma altura). El aire sobre el mar en la zona del punto A', que se ha ido hacia la tierra circulando por debajo, es reemplazado sobre el mar por aire de capas más altas (zona del punto B'). En la zona entre B' y A' tiene lugar el fenómeno de subsidencia (desplazamiento de aire frío desde arriba hacia abajo). De esta manera se cierra un ciclo de flujo circular (o celda de circulación). Notar que la fuerza impulsora es la temperatura que genera cambios en la presión y que los gradientes de presión que gobiernan la circulación son horizontales. La celda de circulación comienza formándose cerca de la interfase costera durante la mañana y se va expandiendo hacia el mar y la tierra simultáneamente, al mismo tiempo que se desarrollan también en altura (Planchon et al., 2006). Dependiendo de las condiciones del Tiempo, el desarrollo de esta celda puede inducir chaparrones o tormentas. La celda puede alcanzar los 100- 200 m de altura y hasta los 2000 m (Oke, 1987) o 4000 m (Celemin, 1984). La brisa de mar puede desarrollar velocidades inherentes de hasta 36 km h^{-1}. El desarrollo horizontal de la celda de circulación puede ir desde unos cientos de metros a unos 20- 50 km (en latitudes medias) y hasta cientos de km en los trópicos (Emeis, 2012). El mecanismo mostrado suele tener, en la medida en que haya presencia de humedad, una cadena de nubes de tipo cúmulos que se forman en la parte

marina superior de la celda y viajan hacia el continente (indicando el frente de la brisa marina). En las latitudes medias (Sección III.9) el fenómeno de las brisas de mar y tierra es observable principalmente durante la estación cálida en zonas anticiclónicas en donde hay baja producción de nubes al mismo tiempo que vientos con bajas velocidades pero en el Ecuador se observan durante todo el año (Oke, 1987).

La presencia de brisas de mar y tierra no solo influye sobre el clima local, el transporte de los contaminantes, el transporte de insectos, pájaros y polen (Gassmann et al., 2002) sino que afecta una variada gama de actividades económicas: agropecuarias, deportivas, y de navegación, etc. (Simpson, 1994; Borque et al., 2008). Ultimamente (Orton et al., 2010), se ha valorado el rol importante que tiene el desarrollo de estas circulaciones en el ciclo global del carbono y en el intercambio gaseoso de los estuarios (ventilación del agua). Cabe agregar que en muchas ciudades (tales como en Los Angeles en EUA o Atenas en Grecia) la brisa marina cumple el rol de exacerbar la contaminación del aire (Jacobson, 2005).

Simpson (1994) menciona un método muy sencillo para obtener un "índice de brisa marina" capaz de decir cuando este fenómeno es posible (ver Nota al final de esta sección).

La contraparte de la brisa marina es la brisa de tierra que tiene su inicio a la noche y su máximo desarrollo en horas de la madrugada, es un fenómeno menos importante (la celda de convección es más pequeña en altura y penetración) debido a las estabilidades nocturnas prevalentes (Oke, 1987) y puede no darse (Wanta, 1968; Simpson, 1994; Emeis, 2012). La velocidad de los vientos de brisa de tierra puede ser de hasta 7 km h^{-1} (Oke, 1987).

Existen diversos recursos instrumentales para medir las brisas de mar y tierra, lo cual implica en muchos casos determinar la factibilidad de su ocurrencia, las horas de desarrollo y caracterizar su penetración en el continente. Estos instrumentos van desde observadores calificados y la medición de parámetros meteorológicos desde tierra (temperatura, humedad, velocidad del viento, etc.) hasta la toma de fotografías, filmaciones, mediciones con globo sonda, radiosondeos, radares, LIDAR ("Light Detection and Ranging"), SODAR ("Sound Detection and Ranging") y el empleo de satélites (Simpson, 1994; Orton et al., 2010).

Nota: Partiendo de un balance de fuerzas se llega a la relación $v^2/\Delta T = R$ donde v es la velocidad horizontal del viento en la costa y ΔT es la diferencia de temperatura entre temperaturas observadas por encima del agua y por encima del terreno; R se debe determinar para la zona de estudio. Si ese cociente sobrepasa un valor crítico de R la brisa marina no se producirá, de lo contrario la misma tendrá lugar.

III.11 Estaciones del año

Se han descripto hasta aquí conceptos generales, algunos de ellos de importancia por su ciclo diario. Cabe ahora mencionar otro de los ciclos sobre cuya base se ha trabajado durante la tesis: el ciclo anual. Las estaciones del año están determinadas por la cantidad de radiación solar que llega al planeta tierra. Esta cantidad depende de la duración del día y el ángulo de incidencia de los rayos del sol en relación a la superficie terrestre (Lazaridis, 2011). En las latitudes medias el verano es definido como la estación más cálida mientras que el invierno como la más fría. Si el año es dividido en cuatro estaciones de igual duración, el verano puede definirse para el hemisferio sur conteniendo a los meses más cálidos, o sea, Diciembre de un año determinado y Enero y Febrero del año siguiente. El invierno corresponderá a los meses de Junio, Julio y Agosto siendo la primavera y el otoño los trimestres intermedios. Algo análogo es posible definir para el hemisferio norte. Esta definición dada es muy característica de la meteorología (Arhens, 2009) y difiere de la definición dada por las ciencias astronómicas. Una descripción detallada de como se producen las estaciones se da en Lutgens y Tarbuck (2013); desde otra perspectiva se explican en Bely et al. (2010).

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem"
John Tukey

"*The use of any knowledge reaches into three areas of the mind: the search for truth, the skill of forecasting and the gift to imagine a future different from the present. There will never be clear-cut rules of procedure.*"
Lazarsfeld and Reitz (1970)

"...we are always searching for something hidden or merely potential or hypothetical, following its traces whenever they appear on the surface"
Six Memos for the Next Millennium, Italo Calvino (1996)

Capítulo IV
Similitud- disimilitud, regresión y tendencia

En este capítulo se presentan y discuten conceptos, recursos gráficos y herramientas de estadística "clásica" y "robusta" (principalmente para los casos univariado y bivariado) que fueron empleados a lo largo de las distintas publicaciones. Se analizan resultados de observaciones de SO_2 en aire en dos períodos y sitios y su relación con los vientos dominantes. Se estudian los ciclos diarios y anuales así como la tendencia de los vientos más importantes para el transporte de los contaminantes. Se analizan las velocidades de los vientos y la estructura de las calmas en la zona y se sugiere la localización de un área para el seguimiento de concentraciones de fondo. Los resultados presentados son, en su mayoría, partes de varios de los trabajos publicados (Rosato et al., 2001; Ratto et al., 2005, 2006, 2009, 2012a, 2012b, 2012c).

IV.1 Datos Atípicos

En el Capítulo I se señaló la importancia de la estadística robusta y, en particular, la de considerar la presencia de potenciales valores atípicos en los datos de trabajo. Conviene realizar aquí (antes de abordar los temas específicos de este capítulo y dada su importancia para el capítulo siguiente) algunos comentarios y precisiones sobre la presencia de tales valores y su detección.

Si bien no hay una definición universalmente aceptada de lo que es un valor atípico o un conjunto de ellos (Hodge y Austin, 2004), existen varias descripciones que en conjunto dan una idea conceptual y amplia del tema. Un valor atípico es en general definido como un punto tal que, en el contexto práctico de otras observaciones, contrasta con ellas (Barnett y Lewis, 1994), se desvía de manera notable de ellas (Grubbs, 1969) o resulta sospechoso de haber sido generado por otro mecanismo (Hawkins, 1980). Cuando los datos forman grupos definidos Aggarwal y Yu (2001) señalan que los valores atípicos pueden considerarse como aquellos que quedan afuera de los grupos y no forman parte del ruido de los datos. Las frases "se desvían", "contrasta", "resulta sospechoso", "quedan afuera" y "no forman parte del ruido" ponen en evidencia el carácter subjetivo de estas definiciones. Una definición más ajustada de valor atípico implica realizar suposiciones acerca de la estructura de los datos y sobre el modelo que se utilizará para detectar los atípicos (Ben-Gal, 2005).

El término "valor atípico" está muy difundido en la literatura pero también suelen utilizarse los términos "dato espurio", "observación no representativa", "dato dudoso", "valor discordante", "dato malo", etc. (Seber, 1984; Barnett y Lewis, 1994; Markatou y Ronchetti, 1997). Una larga lista de aplicaciones prácticas en las que se requiere la detección de dichos valores se halla en Hodge y Austin (2004). El interés por la presencia

de los valores atípicos en los datos ambientales se remonta a mediados del Siglo XIX (Barnett, 2004) pero ¿por qué detectar atípicos es tan importante? Una posible respuesta ha sido enunciada en la Sección I.1.5 (en relación a la distorsión de los estimadores) pero cabe agregar, desde una perspectiva más amplia, que es una responsabilidad del investigador conocer si los datos forman un grupo homogéneo o no, o si contienen errores (Bartkowiak y Szustlewicz, 1997). Barnett (2004) señala que se debe saber que importancia tiene el valor atípico en relación al mecanismo que genera los datos y en relación al modelo que se supone para ellos.
El concepto de valor atípico que se ha delineado hasta aquí se ve reforzado si se introducen en la discusión los conceptos de "valores extremos" y de "datos contaminantes". Si se considera un conjunto de datos que siguen una distribución normal (modelo básico de datos sin contaminación) tal conjunto tendrá valores extremos (por ejemplo, el máximo). Esto es, un extremo puede ser o no un valor atípico. Si el conjunto de datos contiene datos de otro origen (otra distribución) se dice que la muestra se halla contaminada. Un valor atípico podrá no ser un dato contaminante. La Figura IV.1 muestra lo que se acaba de describir.

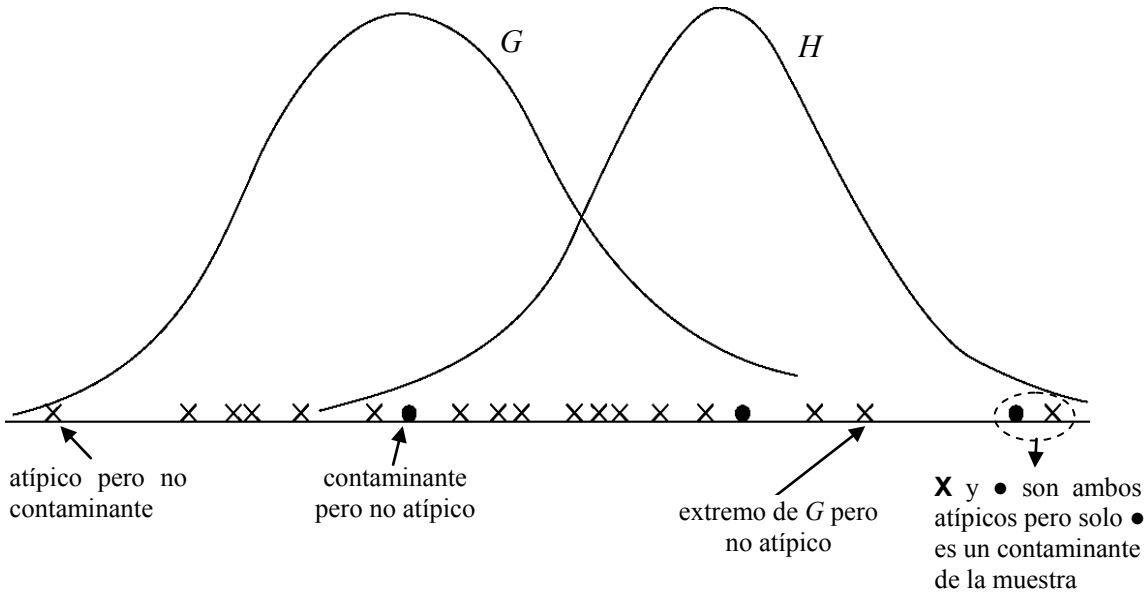


Figura IV.1: Dos curvas de densidad de distribución (tomadas de Barnett (2004) Capítulo 3):
G distribución normal de la que provienen los datos, simbolizados con **X**,
H distribución de la que provienen otros datos, simbolizados con ●

Por lo tanto, el investigador deberá tener en cuenta y tomar una decisión sobre que hacer con los datos contaminados o los valores atípicos (cabe aclarar que algunos autores no hacen discriminación entre atípico y contaminante y los consideran como sinónimos). Según Barnett (2004) existen tres posibilidades: a) eliminar (rechazar) el dato, b) identificarlo para realizar algún tipo de consideración especial o c) tolerarlos utilizando un procedimiento que sea poco influenciable por su presencia (enfoque robusto).

Como se ha señalado, un solo valor atípico puede afectar de gran forma la estimación de parámetros. La Figura IV.2 muestra cómo queda afectado (en valor absoluto y en signo) el coeficiente de correlación ρ de Pearson (Sección IV.2.1) debido a la presencia de dos valores atípicos. Nótese además que la presencia de los mismos afecta la estructura (forma) de la nube de puntos.

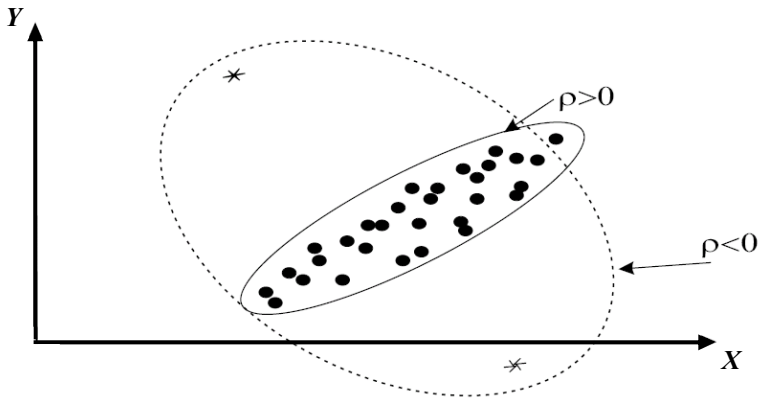


Figura IV.2 :
Nube de puntos y el impacto sobre el coeficiente de correlación para un caso bivariado. (Gráfico tomado de Shevlyakov y Vichelv-sky (2000)).

La Figura IV.3 muestra el caso en que sin quedar afectada la estructura por la presencia de un valor atípico (punto A), quedan afectadas las magnitudes de los estimadores ($\bar{x}$, $\bar{y}$, ρ, etc.) aunque los coeficientes de regresión resulten muy similares (con o sin la presencia del punto A).

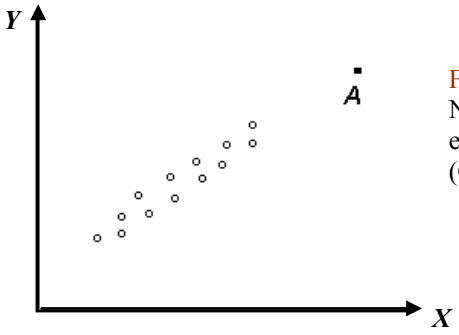


Figura IV.3 :
Nube de puntos y el impacto sobre la magnitud de los estimadores sin afectar la estructura general de los datos. (Gráfico tomado de Bartkowiak y Szustlewicz (1997)).

Ben-Gal (2005) señala que existen alternativas paramétricas, no- paramétricas y técnicas de agrupamiento de datos para detectar valores atípicos. Los procedimientos paramétricos implican la suposición de una determinada función de densidad de distribución en los datos; darán como atípicos aquellas observaciones que se alejen de las suposiciones realizadas. Esto se puede realizar de varias maneras, por ejemplo, empleando gráficos cuantil- cuantil (QQ-Plots- Sección V.5.2.4.1) y/o tests. Los procedimientos no paramétricos se basan en el cálculo de distancias (por ejemplo utilizando distancia de Mahalanobis– Sección V.5.2.4.2). Los procedimientos basados en técnicas de agrupamiento se basan en la idea de que cada dato pertenece o bien a un grupo o es un atípico (Aggarwal, 2013).

Cuando los datos dependen de una única variable (caso univariado) o de dos variables (bivariados) los métodos para detectar valores atípicos resultan sencillos (dada su visibilidad). Pero a partir de tres o más variables la identificación de los atípicos se vuelve más compleja; esto ha dado lugar a una gran variedad de métodos de diagnóstico (identificación) y robustos (Hawkins, 1980; Barnett y Lewis, 1994; Maddala y Rao, 1997; Cohen et al., 2003; Belsley et al., 2004; Maronna et al., 2006; Aggarwal, 2013).

La Figura IV.4 muestra un valor atípico que no es apreciable en las variables marginales (individuales *X* e *Y*); cualquier método de exploración de atípicos o un test deben tener en cuenta este efecto de interacción entre las variables.

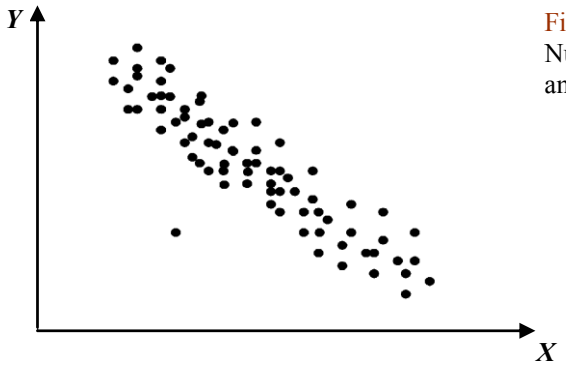


Figura IV.4 :
Nube de puntos y un valor atípico en relación a ambas variables a la vez.

Frecuentemente, el investigador necesita saber sobre la presencia de varios valores atípicos. Existen dos fenómenos importantes a considerar: a) el efecto de enmascaramiento ("masking") y b) el efecto de hundimiento ("swamping"). La Figura IV.5 ejemplifica para el caso univariado la problemática de estos dos efectos.

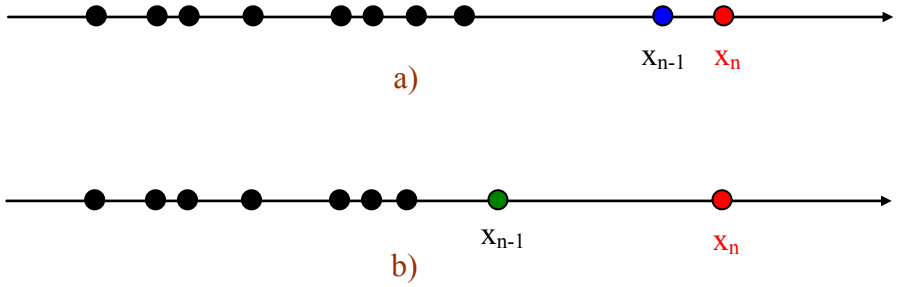


Figura IV.5.: Dos conjuntos de datos para mostrar los efectos de:
a) enmascaramiento y b) hundimiento. Ejemplo tomado de Barnett (2004).

Hay dos posibles enfoques a adoptar para determinar si x_{n-1} y x_n son valores atípicos: realizar un test consecutivo (se testea x_n y luego x_{n-1}) o realizar un test en bloque (x_{n-1} y x_n en conjunto). Ambos enfoques tienen dificultades conceptuales (Barnett, 2004): el consecutivo puede fallar en el primer paso porque dada la presencia de x_{n-1} (Figura IV.5a), x_n no aparecerá como valor atípico (quedando este último "enmascarado" en lugar de revelado); el enfoque en bloque declarará que ambos son valores atípicos.
En la Figura IV.5b el enfoque consecutivo dará como atípico a x_n mientras que el enfoque en bloque dará como atípicos a los dos (dado que considerar a ambos conjuntamente produce el "arrastre" o "hundimiento" de x_{n-1} generándose así un falso valor atípico).
La Figura IV.6 muestra los efectos de enmascaramiento y hundimiento en el plano para una nube de puntos cuando se busca estimar la correlación lineal. En el caso a) los atípicos "inflan" la matriz covarianza y permanecen indetectados. En el caso b) los atípicos no solamente aumentan la matriz de covarianzas sino que la distorsionan a tal punto que los verdaderos valores atípicos quedan indetectados y algunos datos que pertenecen a la mayoría del patrón de datos aparecen como atípicos.
Los riesgos de estos dos efectos pueden reducirse seleccionando el test más apropiado (Barnett y Lewis, 1994) pero cabe aclarar que hasta el momento no existe uno que sea el más abarcativo (Barnett, 2004). En los ejemplos mostrados, el análisis por inspección permite apreciar la distorsión que se generará en la matriz covarianzas; si bien existen otros enfoques (métodos basados en la simulación de datos tipo Monte Carlo) tal como el del elipsoide de menor volumen (MVE- minimum volume elipsoid); estos métodos no resultan del todo confiables (Bartkowiak y Szustlewicz, 1997).

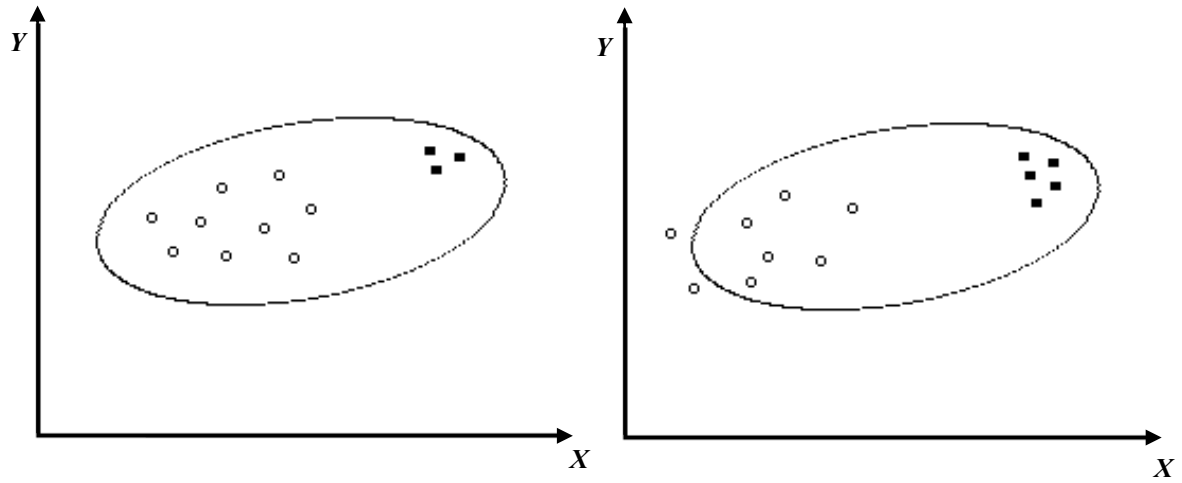


Figura IV.6: a) los valores atípicos (cuadrados rellenos) quedan enmascarados en el contexto del grueso de los datos (círculos) b) dos valores pertenecientes al grueso de los datos quedan afuera de la nube de puntos debido al efecto de hundimiento que producen los verdaderos valores atípicos (cuadrados rellenos). Ejemplo tomado de Bartkowiak y Szustlewicz (1997).

Con lo considerado hasta aquí, se ha intentado mostrar que trabajar con el concepto de valores atípicos resulta fundamental, ya que tales datos pueden tener una importancia especial dentro del fenómeno que se estudia y porque su presencia puede afectar el modelado general de los datos. También es una forma de considerar características de los datos en cuanto a sus posibles orígenes (variabilidad inherente, error de medición, etc.).
Desde el punto de vista práctico el investigador se halla, por lo general, frente a un problema en particular, por ejemplo, determinar los parámetros de regresión, encontrar estructura de grupo en los datos, reducir la cantidad de variables, etc. En lo que respecta al presente capítulo cabe realizar algunas distinciones que favorecerán la interpretación de las aplicaciones que se describen a partir de la Sección IV.6. Cuando el objetivo es encontrar el grado de correlación entre vectores las variables involucradas tienen todas la misma jerarquía, en cambio en regresión las variables se distinguen entre la "respuesta" y las "explicativas". Esta distinción tiene importancia en cuanto a las posibles "vías de detección" de los atípicos. Para el caso de regresión con una variable explicativa, dadas la X (explicativa) y la Y (respuesta) pueden ocurrir valores atípicos en las Y (por ejemplo, cuando la variable X es fija como en el caso de una secuencia de años calendario) y/o valores atípicos en las X (llamados frecuentemente puntos palanca). Ambos pueden tener una fuerte influencia en la obtención de los parámetros de regresión, por lo cual, suelen llamarse atípicos de regresión (datos que desvían la relación lineal dada por la mayoría). Sin embargo, pueden existir casos en que algunos puntos tengan una fuerte desviación tanto en las X como en las Y pero que casi no influyan en la obtención de los parámetros de regresión (punto A de la Figura IV.3), a estos puntos se los suele llamar "buenos" puntos palanca y son también importantes de considerar para caracterizar los datos. Los valores atípicos en las Y se detectan en general (regresión lineal simple) en el análisis de los residuos mientras que los de las X son más difíciles de hallar y se deben emplear varias herramientas (Rousseeuw y Leroy, 1987). El caso multivariado es aún más complejo y se deben utilizar más recursos (Rousseeuw y Van Zomeren, 1990; Gnanadesikan, 1997; Maronna et al., 2006; Aggarwal, 2013).
Cabe agregar que, en la selección de un método robusto, el investigador debe, en la medida de lo posible, conocer la bondad del mismo en relación a cada uno los efectos: enmascaramiento -no detección de atípicos- y hundimiento -detección de falsos atípicos- (Wang y Serfling, 2012); pero también debe ponderar la relación entre eficiencia y

robustez (mencionada en la Sección I.1.5- Capítulo I).
Las tres posibilidades enunciadas arriba sobre la actitud del investigador frente al tema de los valores atípicos, son importantes; en la tesis se trabajó con énfasis en la exploración de atípicos utilizando procedimientos heurísticos (para su identificación) y se aplicaron métodos robustos (tolerancia a los atípicos) solo en los casos en que se consideró necesario dado que la estimación de parámetros robustos lleva asociada una mayor varianza (lo cual implica menor eficiencia (Filzmoser et al., 2009)). Dada la gran variedad de alternativas robustas tanto para correlación como para regresión (Sajesh y Srinivasan, 2013), el trabajo en equipo con un matemático especialista se hace indispensable.

IV.2 Similitud- Disimilitud

La capacidad de juzgar dos situaciones o dos objetos como parecidos depende de la inteligencia y, dado que hoy por hoy es difícil escribir programas que igualen nuestra capacidad de percibir analogías, puede considerarse a la "matematización del parecido" como un arte complejo. Se han desarrollado distintas concepciones sobre lo que es parecido, tales concepciones dan soluciones a algunos problemas y no a otros (Delahaye, 1997; Guthe et al., 2005; Wang et al., 2005; Veltkamp y Latecki, 2006).
Hay dos nociones que se han adoptado en la presente tesis debido principalmente a su interpretabilidad: la correlación como medida de "similitud" y distancia como medida de "disimilitud".

IV.2.1 Correlación

El análisis de correlación busca estimar la relación que tienen, por ejemplo, un par de variables u objetos (vectores) dados. La covarianza es una medida de tal relación, pero al no estar estandarizada la interpretación se dificulta. Surge de aquí la necesidad de operar con coeficientes de correlación que constituyan un punto de partida de varios métodos de análisis multivariado.
El conocido coeficiente de correlación de Pearson ("rho" de Pearson) para una muestra (bivariada) puede expresarse como:

$$\rho = \rho\ (x,y) = \frac{Cov\ (x,y)}{\sqrt{Var(x)}\ \sqrt{Var(y)}} \qquad -1 \le \rho \le 1 \qquad \text{ec. IV.1}$$

donde $$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

es la covarianza estimada entre las variables x e y. $Var(x)$ y $Var(y)$ son las varianzas muestrales de x e y mientras que $\bar{x}$ e $\bar{y}$ son las medias muestrales respectivas.
Este estadístico es ampliamente usado para expresar de manera resumida la relación entre dos variables o grupos de variables que definen a un objeto. Representa el grado de asociación entre dos variables y constituye una medida estandarizada de la dependencia lineal que pueden tener tales variables (Cuadras, 1996). Es un estimador ideal cuando los datos siguen una distribución bivariada normal lo cual no siempre es frecuente; tampoco es práctico demostrarlo. Cuando ρ se halla cerca de 1 o –1, indica que cada una de las variables puede ser predicha de manera lineal por otras de manera bastante exacta. El signo indica la dirección en que la relación tiene lugar, cuando es negativo indica que si una variable crece la otra decrece.
Se han asignado tradicionalmente dos desventajas para la aplicación de ρ (Wilks, 2006; Chatterjee y Hadi, 2006): una de ellas es la sensibilidad de este coeficiente a los valores atípicos, tal como se mostró en la Sección IV.1, debido a que se basa en las medidas de tendencia central y dispersión clásicas (media y desvío estándar) (Croux y Haesbroeck,

1999). La otra desventaja es su inhabilidad para detectar relaciones no lineales. A pesar de estas desventajas el coeficiente de correlación de Pearson es tan popular que cuando no se lo nombra específicamente se da por sentado que la correlación está basada en el ρ (Reimann et al., 2008).

Una alternativa menos sensible a la presencia de valores atípicos la constituye el coeficiente de correlación de rangos de Spearman (*Sr*), cuya aplicación no requiere realizar estimaciones de parámetros muestrales (tales como la media o el desvío). Este coeficiente no supone una relación estrictamente lineal entre las variables involucradas sino que exista entre ellas una relación monótonamente creciente (o decreciente). Esta característica lo hace un poco más tolerante frente a las no linealidades en el crecimiento de las x o de las y, y además, algo más robusto. En el caso en que la muestra siga una distribución normal bivariada *Sr* será menos preciso que ρ (EPA, 2006).

Para poder calcular el *Sr* se requieren al menos cuatro datos; se comienza reemplazando cada dato (x) por su rango (por ejemplo, 1 para el valor más pequeño de las x, 2 para el siguiente más pequeño, etc.) y de igual manera con las y. A los pares (x,y) formados se les calcula el coeficiente de Pearson. Detalles de cálculo pueden verse en EPA (2006) y Corder y Foreman (2014) mientras que una alternativa de cálculo en WHO (1980).

Otro estimador que suele utilizarse como alternativa al ρ de Pearson es el τ de Kendall que es similar al de Spearman en cuanto a que relaciona rangos diferenciándose del mismo en la manera en que se efectúan los cálculos (Wilcox, 2005). Este estimador también guarda cierta insensibilidad a los valores atípicos.

Existen otros estimadores que han sido específicamente diseñados para ser robustos a la presencia de atípicos. En general, se basan en la construcción de una matriz de covarianzas robusta. Existe una gran variedad de estimadores robustos de correlación (Wilcox, 2005; Maronna et al., 2006) y a pesar de que su empleo se va volviendo cada vez más familiar entre los investigadores no matemáticos, estos usuarios se ven enfrentados a una gran variedad de alternativas con distinto grado de sofisticación, que implican el manejo de parámetros de ajuste que no son de interpretación directa (Fauconnier y Haesbroeck, 2009).

El estimador de correlación MCD (Minimum Covariance Determinant- Mínimo Determinante de la Matriz de Covarianzas) propuesto por P. Rousseeuw en 1984 se hizo más conocido solo cuando se encontró una manera eficiente de calcularlo (Rousseeuw y Van Driessen, 1999). El MCD fue adoptado frecuentemente en el trabajo de tesis porque, a pesar de que su cálculo es complejo (involucra fases de análisis combinatorio), posee buenas propiedades matemáticas (Butler et al., 1993; Cator y Lopuhaa, 2010) superando a sus predecesores (tales como el MVE- elipsoide de menor volumen) y además su interpretación es tangible. Desde el punto de vista práctico el MCD aparece incorporado como función de biblioteca en el software *Scout 1.0* de 2008 de la US EPA (United States - Environmental Protection Agency) en la versión de Rousseeuw y Van Driessen (Rousseeuw y Van Driessen, 1999). El algoritmo de cálculo opera con submuestras de h datos (siendo n el número total de datos) con $n/2 < h < n$ buscando minimizar el determinante de la matriz covarianzas de dicha submuestra. Una vez determinada la submuestra óptima, la media y la matriz de covarianza clásicas son los estimadores utilizados en el cálculo del MCD (coeficiente de correlación) de forma análoga a lo expresado en la ecuación IV.1. Un algoritmo determinístico propuesto más recientemente por Hubert et al. (2012) mejora la performance del cálculo del MCD.

Una de las propiedades deseadas en un estimador robusto es su capacidad para tolerar la presencia de valores atípicos (ya sea en cualquiera de las variables individuales, en algunas o en todas ellas). La mínima proporción de datos observados que necesita ser reemplazada por valores atípicos para que los estimadores se distorsionen grandemente se denomina

punto de ruptura (*PR*) (BDP- breakdown point) y se puede expresar en porcentaje (Rousseeuw y Hubert, 2011). El MCD posee un *PR* del 50% cuando $h = (n + p + 1)/2$ siendo p el número de variables. Si n es grande entonces $h \approx n/2$. Si se supone un porcentaje de contaminación en los datos (presencia de atípicos) puede estimarse el *PR* de manera aproximada como $PR = (n-h)/h$ (Fauconnier y Haesbroeck, 2009). Por ejemplo, para $PR = 20\%$, $h \approx 0.8\ n$. La suposición de contaminación de la muestra constituye un parámetro de ajuste para el cálculo del MCD e influirá en el valor obtenido. Croux y Haesbroek (1999) recomiendan un valor de alrededor de $h \approx 0.75\ n$ para mantener una relación óptima entre robustez y eficiencia (ya que a mayor robustez se pierde eficiencia). Sin embargo, es el analista quien deberá, en definitiva, decidir según las características de los datos y el objetivo de estudio. El grado de robustez queda definido al fijar el *PR,* mientras que la eficiencia da cuenta de cuanto se aparta el valor del MCD del obtenido por un método no robusto (por ejemplo, con el ρ de Pearson) cuando no hay valores atípicos en los datos. Cabe recordar que el *PR* del ρ es 0% (Shevlyakov y Vilchevski, 2000).

Otro coeficiente de correlación utilizado en la tesis es el propuesto por Maronna (1976) basado en un estimador-*M* (ver Anexo IV.1, pág. 106).

El siguiente **ejemplo** muestra la información que puede obtenerse al comparar correlaciones calculadas con un estimador clásico y uno robusto.
Si se toma la nube de puntos de la Figura IV.19- Sección IV.6.5 y se considera que las variables X e Y poseen la misma categoría (es decir, se tiene un sistema bivariado) es posible calcular el grado de correlación de los datos. Esto se puede hacer de manera clásica utilizando, por ejemplo, el ρ de Pearson o mediante un estimador robusto de correlación, tal como el MCD; ambos pueden aportar información complementaria.
Como primer paso de exploración Maronna (CP) sugiere tener en cuenta que:

a) Si ρ es alto y MCD es alto entonces no se puede sospechar que los datos tengan atípicos ni tampoco no linealidades.
b) Si ρ es bajo y MCD es bajo entonces se puede sospechar de la existencia de no linealidades
c) Si ρ es bajo y MCD es alto se sospecha de la existencia de valores atípicos. Se puede interpretar que estos atípicos estarán en la dirección de menor variabilidad de los datos (o sea, si se imagina una recta que interpola de manera robusta la nube de puntos, estos atípicos estarán en la dirección perpendicular a la recta).
d) Si ρ es alto y MCD es bajo es factible que pueda suceder:
   d1) que existan atípicos en la dirección de mayor variabilidad de los datos o
   d2) que los datos tengan un fuerte comportamiento no lineal en cuyo caso ningún estimador de correlación lineal (como los utilizados) tendrá sentido.

Es posible profundizar este análisis mediante gráficos que permitan visualizar los datos desde otras perspectivas (QQ-Plots, DD-Plots (diagramas distancia-distancia) (Filzmoser, 2004), etc.). Volviendo al ejemplo se encontró que ρ=0.71 y MCD=0.94 (para h=0.8n) lo cual indica que se estaría aproximadamente en las condiciones del ítem c). El lector podrá apreciar en la Figura IV.19 que existen dos puntos alejados de la nube que muy probablemente sean los responsables de la mayor parte de la discrepancia.

IV.2.2 Distancia

La bien conocida distancia Euclídea al cuadrado puede expresarse como:

$D^2_{(x,y)} = \sum_{i=1}^{p} (x_i - y_i)^2$ siendo x e y dos vectores de p variables. Esta distancia "directa" entre dos (o más) puntos es de aplicación generalizada y da una idea fácilmente

interpretable de las diferencias entre vectores.
Dado que los datos de una nube de puntos suelen tener cierto grado de correlación entre sí es útil recurrir a una distancia que considere este hecho. Mahalanobis (1936) propuso una distancia generalizada dada por:

$$D^2_{(x,y)} = (x_i - y_i)\ \Sigma^{-1}(x_i - y_i)^T$$

donde Σ es la matriz covarianza de todos los datos de la nube de puntos que contiene a los vectores *x* e *y*. Esta distancia se halla "pesada" por la covarianza y será menor en la medida en que los datos se hallen más correlacionados entre sí.
Otra forma utilizada para expresar una distancia entre objetos fue la "suma de los valores absolutos de la diferencia" que nombramos como *SAD* ("sum for the absolute values of the differences") que es un caso particular de la distancia generalizada de Minkowsky (Sección V.3- Capítulo V). Esta distancia mide la diferencia entre vectores y al igual que la distancia Euclídea es una medida de disimilitud puesto que cuanto mayor es su valor más diferencia hay entre los vectores involucrados.

$$SAD = SAD_{x,y} = \sum_{i=1}^{p} \left| x_i - y_i \right|$$

donde $x = x_1, x_2, \ldots x_n$ e $y = y_1, y_2, \ldots y_n$ son los vectores entre los cuales se quiere calcular la distancia cuyas variables van desde $i=1$ hasta p dimensiones.
Operando con vectores cuyas variables están dadas en porcentaje y siendo la suma total de ellas 100% para cada vector, el valor dado por el *SAD* resulta fácil de comprender: por ejemplo, un valor de *SAD* de 15 entre dos vectores indicará que los mismos difieren en un 15%. Esto ejemplifica porque, en algunos casos, se prefirió su empleo frente a la distancia Euclídea al cuadrado. Por otra parte, una propiedad de esta medida de disimilitud resultó muy apropiada para trabajar con valores límite. Una simple prueba (Anexo IV.2, pág. 107) muestra que $\left|x_i - y_i\right| \le SAD/2$ para todo *i*, es decir, un valor dado de *SAD* permitirá que la diferencia entre los valores de una misma variable en dos vectores sea como máximo de la mitad, quedando el resto distribuido en la diferencia entre las demás variables. Por ejemplo, un valor de *SAD* = 10 % implica que la máxima diferencia que puede haber en una variable individual *i* cualquiera entre dos vectores sea del 5%.
A partir de lo presentado en esta subsección y la inmediata anterior es posible mostrar un ejemplo sencillo del aporte que pueden realizar los dos enfoques. Las curvas de la Figura IV.7 constituyen patrones a comparar. La Curva 1 tiene buena correlación lineal con la Curva 2 pero se halla a una gran distancia relativa de la misma. La Curva 1 tiene correlación negativa con la Curva 3 (son imágenes casi especulares, lo cual dará coeficientes cercanos a -1) pero sus distancias son relativamente pequeñas.

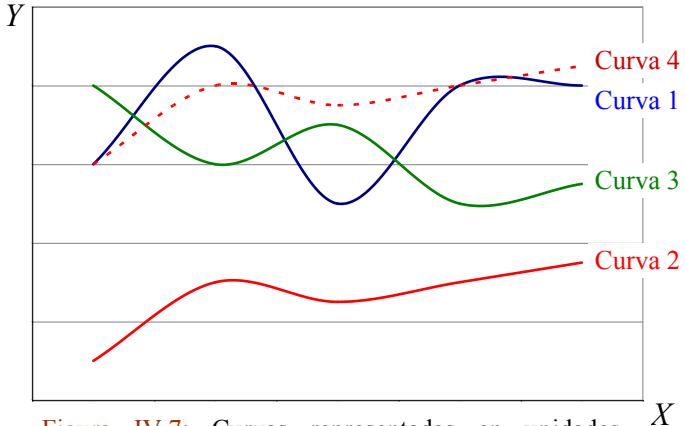


Figura IV.7: Curvas representadas en unidades arbitrarias para mostrar los casos posibles de discriminación utilizando los dos enfoques para estimar similitud o disimilitud entre patrones.

La Curva 1 tiene buena correlación con la Curva 4 y baja distancia.
La Curva 2 tiene baja correlación con la Curva 3 y la distancia entre ambas es alta.
La Curva 2 tiene correlación perfecta con la Curva 4 (una es una combinación lineal de la otra) pero la distancias entre ellas es alta.
La Curva 3 tiene baja correlación con la Curva 4 y la distancia es moderada.

IV.3 Regresión

IV.3.1 Generalidades

El análisis de regresión involucra el estudio de la dependencia entre variables (Weisberg, 2005) y su propósito fundamental consiste en ajustar ecuaciones (modelos) a las variables observadas (Rousseeuw y Leroy, 1987). Tanto el proceso de estimación de los parámetros, como la valoración de lo adecuado del modelo a los datos, se denomina análisis de regresión.

Una curva de regresión describe una relación general entre una o más variables explicativas *(X)* (llamadas también regresores, "carriers" o predictores) y la variable respuesta *(Y)*. Chatterjee y Hadi (2006) desaconsejan llamar a las X variables "independientes" dado a que rara vez lo son (en el sentido de independencia lineal). La función de regresión (Y sobre X) asigna un valor medio a las Y en base a las X. La forma que adquiera la función de regresión da cuenta de lo que se espera para ciertos valores de las X y puede mostrar características entre las X y las Y tales como monotonicidad, unimodalidad, ubicación de los datos en relación al cero y la presencia de valores extremos (Härdle, 1994).

Dado un conjunto de n puntos en el plano (x_i, y_i); $i=1,n$ la relación entre las X y las Y puede plantearse mediante el siguiente modelo:

$$y_i = \beta(x_i) + \varepsilon_i \qquad \text{ec. IV.2}$$

donde:

$\beta(x_i)$ es una función desconocida y

ε_i el término de error aleatorio en las observaciones no incluidas en las x_i

Un diagrama típico de dispersión puede, en algunos casos, dar una idea de la relación entre la variable explicativa y la respuesta (Figura IV.8a), en otros la situación no es tan clara (Figura IV.8b).

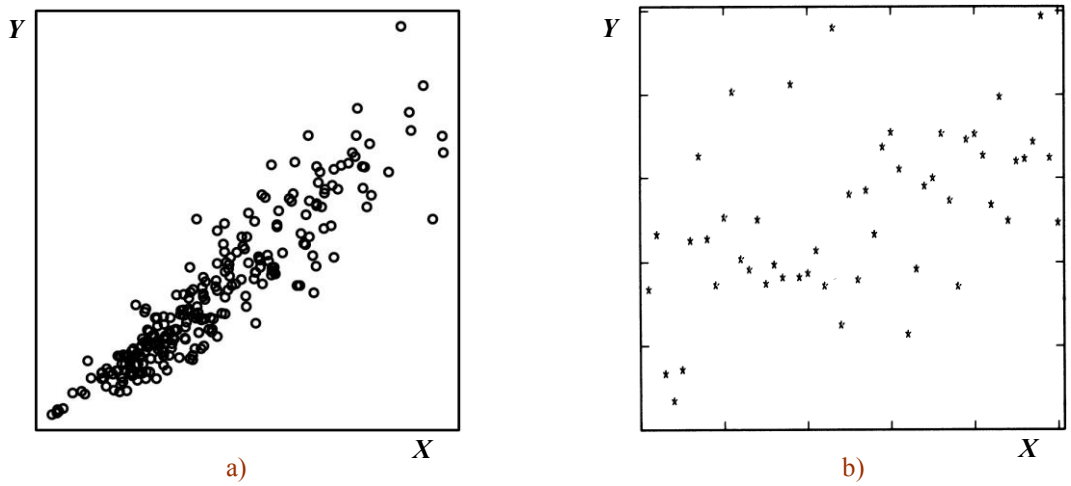


Figura IV.8: a) Diagrama de dispersión tomado de Weisberg (2005) b) Diagrama de dispersión tomado de Cleveland (1979).

En ambos casos es necesario encontrar una relación funcional $\beta(x_i)$ que de cuenta de la dependencia de las Y con las X. Esta relación puede obtenerse de dos formas:

a) paramétricamente: se asume que $\beta(x_i)$ tiene una forma funcional que queda definida por un conjunto de parámetros únicos para describir a todos los datos. Esto puede realizarse a

través de un modelo lineal (recta) o no lineal (polinomio, etc.) en las variables explicativas. A este tipo de regresión suele llamársela regresión global.

b) no paramétricamente: se asume que no existe una única función o familia de funciones que ajusten todos los datos. Se recurre a un conjunto de funciones que, combinadas de manera específica, dan lugar al ajuste o modelado de los datos. Es un ajuste mediante una relación funcional flexible (Härdle, 1994).
Cabe agregar que el término "no paramétrico" suele referir también a métodos que no realizan una suposición explícita sobre la densidad de distribución (por ejemplo de los errores o de la media del predictor).
En esta tesis se recurrió al método de LOESS ("Locally Weighted Scatter Plot Smooth"). Este método es no paramétrico porque los datos no quedan representados por una sola familia de funciones caracterizada por un conjunto de parámetros (tales como la pendiente y ordenada al origen de una recta) sino, por un conjunto de funciones de distintas familias, cada una de ellas ajustando un subconjunto del conjunto total de datos (carácter local).
La sigla LOWESS (que suele aparecer como "sinónimo" de LOESS) denota que el conjunto de funciones son polinomios de grado 1 mientras que el LOESS utiliza polinomios de grado 2 (The MathWork, 2002).
Otros métodos de regresión locales son las "*sp*- lines", las ondeletas, etc. (Loader, 1999; Fox, 2000).
Tanto el enfoque paramétrico como el no paramétrico tienen ventajas y desventajas y la elección depende del caso de estudio; en algunos casos es posible compararlos (Härdle y Mammen, 1993). También existen modelos mixtos llamados semiparamétricos (Härdle, 1994).
Cohen et al. (2003) presentan un ejemplo donde se comparan una solución paramétrica con una no paramétrica para la misma nube de puntos (Figura IV.9). El eje de las *X* representa los años transcurridos luego de obtener el doctorado mientras que el eje de las *Y* el salario.

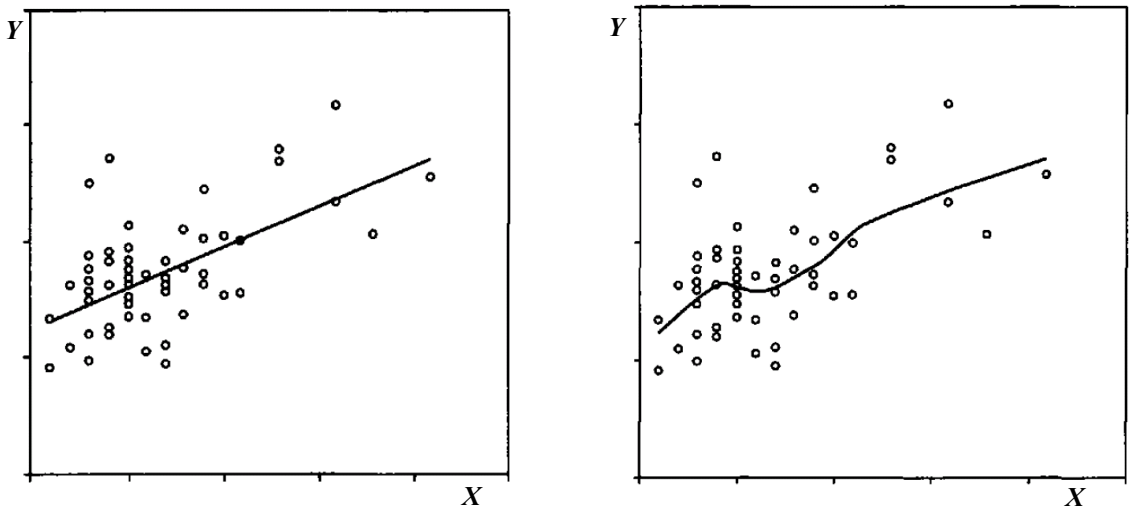


Figura IV.9: a) Regresión lineal simple. b) Regresión no paramétrica realizada con LOWESS. (ambas tomadas de Cohen et al. (2003) Capítulo 4.)

Según el autor ambas soluciones son buenas, la representada en la Figura IV.9a indica que los datos pueden ser representados linealmente, la recta de regresión "caracteriza" a los datos (los salarios aumentan con los años de egreso). La Figura IV.9b deja ver no

linealidades y muestra un "resumen de la tendencia" de los datos (hay un período en donde el crecimiento se estanca para luego reactivarse).
Al emplear métodos de regresión no paramétrica se debe tener en cuenta que la curva obtenida es menos confiable en los extremos (Cohen et al., 2003).
En general, al aplicar un método de regresión paramétrica lo que se busca son los parámetros de la regresión y se evalúa, según la necesidad, la bondad de ajuste. Cuando se aplica un método no paramétrico el énfasis está puesto en el patrón general de la curva obtenida (no en los parámetros de cada porción de la curva) y se evalúa cuan bien queda suavizada la variable respuesta.

En la presente tesis se han aplicado principalmente la regresión lineal simple (utilizando el método de cuadrados mínimos ordinario y un método robusto basado en el estimador-*S* - Sección IV.6.5) y la regresión no paramétrica (utilizando polinomios de segundo grado- Sección IV.3.3.1).

IV.3.2 Regresión global

El modelo clásico lineal paramétrico que regresiona las "*Y*" sobre las "*X*" (se estiman las *Y* a partir de las *X* porque se considera que estas "explican" a las *Y*) busca determinar un conjunto de parámetros que ajusten a todos los datos de la muestra de trabajo (x_i, y_i); $i=1, n$ y asume la forma:

$$\hat{y}_i = \beta x_i + \varepsilon_i \;\; ; \;\; i=1, n$$

donde

$\hat{y}_i$ es el valor de la variable respuesta dada por el modelo.

β es un vector que contiene a los parámetros (ordenada al origen y pendiente para el caso bidimensional).

n es el número de datos (o tamaño de muestra).

x_i es un valor posible que adopta la variable explicativa.

ε_i es el término de error (no incluido en las variable explicativa).

El residuo se define como $r_i = y_i - \hat{y}_i$. El método de cuadrados mínimos ordinario busca encontrar los parámetros β de tal manera que se minimice la sumatoria del cuadrado de los residuos (mín $\sum_{i=1}^{n} r_i^2$). Este método ha sido la piedra angular del análisis de regresión lineal de la estadística clásica, fue creado alrededor del año 1800 por Gauss quien más tarde le asignó al término ε_i el supuesto de distribución normal (Cook y Weisberg, 1999) y es, aún hoy en día, el método más difundido. Un análisis detallado de las suposiciones que implica este método se halla en Belsley et al. (2004). La popularidad del método de cuadrados mínimos se debe, entre otras razones, a que es explícito (los parámetros se obtienen directamente por álgebra de matrices) y es consistente desde el punto de vista teórico (Cook y Weisberg, 1999). Sin embargo, es muy sensible a la presencia de valores atípicos (Hoaglin et al., 1983; Rousseeuw y Leroy, 1987; Maronna et al., 2006).
Una alternativa robusta (estimador-*S*) fue empleada en una aplicación de esta tesis (Sección IV.6.5) para la determinación de los coeficientes de regresión (Ratto et al., 2009).
Dada la complejidad del método y del cálculo el lector interesado puede recurrir a los textos de Maronna et al. (2006) y Filzmoser et al. (2009) en donde además se establecen comparaciones con otros estimadores robustos.

IV.3.3 Regresión local

Los métodos de regresión local han sido desarrollados como una extensión de los métodos paramétricos y deben su solidez a la teoría subyacente en ellos (Loader, 1999). Con

antecedentes en el siglo XIX han proliferado a partir de Cleveland (1979) y Cleveland y Devlin (1988).

Estos métodos se utilizan en general para obtener curvas suavizadas. En el ajuste típico el modelo paramétrico trata de ajustar lo mejor posible todos los datos (por ejemplo, minimizando los residuos, como se ha discutido en secciones anteriores) mientras que en el suavizado se trata de lograr una relación óptima entre el grado en que el modelo se aproxima a los datos y cuanto se puede disminuir el ruido. Logrado este balance el modelo no paramétrico posibilita detectar patrones subyacentes en los datos que la nube de puntos no dejaba ver. La Figura IV.10 muestra los mismos datos que la Figura IV.8b que han sido suavizados por regresión local.

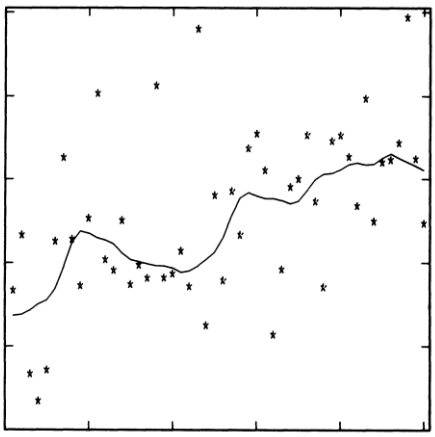


Figura IV.10: Diagrama de dispersión y curva de suavizado tomado de Cleveland (1979).

En una regresión local existen varios componentes que deben ser especificados (Cleveland y Loader, 1996a): a) la ventana de suavizado (llamada también ancho de banda), b) el grado del polinomio local, c) una función de peso y d) el criterio de ajuste.

a) La *ventana de suavizado* (intervalo de las X utilizado para realizar la regresión) tiene un efecto crítico en la regresión local. Si la ventana es muy pequeña habrá pocos datos y el polinomio ajustará bastante bien a cada uno de ellos pero agregará "ruido" produciendo una desvío grande (los valores predichos por el modelo en relación a la media del modelo serán grandes). En este caso el sesgo (dado por la diferencia acumulada entre lo que predice el modelo y los datos) será bajo puesto que la curva estará pasando cerca de cada uno de los datos. Se producirá un sobresuavizado. Por el contrario, si la ventana es muy grande el polinomio no podrá ajustar bien a los datos (el ajuste estará distorsionado -muy aplanado- lo que dará lugar a una sesgo alto). En este caso el desvío será muy pequeño y se producirá un subsuavizado. Por lo tanto y, en términos generales, la ventana deberá adoptarse de tal manera de lograr un compromiso entre desvío (varianza) y sesgo. La razón por la cual es deseable que ambos sean bajos es porque el desvío representa el error aleatorio (da una idea de cuánto se baja el nivel de ruido de los datos) mientras que el sesgo (que da una idea de cuan buena es la aproximación a la función de regresión) representa el error sistemático (debido al modelo elegido) (Loader, 1999). Las ventanas de suavizado se eligen en general como fijas (cuando los datos en las X se hallan equiespaciados) o según una cantidad fija de datos en el eje de las X (vecinos más próximos).

b) El *grado del polinomio* afecta la relación desvío/sesgo. A mayor grado habrá menor sesgo pero más desvío aunque, como se indicó en el párrafo anterior, esto variará según el tamaño de la ventana adoptada. En general los polinomios de grado alto (3 o más) se hacen más inestables en ventanas pequeñas y no producen mucho beneficio (Loader, 1999). Un polinomio de grado cero (constante) da lugar a un promedio móvil pesado (un tipo de estimador "kernel") que suele aplicarse con distintas variantes pero frente al polinomio de grado 1 (recta) o de grado dos es más limitado (NIST, 2012). El polinomio de grado 2 produce menores sesgos que la recta pero aumenta el desvío principalmente en las

fronteras de la ventana (Loader, 1999). Por lo tanto, es tarea del analista observar como es el ajuste.

c) La *función de peso* se elige de tal manera que satisfaga las condiciones de continuidad y simetría, que posea un pico en cero y que esté acotada en [–1,1]. Esta función influirá en la calidad del ajuste de la variable respuesta que es fácil de representar. Existen varias posibilidades pero una elección típica (Loader, 1999) es utilizar la función tricúbica (Anexo IV.3, pág. 108). La idea de trabajar con una función de peso es que para un intervalo dado de la variable explicativa los puntos que se hallan más cerca entre sí se parecen más (entre sí) que los que se hallan más alejados; entonces, los puntos que mejor siguen al modelo local son los que más influyen en la determinación de sus coeficientes.

d) El *criterio de ajuste* implica definir un método, por ejemplo, cuadrados mínimos lineal en los coeficientes, un estimador-*S*, etc. El LOESS que se empleó (Sección IV.6.6) opera con cuadrados mínimos (minimizando los residuos al cuadrado) y se eligió por simplicidad (sigue los mismos criterios que la regresión global aunque también "hereda" su sensibilidad a los valores atípicos). La ventaja de un método basado en regresión local por cuadrados mínimos es que la manera de calcular incertidumbres es la misma que para la regresión global.

Al igual que en regresión global luego de aplicar la regresión local y obtener una curva es posible explorar la bondad de ajuste (por ejemplo, graficando residuos versus predictor) y trabajar con diagnósticos (QQ-Plot, etc.) para ganar más conocimiento sobre los datos de trabajo (Cleveland y Loader, 1996b).

Una de las ventajas de la regresión local es su flexibilidad para adaptarse a datos que no siguen una única curva teórica. Como desventajas puede citarse la necesidad de que los datos sean numerosos (para proporcionar un buen ajuste) y también el hecho de que la curva general obtenida obedece a un conjunto grande de polinomios y es más difícil de transferir a otra personas (NIST, 2012). Esto no sucedería con regresiones no lineales en donde una sola ecuación podrá describir la curva de ajuste.

Una breve descripción del Método LOESS y una forma de detectar tendencias se halla en el Anexo IV.3 (pág. 108).

IV.4 Tendencia

Como señala Simth (2001) el estudio de las tendencias en relación a los gases ambientales observados en una red de monitoreo es un tema que abarca muchos aspectos y los métodos estadísticos cumplen un rol fundamental. En la presente tesis el tema quedó circunscripto a los datos de trabajo, motivo por el cual se describen los métodos adoptados (gráficos y tests) en las secciones correspondientes.

IV.5 Misceláneas

Una variada gama de métodos gráficos fueron utilizados tanto con fines de representación (diagramas de dispersión, series en el tiempo, rosetas de concentración de contaminantes, rosetas de frecuencias de dirección, etc.) como con el objetivo de explorar observaciones y resultados (histogramas, QQ-Plots, etc.). Cada uno de estos recursos se halla explicitado en las secciones correspondientes.

IV.6 Aplicaciones

IV.6.1 Mediciones de SO_2 entre 1996 y 2000

La Figura IV.11 muestra las concentraciones anuales promedio de SO_2 entre 1996 y 2000 y el promedio general (14 ppbv) de los cinco años. También se muestran los valores dados por el lineamiento de la OMS (Organización Mundial de la Salud) del año 2000 (WHO, 2000a) y el límite dado por el Decreto Reglamentario 3395/96 de la Ley N° 5965 de la Provincia de Buenos Aires.

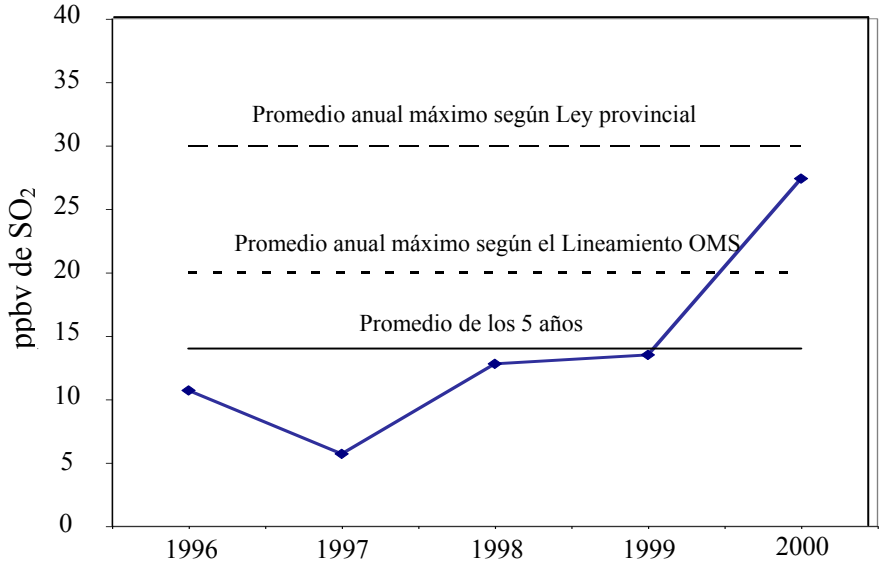


Figura IV.11: Promedio anuales de SO_2 observados en el Punto A (Figura II.6- Capítulo II). Las líneas horizontales muestran el promedio general observado para los años de estudio (línea llena) y los valores límite según distintos referentes.

De una primera inspección surge que los promedios anuales no sobrepasaron el valor de la ley pero sí el lineamiento OMS en el año 2000.

Es importante considerar que una exposición al SO_2 como la que indica el promedio general estará acompañada de la presencia de otros agentes contaminantes (con mayor o menor presencia según la subzona (Colombo et al., 1999; Bilos et al., 2001; Whichman et al., 2009)) provenientes de distintas fuentes (Sección I.1.1- Capítulo I y Sección II.2- Capítulo II) tales como la industria, el parque automotor, la generación de energía y la actividad portuaria. Esta suposición se ve fortalecida por el hecho de que tanto el PM_{10} como el $PM_{2.5}$ son considerados los contaminantes de mayor preocupación en diversas ciudades de América Latina (CAI, 2012). El estudio de Jedrychowski et al. (1999) muestra como valores promedio anuales de 17 ppbv de SO_2 en presencia de material particulado total en valor promedio anual de 52.6 $\mu g/m^3$ (microgramos por metro cúbico) tienen una incidencia significativa en el crecimiento de los niños preadolescentes en Cracovia (Polonia). Según US ATSDR (1998) promedios anuales de SO_2 de 10 ppbv en presencia de material particulado tienen impacto sobre las enfermedades respiratorias. Por su parte, Colombo et al. (1999) encontraron que la presencia de material particulado total en el casco urbano de La Plata se hallaba entre 78 y 219 $\mu g/m^3$ durante una campaña de 7 meses. En OMS (2006) se indica que hay pocas pruebas que indiquen un umbral por debajo del cual no quepa prever efectos adversos sobre la salud por parte del material particulado.

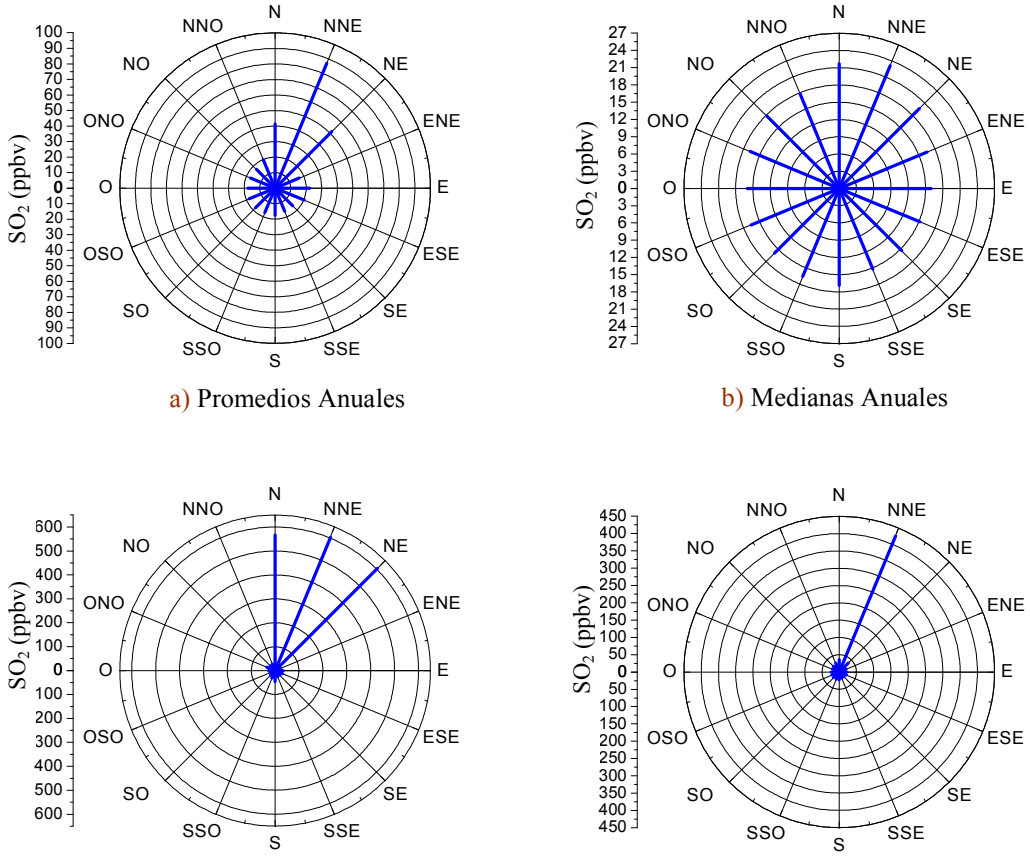
Observando la evolución de las concentraciones de la Figura IV.11 es posible apreciar una tendencia creciente. Con el fin de evaluar, con criterio estadístico, la correlación que existe entre los promedios anuales y la secuencia en que se observaron se aplicó el test de Daniel

-recomendado en WHO (1980) - que se basa en el coeficiente de correlación de rangos de Spearman (Sección IV.2.1). Este test permitió verificar con un 95% de confianza ($\alpha = 0.05$ -test de 1 cola) la H_0 (hipótesis nula) de tendencia creciente de las concentraciones anuales de SO_2 para el período de estudio (Ratto et al., 2006).
En relación a los materiales y objetos culturales Kim et al. (2004) muestran como concentraciones bajas de SO_2 (por debajo de los 10 ppbv) producen distinto grado de impacto. El acero al carbono es uno de los materiales más sensibles a la presencia de SO_2 en el aire, puesto que presenta una tasa de corrosión de 1.446 µm/año/SO_2(ppbv) frente por ejemplo a la de 0.039 del bronce.
Por lo tanto, dados que durante el período de análisis la mayor parte del tiempo se superan las 10 ppbv de SO_2, que los promedios anuales tienen una tendencia creciente, que es verosímil la existencia de otros contaminantes del aire (debido a la intensa actividad industrial y vehicular y a la evidencia proporcionada por otros estudios (ver final de Sección I.1.1- Capítulo I) y que en particular existen registros que dan cuenta de la presencia de material particulado en cantidades significativas, es posible sentar un precedente de la situación de deterioro de la calidad del aire (con efecto sobre la salud y los materiales) en La Plata y alrededores.

IV.6.2 Rosetas de concentración del año 2000

El año 2000 fue el más completo en datos de SO_2 en el Punto A (Figura II.6- Capítulo II) por lo que se procedió a determinar, a modo de ejemplo, la importancia de la presencia de este gas como testigo de actividad industrial. Se discute la metodología en vistas de su aplicación a bases de datos más largas.
Las rosetas de concentración son recursos gráficos en donde se combina información meteorológica (direcciones de viento) con información ambiental (concentración de contaminantes). Estas representaciones se utilizan frecuentemente para ayudar a detectar e identificar fuentes de emisión (WHO, 1980; Henry et al., 2002; Ragosta et al., 2002; Rigbi et al., 2006).
La Figura IV.12 muestra rosetas de concentración elaboradas utilizando distintos estimadores muestrales para el conjunto de datos ya descripto. Una vista panorámica de estas gráficas permite apreciar varios aspectos:
La Figura IV.12a indica que la mayor parte de las concentraciones (evaluadas con el promedio) se hallan por debajo de 20 ppbv. Las direcciones NE y NNE poseen valores considerablemente mayores que el resto. Esto puede deberse a la presencia de máximos influyentes o que esas direcciones sean las que transportan de forma continua niveles más altos de SO_2.
La Figura IV.12b muestra que las medianas están en su mayoría cercanas a las 18 ppbv, exceptuando las direcciones N, NNE y NE en donde se observan valores más altos. Siendo la mediana más robusta que la media, los valores altos de concentración de SO_2 en las direcciones citadas de la Figura IV.12a deben ser considerados como verdaderos (y no valores "inflados" por la presencia de algún potencial atípico).
La Figura IV.12c muestra los valores máximos encontrados en el período según la dirección. En relación a las figuras anteriores esto pone en evidencia que, además de mayor carga continua proveniente del N, NNE y NE, esas direcciones poseen verdaderos "picos" de concentración.
La Figura IV.12d permite consolidar la idea de que existe al menos una dirección "dominante" (el NNE), en la cual el 90% de los datos son menores que 425.6 ppbv pero que hay un 10% de los datos que indican concentraciones algo mayores aún. Es conveniente recordar aquí que, tal como se mencionó en el Capítulo II (Sección II.3.2), los registros efectuados de la dirección NNE dados por la estación meteorológica son algo

defectuosos (esta dirección aparece con frecuencias bastante más bajas que sus vecinas), lo cual produce que haya menos datos en esta dirección a costa de que haya más en las direcciones vecinas inmediatas. Al mismo tiempo, esto permite evidenciar que con pocos registros las concentraciones provenientes del NNE tienden a ser elevadas.



a) Promedios Anuales

b) Medianas Anuales

c) Máximos Anuales

d) Percentil 90

Figura IV.12: Rosas de concentración para el año 2000 observadas en el Punto A de monitoreo. Para cada dirección de viento se acumulan las concentraciones de SO_2 durante el año. Cada dirección implica la dirección desde donde sopla el viento. Luego en cada una de esas direcciones es posible calcular distintos estimadores: a) la media, b) la mediana (o Percentil 50), c) el máximo y d) el Percentil 90 (el 90% de los datos están debajo de determinado valor).

La combinación de estas cuatro gráficas (podría haber otras que tengan en cuenta otros parámetros), permite explorar los datos y concluir (dentro del alcance que permite un año de mediciones) que existen direcciones dominantes en relación al SO_2 y que tales direcciones (observadas desde el Punto A de monitoreo) indican la zona de procedencia.

Es frecuente asociar la contaminación del aire observada en sitios concretos a grupos de direcciones de viento (Cheng y Lam, 1998; Goyal et al., 2002). Inspeccionando la Figura II.6 (Capítulo II) por simple geometría es esperable que las direcciones de viento halladas utilizando las rosetas de concentración, sean las principales responsables de las concentraciones observadas. Si a las direcciones N, NNE y NE se le agrega -siguiendo la geometría y para completar las direcciones más probables por su impacto- la dirección NNO queda conformado un grupo de direcciones (NNO- N- NNE- NE) relevantes para el seguimiento de las concentraciones en el Punto A. A este grupo de direcciones se lo llamó "sector de interés" (Ratto et al., 2006) y Sector 1 en sucesivos reportes.

Queda definido así un grupo de direcciones de viento que será muy importante para estudiar el transporte de los contaminantes desde el área industrial hacia el casco urbano.

La asignación de las fuentes industriales como únicas causantes de las altas concentraciones observadas de SO_2 puede justificarse, en principio, por las características de la zona industrial de La Plata y alrededores (Capítulo II). Esta observación se refuerza por el hecho de que las rosetas de la Figura IV.12 no indican direcciones fuera del Sector 1 con concentraciones llamativas. Ha de consignarse que en las inmediaciones del Punto A hay avenidas de alto tráfico, en particular la Avenida 60 que dista aproximadamente 30 metros en dirección SE (sudeste) de donde se hallaba instalada la unidad analizadora de SO_2. En relación al aporte vehicular ha de tenerse en cuenta por un lado el bajo contenido de azufre de las gasolinas en Argentina (IAA, 2006); Aramendía y colaboradores (Bogo et al., 1999) encontraron valores de SO_2 promedio entre 2 y 7 ppbv en 21 sitios de monitoreo entre Mayo y Julio de 1994 en la Ciudad Autónoma de Buenos Aires (una ciudad con aprox. 3 millones de residentes pero con un intenso tráfico vehicular, cuyo principal aportante es un conurbano de más de 9 millones de habitantes al 2001). Por otro lado, debe tenerse en cuenta que el combustible diesel en Argentina tenía entre 1500 ppm (máx. para vehículos pesados hasta el 2012) y 750 ppm de azufre en promedio (500 ppm en promedio a partir de 2012) lo cual implica valores atendibles de inmisión (Dawidowski, CP). Por lo tanto la discriminación de los aportes de los vehículos diesel podría ser un tema de futuras investigaciones.

En relación al Puerto de La Plata ubicado a aprox. 8 km del Punto A (Punto M - Figura II.6 - Capítulo II) no se encontraron registros de SO_2 pero mediciones de material particulado total y metales en aire (Colombo et al., 1999; Bilos et al., 2001) muestran que dicha zona debe ser tenida en cuenta para el monitoreo sistemático.

Otra fuente potencial de SO_2 la constituyen los aeropuertos (Yu et al., 2004) pero en el caso local (aeropuerto de baja circulación) no se conocen estudios. Además dados su tamaño y ubicación en conjunto con su distancia al Punto A y los vientos dominantes, las emisiones del aeropuerto no puede considerarse una fuente importante a ser detectada desde el Punto A.

Una limitación potencial de las rosetas de concentración la constituye el hecho de que se asume que el viento que se observa en el sitio de monitoreo es muy similar al viento en el que se sumergen las especies contaminantes a partir de sus fuentes (Cosemans et al., 2008). Esto puede no cumplirse debido a: fuentes altas, turbulencia importante u obstáculos entre la fuente y la zona de observación. Otra limitación de estas rosetas es que las mismas no podrían distinguir, para una dirección dada, si un pico registrado se debe a una pequeña fluctuación en una fuente cercana o a una gran fluctuación en una fuente alejada.

Tanto la influencia de algunas fuentes altas (que se hallan presentes en la Refinería, Punto E de la Figura II.6- Capítulo II) como las características de la turbulencia en la zona constituyen un motivo de futuros estudios, que darán más información para el modelado. Sin embargo, la información proporcionada por las rosetas de concentración resulta valiosa dado que no hay obstáculos entre las fuentes y la zona de observación y que no existían en el período de estudio otras fuentes importantes de emisión de SO_2. Cabe aquí considerar que en 2012 se puso en marcha una gran central de generación de energía (Sección II.2- Capitulo II) en las vecindades del complejo industrial de Ensenada (Punto L de la Figura II.6- Capítulo II), situación que refuerza desde entonces, la necesidad de realizar el seguimiento de este gas en el casco urbano y sus alrededores.

IV.6.3 Similitud y disimilitud entre direcciones de viento observadas en distintos sitios

El objetivo de esta sección es realizar una comparación entre curvas horarias de direcciones de vientos observadas en los puntos A y J durante el período 1998- 2003 (tiempo en que ambas estaciones registraron datos simultáneos con buena completitud). Se emplean dos enfoques, uno de orden cualitativo basado en la inspección visual y otro de

orden cuantitativo basado en el uso de dos herramientas: la correlación (similitud) (Sección IV.2.1) y la distancia (disimilitud) (Sección IV.2.2). Ambos enfoques buscan obtener un conocimiento más profundo de las observaciones (respecto del que se pueda obtener con cada uno por separado) y proveer una base para explicar los fenómenos físicos involucrados.

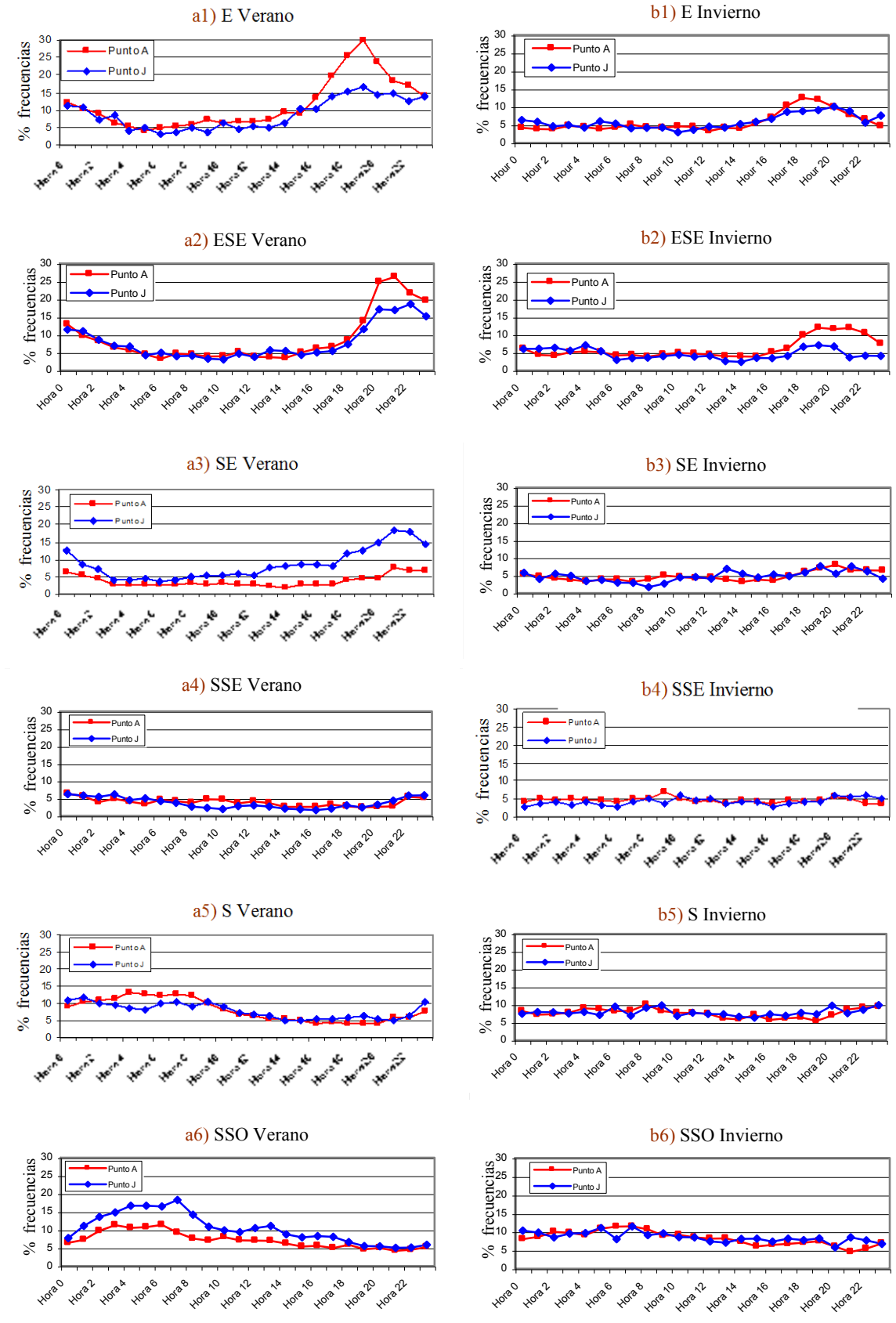


Figura IV.13 (continúa en la página siguiente).

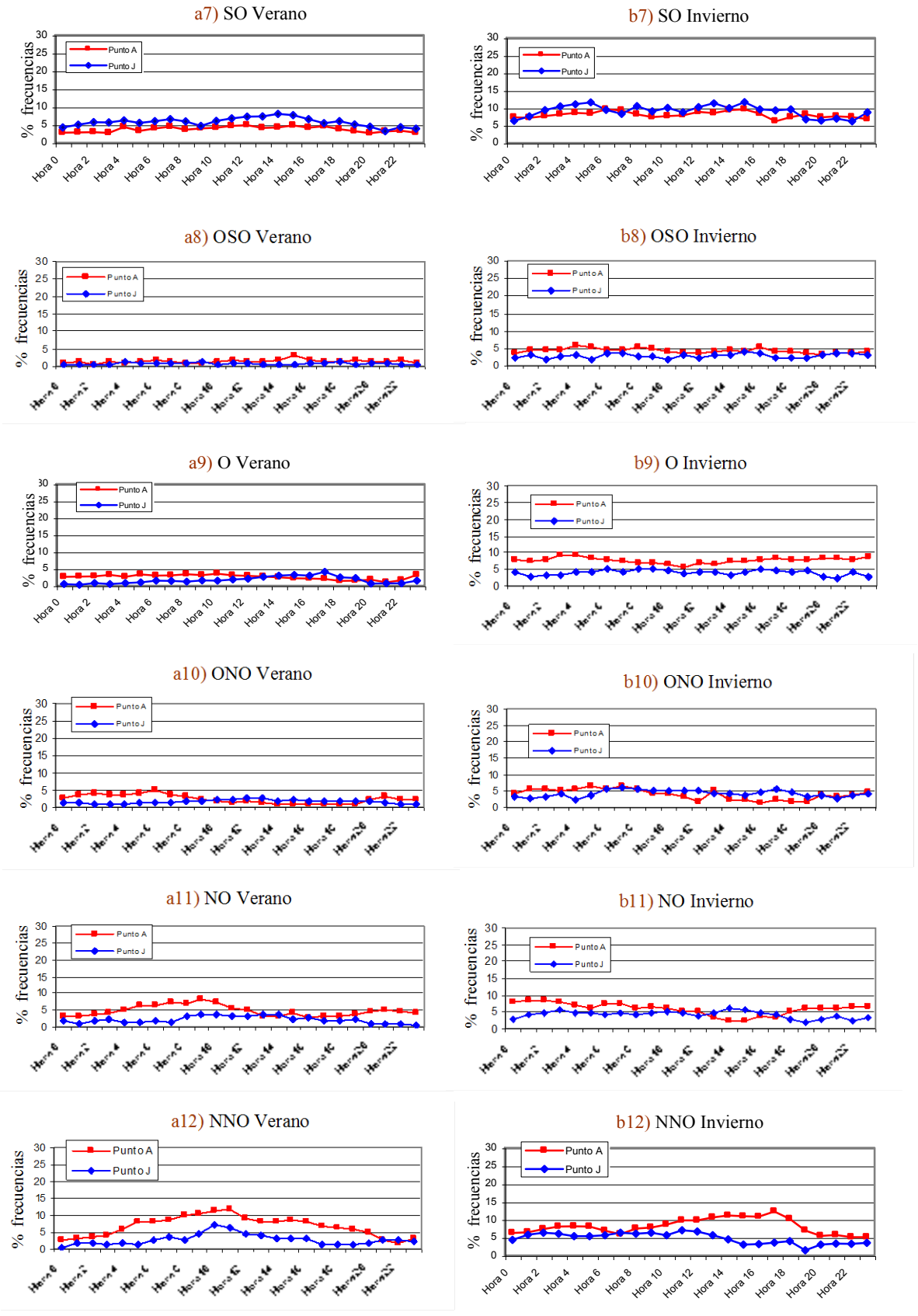


Figura IV.13 (continúa en la página siguiente).

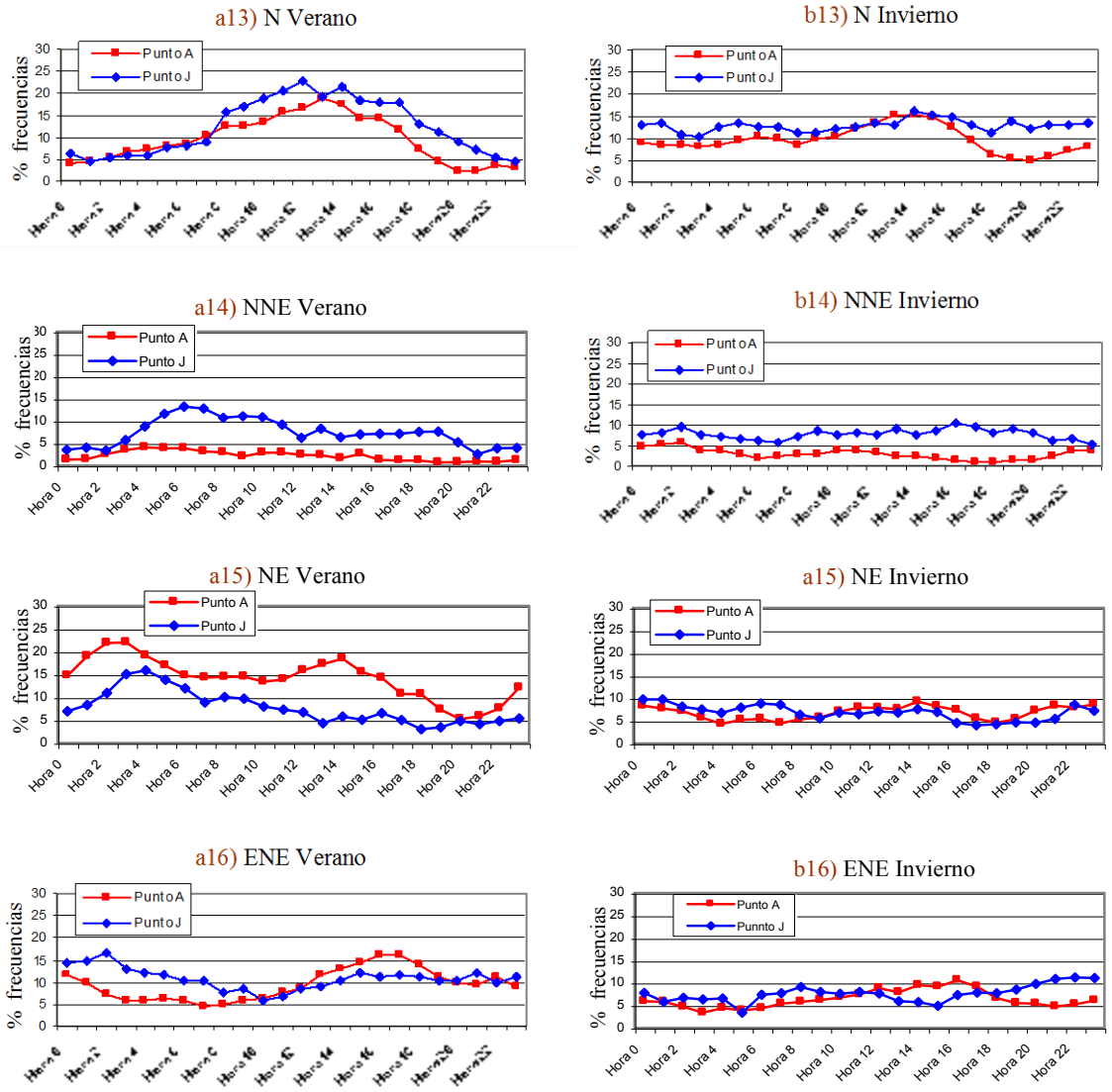


Figura IV.13: Frecuencias acumuladas observadas durante el período 1998- 2003 promediadas por hora en los puntos A y J de monitoreo para la estación verano (a1 a a16) e invierno (b1 a b16) para las 16 direcciones de viento adoptadas. El eje *Y* indica el porcentaje de ocurrencias para una dirección y hora del día particulares respecto del total de ocurrencias para la hora en particular (o sea, la suma de las frecuencias para una hora dada a lo largo de una estación da 100%). El eje *X* indica la hora del "día" en Hora Local (según lo indicado en el Capítulo II- Sección II.3.2).

Considerando la carencia de estudios sobre las ocurrencias horarias de los vientos de la zona de La Plata, el fenómeno de las brisas de mar y tierra parece ser la única fuente local de variabilidad atmosférica (Berri et al., 2010). Este fenómeno fue descripto de manera general en el Capítulo III (Sección III.10), en donde se señaló su importancia en relación a los contaminantes del aire. Es lógico considerar que su mayor influencia sea en verano cuando hay más contraste de temperatura entre la tierra continental y el gran cuerpo de agua que representa el Río de La Plata. Por esta razón, se considera al verano como la estación "líder" para llevar a cabo el análisis.

Se realizaron las gráficas (curvas de evolución horaria según cada dirección de la brújula) correspondientes a las cuatro estaciones del año para los dos sitios de monitoreo y se observó por inspección que, en general, ambas sitios muestran patrones similares. Por cuestiones de espacio solo se muestran (Figura IV.13) las gráficas correspondientes al verano y al invierno (estaciones extremas) pero cabe aclarar que las restantes (otoño y primavera) muestran patrones intermedios.

a) Análisis por inspección visual

Una vista panorámica de la Figura IV.13 (y de las gráficas no mostradas del resto de las estaciones) permite reconocer que las frecuencias del E (por ejemplo, Figura IV.13 (a1)), del N (por ejemplo, Figura IV.13 (a13)) y del NE (por ejemplo, Figura IV.13 (a15)) son más altas que las del resto de las direcciones a través de las estaciones del año. Esto es consistente con las observaciones realizadas en el Aeropuerto de La Plata (Punto K en Figura II.6- Capítulo II) durante las cinco décadas comprendidas en el período 1961- 2010. Según lo visto en la Sección II.1.2 estas tres direcciones dominantes (Figura II.4c) se originan en el flanco oriental del anticiclón subtropical del Atlántico Sur y son importantes para toda la cuenca del Río de La Plata. Durante la noche (entre las horas 0 y 8) las direcciones S (por ejemplo, Figura IV.13 (a5)) y SSO (por ejemplo, Figura IV.13 (a6)) tienen frecuencias de ocurrencias mayores que durante el resto del día. Esto es atribuible a la brisa de tierra por ser estas direcciones algo perpendiculares a la línea costera. Durante las primeras horas de la mañana estas direcciones decrecen notablemente y, en la medida que lo hacen, las direcciones N (por ejemplo, Figura IV.13 (a13)), NNE (por ejemplo, Figura IV.13 (a14)) y NE (por ejemplo, Figura IV.13 (a15)) comienzan a ganar importancia (tener en cuenta que los bajos valores del NNE se deben a mediciones defectuosas- Sección II.3.2- Capítulo II). Estas tres direcciones se hallan relacionadas con la primera etapa del desarrollo de la brisa marina que ocurre cuando el viento comienza a soplar desde el río hacia la tierra. Luego, el viento incrementa su componente N (Berri et al., 2010). Los vientos de la brisa marina siguen un patrón rotacional (Simpson, 1994) en dirección de las agujas del reloj. Esta observación coincide con el estudio preeliminar de Borque et al. (2008) donde se detecta rotación del NE al E entre el mediodía y el atardecer. En una segunda etapa del desarrollo de la brisa marina se observa el decrecimiento de las direcciones N y NE desde la Hora 16 en adelante (por ejemplo, Figura IV.13 (a13) y Figura IV.13 (a14)) mientras que el ENE (por ejemplo, Figura IV.13 (a16)), el E ((por ejemplo, Figura IV.13 (a1)) y el ESE (por ejemplo, Figura IV.13 (a2)) se van volviendo dominantes hasta que alcanzan un pico durante el anochecer (alrededor de las horas 20 y 21).

Las diferencias observadas en las direcciones debidas a la brisa de tierra entre los puntos A y J son más pequeñas que las debidas a la brisa de mar. Es esperable que esto sea así, debido principalmente, a las estabilidades nocturnas (Berri et al., 2010) aunque se puede incluir la rugosidad de la ciudad que puede inhibir el flujo de viento desde tierra adentro hacia el cuerpo de agua. Las direcciones de viento involucradas en la brisa de tierra también aparecen más restringidas (poseen menores porcentajes de ocurrencia en máximos y en promedios) que las direcciones involucradas en la brisa de mar. La penetración de la brisa marina aparece como un tema relevante para ser encarado en futuros estudios.

b) Análisis utilizando distancia y correlación

La inspección visual de la Figura IV.13 indica que las mayores diferencias entre sitios se observan para las direcciones NNE, NE y SE en verano mientras que para NNE, NNO y N en invierno. Una forma de objetivar las diferencias entre observaciones es recurrir al cálculo de la distancia Euclídea al cuadrado (D_E^2); los resultados se muestran en la Tabla IV.1. Esta métrica provee una estimación general de diferencias entre patrones pero no distingue si las mismas están concentradas en unas pocas horas o si se hallan distribuidas a lo largo del día. Por lo tanto, las D_E^2 más grandes entre curvas son analizadas individualmente buscando la hora del día en que la diferencia se hace máxima, de esta manera es posible enriquecer el abordaje por distancias.

El NE y el NNE tienen distancias relativamente altas a través de las estaciones, frecuentemente se hallan entre uno y dos desvíos estándar de la media (promedio).

Considerando que el NNE ha sido medido deficientemente en el Punto A (Sección II.3.2) y que las direcciones son porcentuales en las horas, es posible considerar que la distorsión en esta dirección afectará preferentemente a las vecinas, o sea, al NE y al N. Considerando solo estas tres direcciones, los máximos individuales a través del día son 12.9% en el verano a la Hora 13 para el NE (Figura IV.13 (a5)), 8.3% en el otoño a la Hora 12 para el NE, 9% en el invierno a la Hora 16 para el NNE (Figura IV.13 (b14)) y 10.4% en la primavera a la Hora 11 para el NE.

Tabla IV.1

Direcciones	Verano	Otoño	Invierno	Primavera	Promedio
E	506,0	314,3	59,4	279,1	**289,7**
ENE	383,9	72,8	227,7	311,2	**248,9**
NE	1256,2	344,6	89,6	1070,3	**690,2**
NNE	808,2	534,5	652,3	447,0	**610,5**
N	384,4	364,2	393,2	113,7	**313,9**
NNO	464,2	590,2	404,0	431,3	**472,4**
NO	243,2	290,3	191,8	116,3	**210,4**
ONO	79,2	48,4	81,8	52,7	**65,5**
O	62,0	24,9	380,1	51,7	**129,7**
OSO	18,1	22,8	62,6	21,8	**31,3**
SO	112,5	180,6	80,1	454,3	**206,9**
SSO	353,1	62,5	55,1	149,2	**155,0**
S	92,0	110,1	34,0	36,2	**68,1**
SSE	31,8	54,0	28,4	32,2	**36,6**
SE	797,2	617,5	46,5	1138,1	**649,8**
ESE	201,2	220,3	204,2	773,2	**349,7**
Promedio Estacional	**362,1**	**240.8**	**186.9**	**342.6**	**283,0**
Promedio + 1 DE*	**707,8**	**445.9**	**368.5**	**706.3**	
Promedio + 2 DE*	**1053,5**	**651.1**	**550.2**	**1070.3**	

Tabla IV.1: Distancias Euclídeas al cuadrado entre patrones observados en los Puntos A y J de monitoreo cubriendo todas las direcciones de la brújula con una resolución de 22.5°.
(*) DE: desvío estandar

Excluyendo estas tres direcciones (N, NNE y NE) las máximas diferencias generales involucran al SE y S en verano, al SE y NNO en otoño, al ENE y NNO en invierno y al SE y SO en primavera (Tabla IV.1). Excluyendo al N, NNE y NE las máximas diferencias individuales son 13.4% en el verano a la Hora 17 para el E (Figura IV.13 (a1)), 10.3% en el otoño a la Hora 18 para el E, 8.7% en el invierno a la Hora 17 para el NNO (Figura IV.13 (b12)) y 16.4% en la primavera a la Hora 20 para el ESE.
Tal como se describió en la parte **a)** de esta sección y considerando que las direcciones más influenciadas por la brisa marina comprenden NNO- ESE en la dirección de las agujas del reloj, la mayor parte de las diferencias descriptas pueden ser atribuidas a este mecanismo. Según Oke (Oke, 1987) un viento paralelo a la costa, por ejemplo el SE, es esperable cuando la brisa marina decrece; esto solo se observa débilmente. Además, y de forma contraria a la esperada, el SE es más importante en el Punto J que en el Punto A, lo cual sugiere la ocurrencia de algún mecanismo más complejo.
Las direcciones comprendidas entre SSE y NO (en el sentido de las agujas del reloj) son, en general, cercanas entre los dos sitios de monitoreo para todas las estaciones (todos los valores se hallan por debajo de la media general (283,0) (ver Tabla IV.1). Considerando que el área de estudio es una llanura, que la brisa de tierra es débil y que las direcciones involucradas en SSE- NO no se hallan influenciadas por la brisa marina es apreciable una buena similitud entre patrones de ambos sitios.

Desde otra perspectiva y haciendo uso de la correlación entre estas mismas curvas (Figura IV.13) se recurrió al uso del MCD (Sección IV.2.1) cuyos valores se muestran en la Tabla IV.2.

Tabla IV.2

	Verano	Otoño	Invierno	Primavera
E	0,893	0,776	0,294	0,694
ENE	0,272	0,792	-0,083	-0,304
NE	0,522	-0,427	0,468	0,143
NNE	0,878	-0,499	-0,602	-0,106
N	0,958	0,357	-0,018	0,897
NNO	0,793	0,850	0,555	0,795
NO	-0,358	0,129	0,272	0,036
ONO	-0,695	-0,484	-0,242	-0,870
O	-0,606	0,394	-0,202	0,531
OSO	0,163	0,369	0,151	-0,365
SO	0,881	0,885	0,624	0,876
SSO	0,946	0,953	0,686	0,916
S	0,921	0,855	0,393	0,930
SSE	0,904	0,789	0,540	0,717
SE	0,403	0,293	0,219	0,101
ESE	0,897	0,562	0,308	0,741

Tabla IV.2: Valores del estimador robusto de correlación MCD (Sección IV.2.1) calculados utilizando el software *Scout 1.0*. Este estimador ha sido ajustado para h=0.8 lo que implica que se supone que cada submuestra contiene 19 datos sin contaminación (respecto de los 24 datos totales para una dirección dada). O sea, el punto de ruptura tolerará hasta 5 valores atípicos en cada submuestra. Una estimación posterior mostró que el número de potenciales datos atípicos nunca pasó de 3 para los 4 x 16 casos.

Una vista general de esta tabla da cuenta de la existencia de relaciones lineales entre algunos patrones y de relaciones no lineales entre otros. El verano aparece como la estación más correlacionada mientras que el invierno es la menos correlacionada. Los valores negativos del MCD, tales como el correspondiente al NNE en invierno (Figura IV.13 (b14)) indican, predominantemente, que cuando una de las variables crece la otra decrece. Observar que entre las horas 15 y 22 las formas de las respectivas curvas son imágenes especulares una de otra (recordar la Figura IV.7). Valores de MCD cercanos a cero, tales como el ENE para el invierno (Figura IV.13 (b16)) implica que no hay una relación lineal entre patrones (curvas). A través de las estaciones existe un grupo de direcciones de viento entre el OSO y el NO (sentido horario) que se hallan pobre o negativamente correlacionadas mientras que en el grupo SSE-SO (sentido horario) las curvas se hallan altamente correlacionadas.

Teniendo en cuenta ambos criterios de comparación surge que los vientos entre el SSE y el SO (sentido horario) se hallan relativamente próximos y altamente correlacionados a lo largo de las estaciones para ambos sitios. Por otro lado, el NE y el NNE tienen ambos poca proximidad y correlaciones bajas. Además, el NO se halla pobremente correlacionado pero muy próximo mientras que el NNO se halla altamente correlacionado pero la distancia entre sitios es relativamente alta.

Como se expresó anteriormente, se espera que dos sitios ubicados en la llanura produzcan curvas muy similares tanto en proximidad como en correlación lineal. Sin embargo, el Punto A se halla algo más cercano a la costa del río en una zona urbana de edificios bajos mientras que, el Punto J se halla tierras adentro en una zona semi-rural (Figura II.6- Capítulo II) de baja rugosidad de terreno (Sección II.3.3- Capítulo II).

Por cercanía a la costa se espera que el efecto de la brisa marina sea más pronunciado en el Punto A que en el Punto J, además cabe recordar que los registros de ambos sitios tienen diferencia en la calidad de los datos (Sección II.3.2- Capítulo II). Estas tres circunstancias explican, en términos generales, las diferencias observadas. Mientras las distancias muestran un panorama de similitud general las correlaciones muestran un panorama irregular. Esto último implica que, para algunas direcciones de viento en particular, no será

posible "predecir" el patrón horario de uno de los sitios a partir del observado en otro (correlación lineal pobre). Esto debe ser considerado cuando concentraciones horarias medidas en cualquier lugar de la ciudad necesiten ser correlacionadas con las frecuencias de ocurrencia horarias de vientos por dirección según los puntos A o J.

IV.6.4 Concentraciones de SO_2 durante una campaña corta en un sitio alejado de las fuentes y su relación específica con algunas direcciones de viento

Se llevó a cabo una campaña de monitoreo continuo de SO_2 en el CIOp (Punto D de la Figura II.6- Capítulo II) entre el 1 de Septiembre y el 21 de Diciembre de 2005 (92 días de mediciones). La distancia directa entre el Punto D y la zona industrial es de alrededor de 6 km. Puesto que el SO_2 es un gas muy reactivo, el objetivo era determinar cuánto podía encontrarse en un sitio alejado de las fuentes.

La unidad analizadora de gases empleada fue el equipo Lear Siegler ML 9850® mientras que la estación meteorológica fue la Davis Weather Monitor II Euro Version® (Sección II.3.3- Capítulo II). Las mediciones de los parámetros meteorológicos se efectuaron con algunas dificultades técnicas que hicieron que se perdieran algunos registros.

Las mediciones de SO_2 (ppbv) se presentan como promedios diarios y horarios. Estas escalas de tiempo fueron adoptadas a modo de ejemplo para poder comparar con algunos de los estándares (base horaria y/o diaria) y para establecer su relación con las frecuencias de ocurrencias de algunos vientos que adopta como unidad la hora.

Algunos autores (Bencalá y Seinfeld, 1976; WHO, 1980; Gilbert, 1987) señalan que puede esperarse que la distribución de promedios diarios de los contaminantes del aire sea lognormal. Se probó la H_0 (hipótesis nula) de normalidad para α=0.05 mediante el test de rangos "estudentizados" sugerido en el Capítulo 4 de EPA (2006) para muestras de $n \leq 1000$. Se comprobó que los datos siguen a la distribución normal.

Para visualizar esto se graficaron un histograma (Figura IV.14) y un gráfico cuantil- cuantil (QQ-Plot (Figura IV.15)).

En el eje *X* de la Figura IV.14 se representan mediante barras los intervalos de clase mientras que en el eje *Y* se hallan representadas las cantidades de datos que hay en cada intervalo de clase.

La curva continua (roja) es la distribución normal teórica, dada por el software *Statistica 8.0,* a la que los datos aproximan.

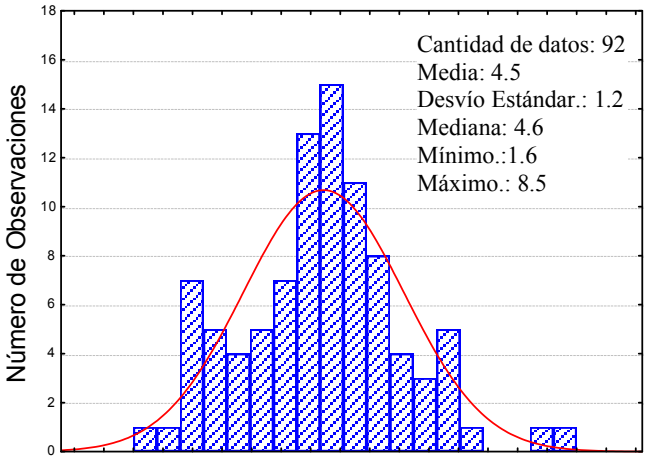


Figura IV.14: Densidad de distribución para las observaciones (histograma) y para la curva teórica ajustada (normal) correspondiente a los promedios diarios de SO_2.

El eje *X* inferior de la Figura IV.15 muestra los valores de la distribución normal estándar (media cero y varianza unidad) que corresponden a cada uno de los percentiles cuyos valores de referencia se muestran en el eje superior de las *X*.
El eje de las *Y* corresponde a los valores (percentiles) de los datos de trabajo. La recta roja continua representa la curva "ideal" que daría si los datos estuvieran perfectamente distribuidos según la distribución normal.
Este diagrama permite apreciar la ausencia de valores atípicos (la metodología se detalla en el Capítulo V- Sección V.5.2.4.1).

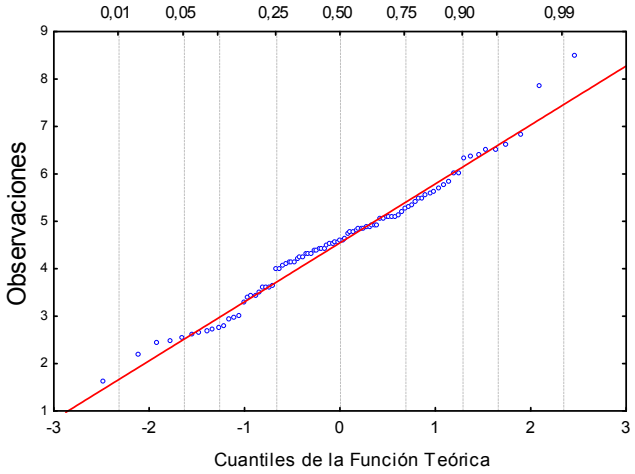


Figura IV.15: Diagrama cuantil-cuantil (QQ-Plot) correspondiente a los promedios diarios de SO_2. Eje *X* inferior: valores de los percentiles de la Distribución Normal Estándar (teórica). Eje *X* superior: percentiles expresados como probabilidad. Eje *Y*: valores observados.

Tanto la Figura IV.14 como la Figura IV.15 permiten visualizar lo demostrado mediante el test, o sea que los datos se distribuyen de forma aproximadamente normal. Esto implica que el promedio (media aritmética) y la varianza (desvío estándar al cuadrado (S_D^2) son buenos estimadores de posición y dispersión (escala) de los datos. Como estos dos parámetros pueden cambiar de una campaña a otra puede resultar útil estimar el intervalo de confianza (IC) de la media. Según WHO (1980) el mismo se pude calcular como:

$$IC = \overline{X} \pm t_{(1-\alpha/2)} \frac{S_D}{\sqrt{n}}$$

donde:
t es el "*t*" de Student.
n es el número de promedios diarios (cantidad de datos).
α es el nivel de significación.

Recurriendo al Anexo II de Gilbert (Gilbert, 1987) para α=0.05 se obtiene un $t_{0.975}$=1.99.
Siendo $\overline{X} = 4.5$ y $S_D = 1.2$ entonces $IC = 4.5 \pm 0.25$ (ppbv)

Este intervalo da un rango en donde se puede encontrar la media si las condiciones en que se miden los datos se mantienen estacionarias.

La Figura IV.16 muestra la serie de promedios diarios (curva a rayas) para la campaña completa (92 días). El valor más alto se da en el día 30 (8.5 ppbv) mientras que el más bajo en el día 71 (1.6 ppbv). Para averiguar si estos valores extremos son atípicos se recurrió al test de Rosner que permite evaluar varios potenciales atípicos al mismo tiempo, se siguieron los lineamientos de cálculo dados en EPA (2006). Para α =0.05 se rechazó la H_0 de la existencia de atípicos.

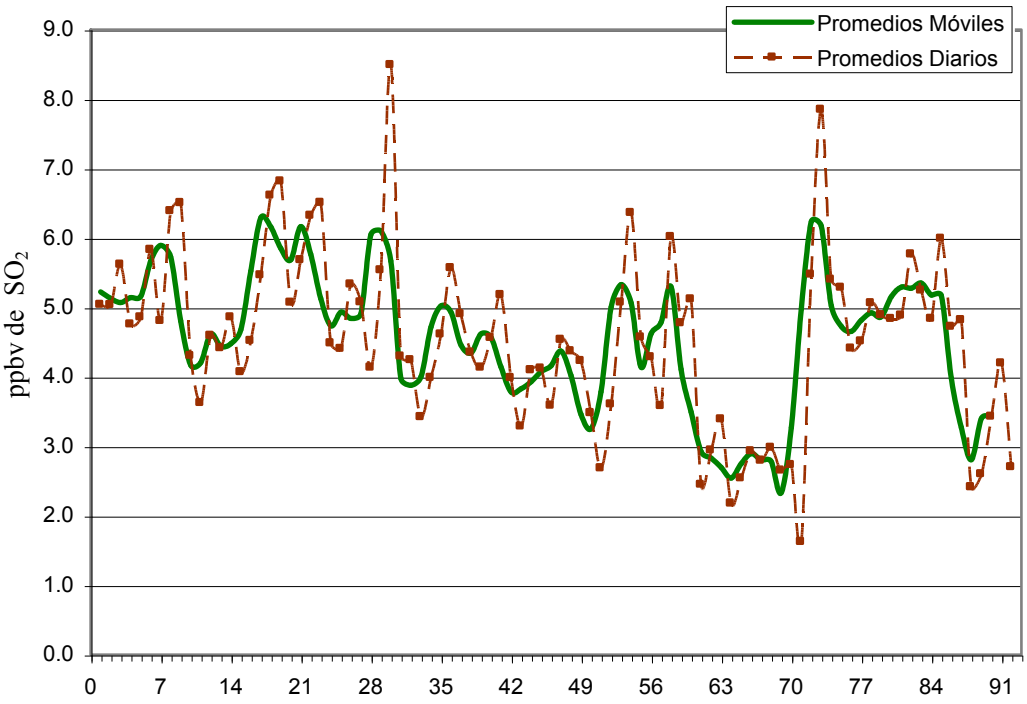


Figura IV.16: Promedios diarios de SO_2 (ppbv) registrados en el Punto D (CIOp) durante una campaña de 92 días (curva a rayas). La curva llena muestra los promedios móviles tomados de a tres días.

Con la finalidad de disminuir el "ruido" de los datos y hacer más visible el patrón subyacente se procedió a suavizar la curva de promedios diarios (curva continua en la Figura IV.16) mediante los promedios móviles tomando como ventana tres días. Los valores suavizados dan cuenta de una especie de "memoria" que tiene el ambiente en relación a la carga de contaminantes (Berthouex y Brown, 2002). Además, el suavizado muestra de manera más clara que los datos parecen tener una tendencia decreciente. Para verificar esta suposición se recurrió al test de Mann- Kendall (Gilbert, 1987). Para α =0.05 se rechazó la H_0 de no existencia de tendencia. La tendencia decreciente puede deberse, por ejemplo, a una disminución en las fuentes emisoras o a un incremento de la humedad pero esto no fue posible de verificar. Otra variable relevante es la altura de capa de mezcla que, si bien no pudo ser evaluada en este trabajo de tesis, presenta estacionalidad. Es decir, dada su evolución característica creciente de invierno al verano (Mazzeo et al., 1971), y por lo tanto su mayor capacidad de mezclado, su crecimiento a medida que se avanza en la primavera hacia el verano, sería un factor a correlacionar con el decrecimiento observado del contaminante.

Los promedios diarios de SO_2 durante toda la campaña estuvieron por debajo de lo establecido por el Decreto Reglamentario 3395/96 de la Ley N° 5965 de la Pcia. de Buenos Aires (140 ppbv) y también por debajo del lineamiento de la Organización Mundial de la Salud (WHO, 2000a) cuyo límite es 48 ppbv (125 $\mu g/m^3$) aunque en dos ocasiones se superó el valor de OMS (2006) -actualización mundial de los lineamientos- que es de 7.6 ppbv (20$\mu g/m^3$). Esto implica que sobre 92 días este límite máximo recomendable fue superado el 2.2 % de las veces, que extrapolado, equivale a aproximadamente 8 días al año en que se supera el valor sugerido por el lineamiento. Una campaña corta (Marzo- Junio de 2010), llevada a cabo en el Punto A (Orte, 2011), mostró un promedio general de 13 ppbv de SO_2 (similar al promedio general de la Figura IV.11) sobrepasando el nivel recomendado por el lineamiento OMS todos los días. Además se detectaron picos cortos de

concentración de 50 y 170 ppbv. Dada la escasez de registros en la zona, estas últimas mediciones sustentan la idea de un factor de dilución observable (entre el Punto A y el D) al mismo tiempo que refuerzan la necesidad de registros permanentes.
La Tabla IV.3 muestra los tres primeros máximos (promedios horarios) encontrados en cada mes de campaña. Según normativa de la EPA (Environmental Protection Agency) de EUA (Lutgens y Tarbuck, 2013) no se sobrepasa el valor límite (75 ppbv) para promedios horarios en ningún caso. Por otra parte y debido a la ubicación del Punto D de monitoreo en relación a las fuentes industriales, es esperable que las direcciones de viento ESE, E y ENE sean las que más estén asociadas a picos de concentración observables. Nótese que los promedios horarios más altos dan cuenta de esta suposición tal como lo muestra la Tabla IV.3.

Tabla IV.3

Septiembre					
Máximos Horarios	**Día N°:**	**SO_2 (ppbv)**	**Hora**	**Dirección prevalente (*)**	**Fecha**
1ro	8	20,3	21	SE- ESE	8
2do	9	16,7	12	ESE	9
3er	3	12,9	11	ESE- E	3
Octubre					
1ro	30	20,3	13	no disponible	20
2do	36	10,1	16	no disponible	8
3er	35	8,5	18	no disponible	7
Noviembre					
1ro	58	18,7	21	ESE	12
2do	59	17,1	17	no disponible	8
3er	63	15,4	17	ENE	17
Diciembre					
1ro	85	25,9	5	E	14
2do	87	23,3	0	E- ENE	16
3er	86	18,3	18	E	15

Tabla IV.3: Registro de concentraciones de SO_2 según el día de campaña, fecha y hora junto a las direcciones dominantes dentro del intervalo horario.
(*) La toma de datos se realizó cada 15 minutos (4 registros horarios); cuando se indica una sola dirección implica que los cuatro registros pertenecen a dicha dirección, en los casos en que hay dos direcciones es porque hubo dos registros de cada una de ellas durante la hora de medición.

Esto implica que la mayoría de las veces en que se observan picos existen vientos provenientes de la zona industrial. Por lo tanto, y de forma análoga a lo realizado en la Sección IV.6.2, puede definirse un grupo de direcciones: ESE-E-ENE, que son de particular interés debido al transporte de los contaminantes industriales. Estos vientos transportan a dichos contaminantes hacia barrios residenciales del área del Gran La Plata. A este grupo de direcciones se lo identificó como Sector 2 (Figura II.6- Capítulo II).
La Figura IV.17 muestra los promedios horarios para cada hora del día para todo el período de campaña. Los picos entre las horas 15 y 18 y entre las horas 21 y 22 se hallan comprendidos entre $\overline{X} + S_D$ (media y desvío estándar) y $\overline{X} + 2S_D$ mientras que el valle de las horas 4 a 7 entre $\overline{X} - S_D$ y $\overline{X} - 2S_D$. Si bien el rango en el eje de las Y es pequeño (aproximadamente 2 ppbv) es posible correlacionar los valores de concentración con las direcciones de viento. Durante los picos prevalecieron los vientos provenientes del E y ESE mientras que en los valles las direcciones dominantes eran del S y SSO (alternándose aunque con menores frecuencias con vientos del E, ENE y ESE).

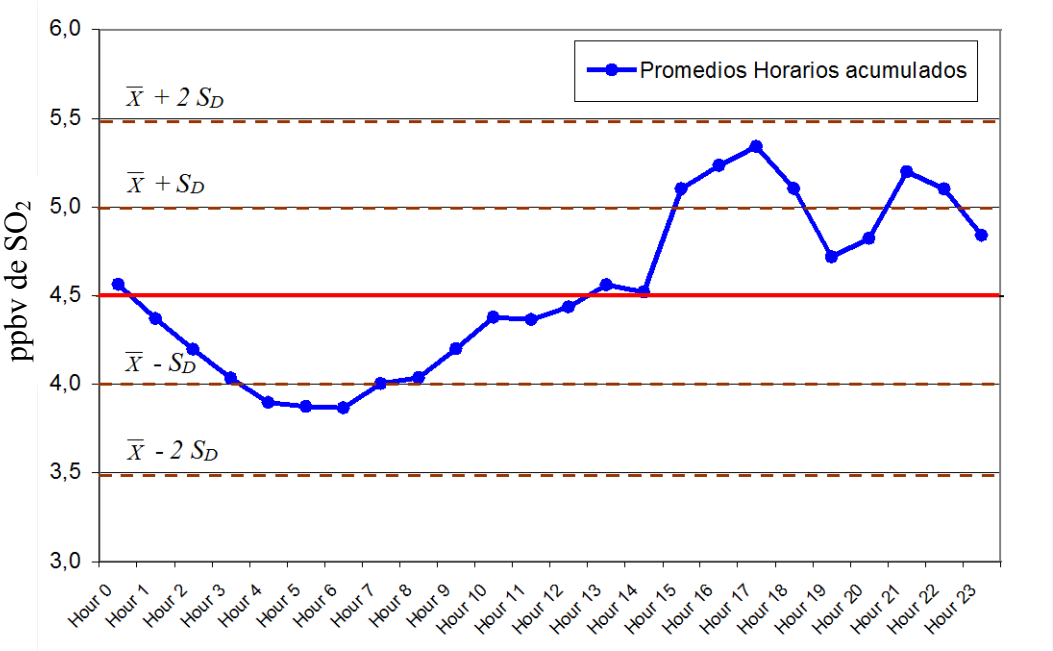


Figura IV.17: En el eje de las *X*, las horas del día implican bloques horarios, por ejemplo Hora 0 (00:00- 00:59 hs.). El eje de las *Y* contiene los promedios de las concentraciones horarias de SO_2 para todos los días de campaña. Se muestran además, con rectas punteadas $\overline{X} \pm S_D$ y $\overline{X} \pm 2\ S_D$. La línea recta horizontal llena (roja) indica el promedio general (4.5 ppbv).

No siendo los datos meteorológicos tomados en el CIOp (Punto D) suficientes como para poder correlacionar las frecuencias del Sector 2 con las concentraciones de la Figura IV.17 se recurrió a registros históricos tomados en otros sitios de la ciudad. La Figura IV.18 muestra los datos de la Figura IV.17 junto a los valores de las frecuencias observadas para el Sector 2 durante las primaveras en los puntos A y en J para distintos períodos.

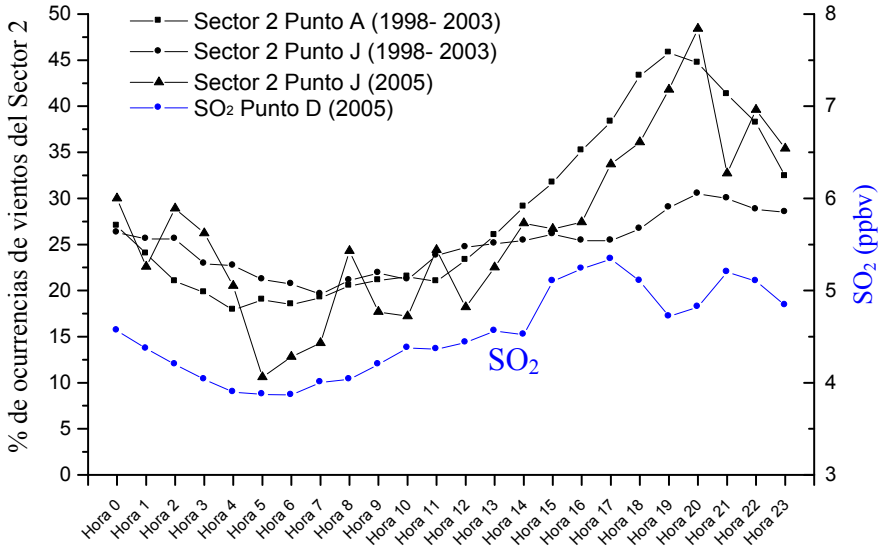


Figura IV.18: El eje *Y* izquierdo refiere a las ocurrencias de vientos del Sector 2 observadas en los puntos A y J en primaveras de distintos períodos. El eje *Y* derecho indica la escala de las concentraciones horarias de SO_2 observadas en el Punto D durante una campaña corta en la primavera de 2005.

A simple vista se observa una buena correlación entre la curva de SO_2 y las correspondientes al Sector 2 (ENE- E- ESE) en los distintos períodos de tiempo y sitios. Para objetivar el grado de correlación se recurrió al cálculo del MCD cuyos valores se muestran en la Tabla IV.4.

Tabla IV.4	
Correlaciones entre frecuencias del Sector 2 y concentraciones de SO_2	
Sector 2 2005 (Punto J) - SO_2 (Punto D)	0.813
Sector 2 1998- 2003 (Punto A) - SO_2 (Punto D)	0.967
Sector 2 1998- 2003 (Punto J) - SO_2 (Punto D)	0.916
Sector 2 1998- 2009 (Punto J) - SO_2 (Punto D)	0.926

Tabla IV.4: Valores de MCD obtenidos al correlacionar concentraciones de SO_2 observadas en el Punto D durante la primavera de 2005 con frecuencias de vientos del Sector 2 en distintos sitios y escalas de tiempo correspondientes a primaveras. Notar que en esta tabla se agrega información (última fila), respecto de la Figura IV.18, para enriquecer el análisis.

Estos valores son lo suficientemente altos como para dejar ver el carácter lineal que tienen las concentraciones de SO_2 (originadas en la zona industrial de Ensenada) observadas en el Punto D con los vientos del Sector 2.
Por otra parte, al correlacionar las curvas de frecuencias del Sector 2 visto desde el Punto A y desde el Punto J durante el período 1998- 2003 se encontró que el MCD era de 0.795, valor alto si se comparan con los hallados para las direcciones individuales que forman parte de este sector (Tabla IV.2). Esto último sugiere, a la luz de las observaciones analizadas, que los vientos del Sector 2 poseen un patrón más generalizable (que las direcciones que lo componen) espacialmente.
Con el objeto de proveer un contexto a las direcciones de viento y las concentraciones analizadas, caben agregar, algunos valores de las velocidad involucradas: las velocidades observadas correspondientes al Sector 2 durante primavera en el Punto A (1998- 2003) fueron de 8.2 km h^{-1}, en el Punto J (1998- 2003) de 7.5 km h^{-1} y en el Punto J (1998- 2009) de 8.0 km h^{-1}. Puesto que el Punto A registró a 12 m y el Punto J a 5 m de altura las correcciones llevadas a 10 m hacen que las velocidades sean muy similares. Un tratamiento más completo se discute en Ratto et al. (2012b).

IV.6.5 Criterio alternativo de muestreo de SO_2 basado en el uso de un estimador robusto de regresión

Cuando luego de un determinado tiempo de monitoreo continuo en un sitio dado, es posible suponer que los valores del contaminante en cuestión resultan redundantes en relación a sitios vecinos (Borge et al., 2014), o muy bajos, o bien se han agotado los objetivos del monitoreo, no es aconsejable abandonar totalmente el seguimiento de dicho contaminante, en particular si el mismo es relevante desde el punto de vista de los lineamientos o leyes. Resulta lógico considerar la posibilidad de reemplazar a la unidad analizadora continua por un método discontinuo de bajo costo que sirva como referencia. En el caso del SO_2, este método podría ser el de la Pararosanilina (EPA, 2010). Esta decisión se realiza haciendo consideraciones de tipo económicas y en el caso de las redes de monitoreo existen ciertos requisitos que se deben cumplir (EPA, 1980).
El objetivo de esta sección consiste en mostrar, a modo de ejemplo y utilizando los datos de una campaña corta (en donde no se satisfacen los lineamientos de la EPA), un método que permita realizar el reemplazo (de un sistema de medición continuo por uno discreto) de una manera controlada (utilizando criterios estadísticos). Para ello se tienen en cuenta los datos registrados de SO_2 en el Punto D durante la campaña de primavera de 2005 (Sección IV.6.4).

Si se considera la relación entre los promedios diarios de todos los días (92 datos) y los promedios horarios de una hora dada para todos los días (92 datos) será posible encontrar aquellas horas del día que representan mejor los promedios diarios. Esto implica operar con 24 nubes de puntos, una para cada hora del día acumulado.
Luego será deseable conocer la frecuencia de muestreo, o sea, cada cuanto se debe efectuar una medición con el método discontinuo durante una hora para que represente el promedio de un período. Este segundo objetivo hace necesario recurrir a una regresión entre los datos, puesto que la ordenada al origen y la pendiente serán de utilidad.
Se recurrió a la implementación de un estimador-*S* como una de las alternativas robustas posibles. La Figura IV.19 muestra la nube de puntos correspondiente a la Hora 13 (seleccionada como ejemplo). La recta de regresión llamada RR es la realizada con el estimador-*S* según lo descripto en Ratto et al. (2009). La recta de regresión llamada CM es el ajuste realizado por el método clásico de cuadrados mínimos. Es evidente como valores atípicos en el eje de las *X* (puntos palanca) distorsionan la pendiente de CM.
La Tabla IV.7 permite seleccionar de entre todas las nubes de punto la de menor residuo. El menor valor de las *S* indicará cual es la nube de puntos que posee el mayor poder predictivo. En este caso es el que corresponde a la Hora 13 (Figura IV.19). Por lo tanto, la Hora 13 es la más representativa de los promedios diarios.

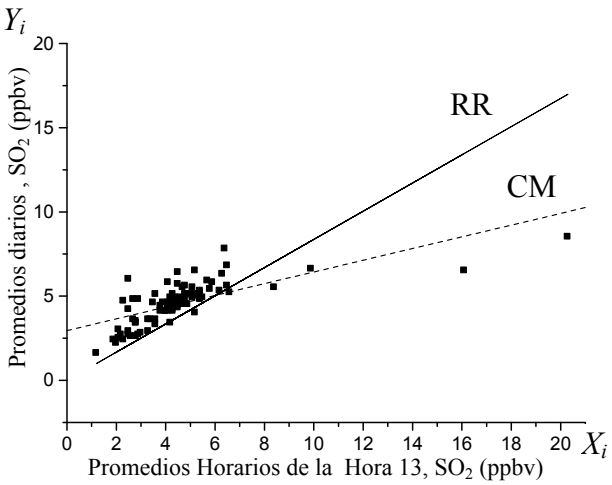


Figura IV.19: RR es la recta obtenida mediante un método robusto. CM es la recta obtenida mediante cuadrados mínimos.

Resta ahora determinar la frecuencia de muestreo.
Sea $\hat{y}_i = a\, x_i + b$ un modelo de regresión lineal. Sean:
x_i= promedios de la Hora 13 para el día i
y_i= promedios diarios de cada día de la campaña
$\hat{y}_i$= valor que predice el modelo
$\bar{\hat{y}}$ = promedio de los días dado por el modelo
$\hat{\mu}$ = estima de la media en el eje "y"

El desvío estándar de los promedios encontrados para la Hora 13 *DS(x)*= 2.52 (ppbv de SO_2), $|a_{RR}|$ para la Hora 13 es 0.84 (Tabla IV.5) y considerando $\hat{\mu} = \bar{\hat{y}}$ es entonces posible calcular el desvío estándar del modelo *DS(ŷ)* como:

$$DS(\hat{y}) = |\boldsymbol{a_{RR}}|\, DS(x)$$

Puesto que se tiene que $\bar{\hat{y}}$ = 6.06 ppbv es razonable que el desvío estándar de la media $DS(\hat{\mu})$ no difiera más de un 10% del valor de $\bar{\hat{y}}$.
Si la ecuación dada en Gilbert (1987) para las varianzas se reescribe para los desvíos estándar entonces queda:

$$DS(\hat{\mu}) = DS(\hat{y})/\sqrt{n}$$

de donde resulta que *n*=12 (tamaño de muestra o número de veces que se debe muestrear según los datos de la presente campaña).

Tabla IV.5: Resultados de la regresión robusta.

Primera columna: Horas del día en las que han sido acumuladas los promedios diarios de la campaña de primavera de 2005 en el CIOp.

Segunda y tercera columnas: pendiente (a_{RR}) y ordenada al origen (b_{RR}) obtenidas con un método de regresión robusta (RR) para cada nube de puntos que vincula los promedios diarios con los promedios horarios para cada día de campaña.

Tercera columna: mediana del valor absoluto de los residuos (*S)* que aparece multiplicada por 1000 para mayor claridad.

Tabla IV.5

Hora (acum.)	a_{RR}	b_{RR}	S (x 1000)
Hora 0	0.8125	0.0008500	0.4833
Hora 1	0.7700	0.0011080	0.4671
Hora 2	0.8750	0.0006917	0.4271
Hora 3	0.9266	0.0004877	0.4169
Hora 4	0.9568	0.0004527	0.4154
Hora 5	0.9664	0.0004068	0.3804
Hora 6	0.9479	0.0005062	0.3755
Hora 7	0.9954	0.0002199	0.2896
Hora 8	0.9250	0.0004567	0.3475
Hora 9	0.9611	0.0002806	0.4073
Hora 10	0.9625	0.0003358	0.3788
Hora 11	0.9417	0.0004358	0.3117
Hora 12	0.8737	0.0007597	0.3144
Hora 13	0.8375	0.0007704	**0.2758**
Hora 14	0.6944	0.0012350	0.3403
Hora 15	0.8409	0.0007614	0.3153
Hora 16	0.7869	0.0009251	0.3886
Hora 17	0.8542	0.0006229	0.3250
Hora 18	0.8431	0.0006963	0.3131
Hora 19	0.7813	0.0008198	0.3590
Hora 20	0.8623	0.0004123	0.3715
Hora 21	0.7708	0.0007958	0.4302
Hora 22	0.8866	0.0004900	0.4530
Hora 23	0.7969	0.0009375	0.4145

Esto implica que muestreando con una técnica discontinua como método alternativo será necesario hacerlo al menos cada 7 días (7 x 12 = 84). De efectuarse dicha medición a la Hora 13 se estará en presencia de un desvío menor o igual al 10% respecto de los valores dados por la técnica continua.
Por lo tanto, se ha mostrado, a modo de ejemplo, un método alternativo para reemplazar una unidad analizadora continua. Cabe destacar que el foco estuvo puesto en valores medios, pero que la misma metodología robusta puede llevarse a cabo si se quiere determinar, por ejemplo, la hora y la frecuencia de muestreo con mayores probabilidades de detectar picos de concentración (la Figura IV.19 sería reemplazada por una que tenga en cuenta los máximos horarios y diarios).

IV.6.6 Influencia estacional (ciclo anual) y horaria (ciclo diario) en los sectores 1 y 2 y sus tendencias en el tiempo

De acuerdo a lo hallado en secciones anteriores existen direcciones de viento que son de particular importancia. El Sector 1 (NNO- N- NE- NNE) transporta a los contaminantes de origen industrial hacia el casco de la ciudad mientras que el Sector 2 (ENE- E- ESE) lo hace hacia los barrios residenciales (Figura II.6- Capítulo II).

La Tabla IV.6 muestra los porcentajes de ocurrencia promedio de estos sectores en los puntos A y J durante el período 1998- 2003 y en el Punto J durante el período 1998- 2009. Tales lapsos se seleccionaron debido a la disponibilidad de los datos y la calidad de los mismos.

Tabla IV.6		
	Sector 1 (%)	Sector 2 (%)
Punto A $^{1998-2003}$	28.9	25.4
Punto J $^{1998-2003}$	27.6	23.0
Punto J $^{1998-2009}$	28.4	23.7

Tabla IV.6: Porcentaje de ocurrencia de los sectores 1 y 2 según distintos sitios de monitoreo y escalas de tiempo.
El promedio del Sector 1 para A y J durante 1998- 2003 es de 28.3 % mientras que para el Sector 2 es de 24.2 %.

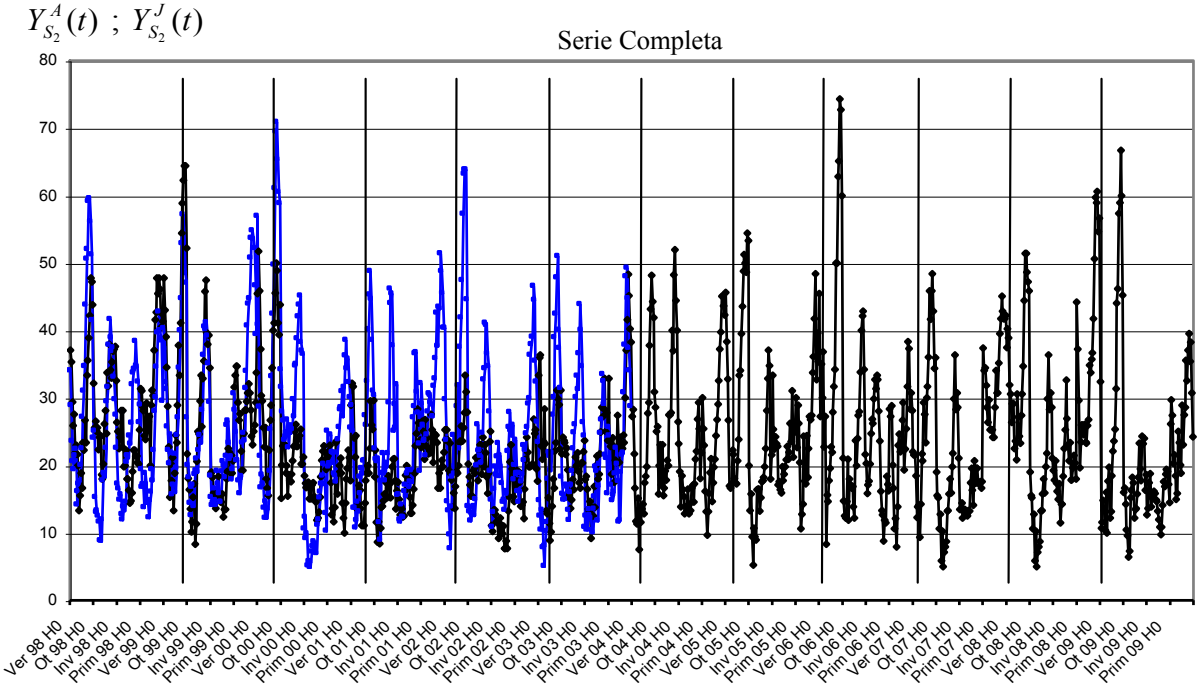
Ambos sitios dan valores similares. Si se suman las frecuencias de ambos sectores para un período y sitio dado, el porcentaje de ocurrencia de ambos sectores es mayor al 50% en todos los casos. Esto indica que la mayor parte del tiempo los vientos transportan a los contaminantes hacia donde más población se halla expuesta. Para profundizar en el conocimiento de estos sectores se determinará cuanta variación presentan según las estaciones del año y según las horas del día, al mismo tiempo se investigará si estos sectores presentan algún tipo de tendencia durante los períodos de estudio.

La Figura IV.20 resume el análisis llevado a cabo para el Sector 2 durante el período 1998-2003 en los puntos A y J y durante el período 1998- 2009 en el Punto J. Este sector se seleccionó, a modo de ejemplo, para mostrar los distintos pasos del análisis.
La Figura IV.20a muestra la evolución de las ocurrencias del Sector 2 en el Punto A ($Y_{S_2}^A(t)$) y en el Punto J ($Y_{S_2}^J(t)$). Cada punto de la figura representa una frecuencia de ocurrencia de este sector para una hora del día correspondiente a una estación del año para un año determinado. Un año en particular puede ser recorrido a través de las estaciones en el orden: verano, otoño, invierno y primavera. Cada estación está representada por 24 puntos (que corresponden a las 24 horas del día). Hay dos contribuciones que requieren ser discriminadas en estas series; la influencia horaria (ciclo diario) y la de las estaciones del año (ciclo anual). La metodología empleada hasta el final de esta sección fue propuesta por Maronna (CP), más detalles se pueden apreciar en Ratto et al. (2012b).

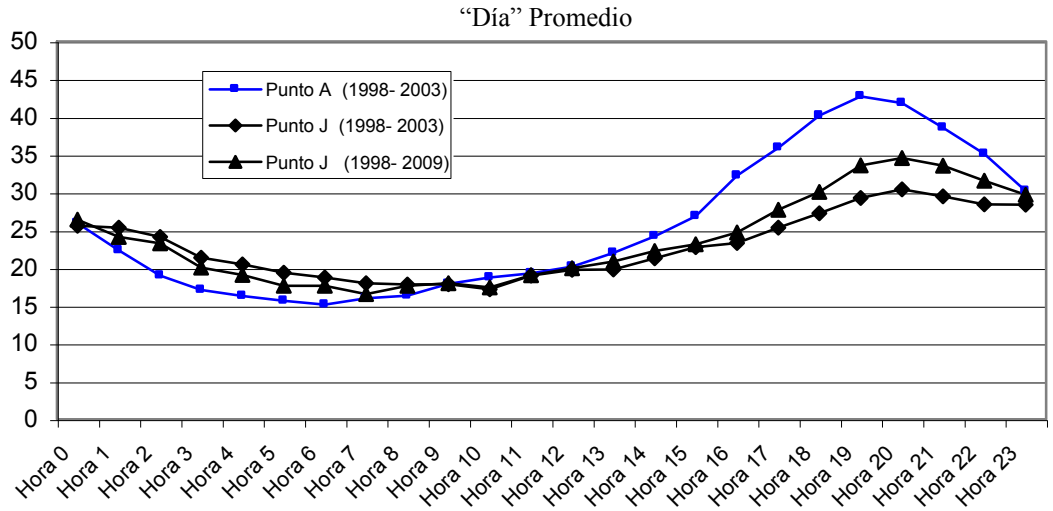
Ciclo diario y ciclo anual
Sustrayendo el día promedio observado en el Punto A (Figura IV.20b) a la serie original $Y_{S_2}^A(t)$ (Figura IV.20a) la nube de puntos resultante $C_{S_2}^A(t)$ (no mostrada) no tendrá la influencia del día. Pero $C_{S_2}^A(t)$ todavía tiene la influencia de las estaciones. Se obtienen los promedios de las estaciones cuyos valores son 3.44 para el verano, -2.96 para el otoño, -1.57 para el invierno y 1.09 para la primavera (Figura IV.20c). Sustrayendo esos promedios estacionales a $C_{S_2}^A(t)$ la nueva nube de puntos resultante (residuos del Sector 2 en el Punto A, o sea, $R_{S_2}^A(t)$) no tendrá la influencia ni de las horas del día ni de la estación del año (ver la nube de puntos en la Figura IV.20d).

Un procedimiento análogo se siguió para los datos del Punto J en los períodos 1998-2003 y 1998- 2009, lo mismo fue realizado para el Sector 1.

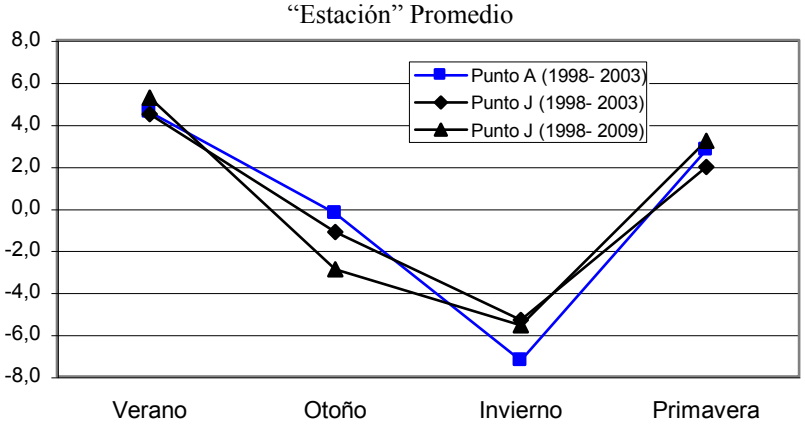
$Y_{S_2}^{A}(t)\ ;\ Y_{S_2}^{J}(t)$



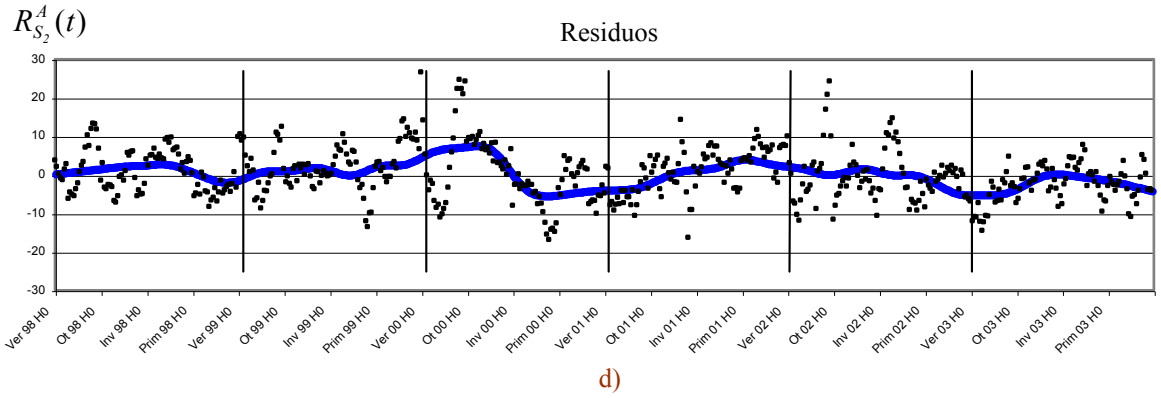
a)



b)



c)



d)

Figura IV.20**:** Serie original del Sector 2. Influencia diaria y estacional sobre el Sector 2 en el Punto A (1998- 2003) y en el Punto J (1998-2003; 1998-2009). Residuos del Sector 2 en el Punto A y la curva de suavizado correspondiente.

a) $Y^A_{S_2}(t)$ representa la frecuencia de ocurrencias de los vientos del Sector 2 observadas en el Punto A respecto del total de ocurrencias durante el período 1998- 2003 (curva azul). $Y^J_{S_2}(t)$ ídem para el Punto J pero cubriendo el período 1998- 2009. Cada punto del gráfico representa la frecuencia de vientos soplando desde el Sector 2 para una determinada hora (*t*) del día para una particular estación del año y para cada año del período especificado. Los valores de t están identificados cada 24 datos y están expresados de forma abreviada, por ejemplo, Ver 00 H0 indica la Hora 0 del Verano del año 2000. La cantidad total de datos es de 576 puntos para el Punto A (que cubre 6 años de observaciones) mientras que de 1152 datos para el Punto J (que cubre 12 años).
b) El eje de las *Y* representa el porcentaje de ocurrencias del día promedio para el Sector 2 desde el punto de vista de los puntos A (líneas azules) y J para los dos períodos de estudio (líneas negras). El eje de las *Y* fue construido promediando cada hora acumulada según los años y las estaciones del año.
c) El eje *Y* representa el porcentaje de ocurrencias del promedio de las estaciones.
d) Residuos de la serie de la Figura IV.20a en el Punto A. La curva suavizada fue obtenida mediante la aplicación de un método de regresión local (LOESS) (Sección IV.3.3). Las líneas verticales señalan el inicio de año.

Para evaluar la contribución del ciclo diario y del ciclo anual se recurrió a cuantificar las varianzas involucradas en cada paso del procedimiento descripto.
Por ejemplo, si la varianza de la serie original $Y^A_{S_2}(t)$ es 147.0 y la varianza de los datos remanentes al restar el día promedio $C^A_{S_2}(t)$ es 64.0 entonces la diferencia (83) que representa el 56.5% de la varianza de la serie original será influencia del ciclo diario (*ICD*) para el Sector 2 en el Punto A. Si a la varianza de $C^A_{S_2}(t)$ se le resta la varianza de los residuos $R^A_{S_2}(t)$ (Figura IV.20d) se obtiene 20.5 que representa el 13.9% de la varianza de la serie original $Y^A_{S_2}(t)$, esta será la influencia del ciclo anual (*ICA*) dado por la presencia de las estaciones. Finalmente, la varianza de los residuos $R^A_{S_2}(t)$, que representa el 29.6% de la varianza de la serie original, constituye la fracción inexplicada de la variación total (*FIVT*).

La Tabla IV.7 resume los aportes a la variación total para los distintos períodos y sitios de monitoreo.

Tabla IV.7			
Sector 1			
	Punto A 1998- 2003	Punto J 1998- 2003	Punto J 1998- 2009
ICD	51,3	29,6	25,6
ICA	6,4	15,4	22,3
FIVT	42,3	55,0	52,1
Sector 2			
	Punto A 1998- 2003	Punto J 1998- 2003	Punto J 1998- 2009
ICD	56,5	20,3	29,2
ICA	13,9	16,0	19,2
FIVT	29,6	63,8	51,6

Tabla IV.7: % de variación atribuida a la influencia de las horas día (ciclo diario), de la estación del año (ciclo anual) y la fracción inexplicada respecto de la variación total de la serie original.

ICD : influencia del ciclo diario (%).
ICA : influencia del ciclo anual (%).
FIVT : fracción inexplicada de la variación (%).

Una vista general de esta tabla muestra que, independientemente de la estación del año, el Punto A tiene más variación que el Punto J en el ciclo diario (*ICD*) siendo esta variación más pronunciada en el Sector 2 que en el Sector 1. Como se señaló en el Capítulo II (Sección II.1.2) los vientos a escala sinóptica sobre el Río de La Plata están originados principalmente por el anticiclón subtropical del Atlántico Sur (Sección II.1.2- Capítulo II) coexistiendo con la circulación local de tipo brisa de mar- tierra (Berri et al., 2010).
La influencia del anticiclón será la misma para ambos sitios puesto que solo difieren en la rugosidad de los terrenos, pero la brisa marina influirá más sobre el Punto A (más próximo al río) que sobre el Punto J (más alejado de la costa) (ver Figura II.6- Capítulo II). La circulación de la brisa de mar- tierra influye más en algunas direcciones que en otras (preferentemente en las del Sector 2 donde se observa un fenómeno de rotación durante la tarde).
En conclusión, en todos los casos el ciclo diario es más pronunciado que el ciclo anual. En el Punto A esta diferencia es más pronunciada para el Sector 1 (del orden de 8 veces) que para el Sector 2 (del orden de 4 veces) mientras que en el Punto J -durante el mismo período- para el Sector 1 la proporción es del orden de 2 mientras que para el Sector 2 el ciclo diario es solamente algo mayor al anual. Si se considerara un contexto de emisiones industriales constantes, esto mostraría por ejemplo, que las variaciones diarias de concentración a nivel de calidad de aire en el Punto A (que es un buen detector de emisiones transportadas por el Sector 1) tengan mayores amplitudes (rangos, varianzas, etc.) que las observadas entre las estaciones del año. De manera análoga y en relación al Punto J (que es un buen detector de emisiones transportadas por el Sector 2), no sería esperable una gran diferencia entre las amplitudes diarias y las estacionales.

Tendencia

Los residuos resultantes de haber sustraído la influencia de los ciclos diarios y anuales (por ejemplo, nube de puntos de la Figura IV.20d) pueden contener aún algún patrón. Con el objeto de investigar esta posibilidad se recurrió al método de LOESS (Sección IV.3.3.1). Siguiendo con el ejemplo del Sector 2 en el Punto A, la Figura IV.20d muestra la curva de suavizado (línea llena) obtenida con LOESS (Anexo IV.3, pág. 108). A pesar de que no es observable ningún patrón periódico la parte final de la curva muestra una leve tendencia decreciente. Para analizar problemas de este tipo (que podrían aparecer en cualquiera de las otras curvas no involucradas en este estudio) se recurrió al siguiente procedimiento (Anexo IV.3, pág. 108): se adoptaron "ventanas" de 48 datos (por prueba y error), se calcularon la media, el coeficiente de autocorrelación de primer orden y el desvío de la media para cada una de las ventanas.

La Tabla IV.8 muestra estos valores para los datos de la Figura IV.20d. Puesto que las diferencias entre las medias de las ventanas consecutivas son, en general, menores que los desvíos de la media no hay evidencia de que haya una tendencia.

Los datos de los residuos correspondientes al Sector 2 para el Punto J correspondientes a los períodos 1998- 2003 y 1998- 2009 así como los datos del Sector 1 para el período 1998- 2003 fueron tratados de forma análoga no encontrándose tendencia ni creciente ni decreciente.

Tabla IV.8

Ventana rango de datos	Media aritmética	Autocorrelación (coeficiente de primer orden)	Desvío de la Media
1- 48	0,4872	0,829	2,8801
49- 96	0,4726	0,866	2,9388
97- 144	1,0267	0,745	2,1585
145- 192	2,8601	0,899	4,2724
193- 240	5,4809	0,839	4,9590
241- 288	-4,6920	0,844	2,9512
289- 336	-2,4045	0,602	2,0300
337- 384	3,5622	0,671	1,5209
385- 432	-0,4920	0,693	3,0246
433- 480	-0,4524	0,851	3,2849
481- 528	-4,0983	0,787	2,2347
529- 576	-1,7503	0,725	1,8401

Tabla IV.8: Criterio para reforzar la discriminación de tendencias en la series según Maronna (CP). En esta tabla se muestra el coeficiente de autocorrelación utilizado para calcular el desvío de la media (Anexo IV.3, pág. 108).

Este resultado se halla en concordancia con estudios que analizan la variabilidad interanual de distintas variables meteorológicas en la costa y el estuario del Río de La Plata (Escobar et al., 2003; Berri et al., 2010; Dragani et al., 2010).

IV.6.7 Análisis de calmas utilizando un estimador-*M* de correlación

La ocurrencia de calmas constituye un fenómeno importante en relación al estancamiento de los contaminantes (McCormik, 1968). Horas consecutivas de calmas pueden constituir una condición meteorológica propicia para la acumulación de grandes cantidades de contaminantes del aire en las cercanías de las fuentes emisoras. Este fenómeno, llamado "efecto de acumulación" (Alvarez Morales y Alvarez Escudero, 2000; Alvarez Escudero et al., 2007) puede ser caracterizado, en principio, estudiando la localización horaria de las calmas y sus duraciones.

En la presente sección se describe la estructura de las calmas en los puntos A (1997- 2003), J (1997- 2006) y K (1995- 2005) (Figura II.6- Capítulo II) y se extraen conclusiones sobre la similitud de los patrones estacionales observados utilizando un estimador robusto de correlación (Sección IV.2.1).

La Figura IV.21 muestra una curva típica de la ocurrencia de calmas según la hora del día. Es evidente la presencia de dos máximos, uno cercano a la salida del sol (comienzo de la construcción de la capa límite diurna) y otro en el anochecer (comienzo de la construcción de la capa límite nocturna). Durante la noche se observan valores relativamente altos de calmas en coincidencia con las estabilidades nocturnas. Un amplio valle entre los máximos da cuenta del crecimiento de la capa de mezcla durante el día (Sección III.9- Capítulo III).

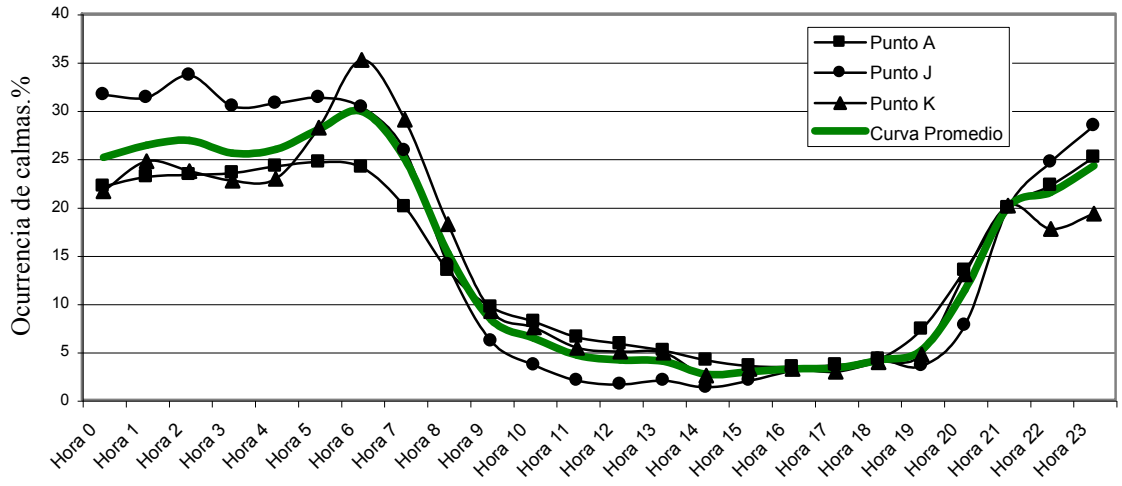


Figura IV.21: Distribución horaria de las calmas en distintos sitios de monitoreo para la estación verano elegida como referente y por cuestiones de espacio. El eje de las *Y* representa los promedios de frecuencias de ocurrencia de calmas en relación al total de ocurrencias expresadas en %. La curva llena suavizada (verde) representa el promedio de los tres sitios.

El promedio general de ocurrencia de calmas en los tres sitios es de 14.7 % en verano (promedio de la curva llena en la Figura IV.21), 19.1% en otoño, 12.8% en invierno y 11.6% en primavera. Con el objetivo de detectar similitudes en los patrones en todas las estaciones del año se recurrió a la aplicación de un estimador-*M* (Anexo IV.1, pág. 106). La Tabla IV.9 muestra las correlaciones entre todos los sitios de monitoreo a lo largo de las estaciones del año.

Tabla IV.9

Sitios que se correlacionan	Verano	Otoño	Invierno	Primavera	Promedio
Punto A- Punto J	0,9846	0,9567	0,9293	0,9742	**0,9612**
Punto A- Punto K	0,9891	0,9766	0,9112	0,9925	**0,9674**
Punto J- Punto K	0,9520	0,7415	0,8020	0,9752	**0,8677**
Promedio	**0.9752**	**0.8916**	**0.8808**	**0.9806**	**0.9321**

Tabla IV.9: Coeficientes de correlación utilizando el estimador-*M* mencionado en la Sección IV.2.1 y descripto en el Anexo IV.1 (pág. 106).

La misma permite apreciar que tanto los sitios como las estaciones del año se hallan altamente correlacionados. En particular, las estaciones cálidas (verano y primavera) tienen coeficientes más altos que las estaciones frías (otoño e invierno). Estos hallazgos permiten establecer la existencia de un patrón generalizado de calmas en la zona.

Para profundizar en el análisis es conducente conocer cómo están distribuidas las calmas según su duración, para ello se establecieron intervalos de duración de 1 hora. La calma más larga encontrada fue de 20 horas.

La Figura IV.22 muestra las curvas de distribución de las calmas según su duración para cada estación del año. Cada punto de esta curva representa la frecuencia de calmas encontradas de una determinada duración (por ejemplo, 1 hora, 2 horas, etc.) respecto del número total de ocurrencia de calmas (todas las duraciones) para una dada estación del año. Las calmas con duración de 1 hora representan en promedio el 50.6 %, las que duran 2 horas 20.1%, las que duran 3 horas 9.5 %, las que duran 4 horas 6.2 % y las que duran 5 horas 3.7%; el resto de las duraciones (hasta 20 horas) representan solo el 2.2 %. Puesto que las calmas cuya duración es de 5 horas o menos representan el 90.1% de las ocurrencias totales se adoptaron estas 5 duraciones para continuar el análisis.

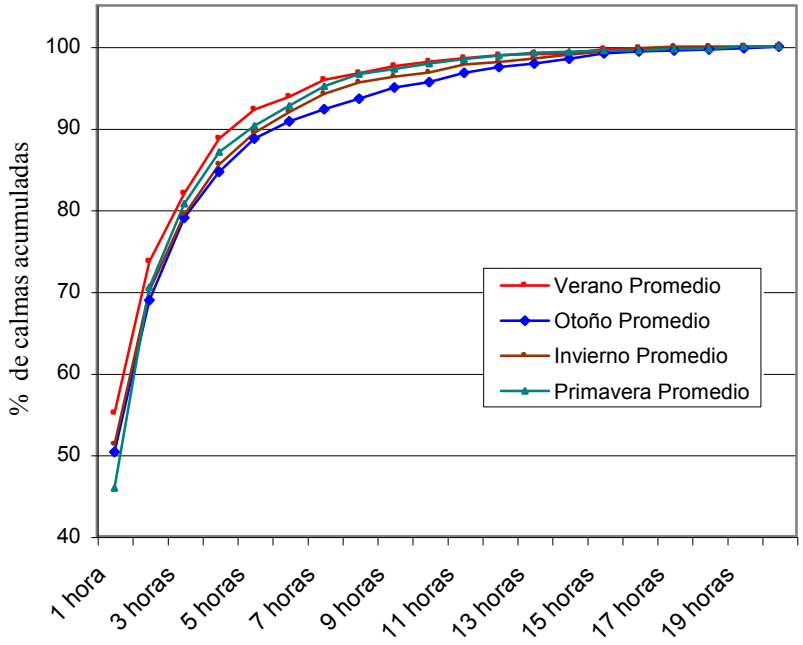


Figura IV.22: Calmas acumuladas (%) en intervalos de 1 hora para cada estación del año en los puntos A, J y K. Los porcentajes están expresados respecto del total de duraciones y horas del día.

Si bien se realizaron las curvas de ocurrencias de calmas para cada una de las duraciones antedichas por estación del año, no se halló una estructura en las mismas. Por esta razón y por cuestiones de espacio, se prefirió continuar el análisis involucrando a todas las estaciones del año. La Figura IV.23 muestra el porcentaje de calmas observado (eje *Y*) para cada hora del día (eje *X*) y una duración determinada (parámetro) respecto del total de calmas (todas las duraciones). Por ejemplo, a la Hora 9 (ver línea vertical de rayas (roja) en la Figura IV.23a) el 72.2% de las calmas tienen una duración de 1 hora, el 15.2% (2 horas), el 6.2% (3 horas), el 2.4% (4 horas) y el 1.5% (5 horas).

Estas 5 duraciones suman el 97.5% (el resto son duraciones mayores). La Figura IV.23a muestra que a través del día existe un rango horario (entre la Hora 7 y la Hora 12) que contiene a los principales picos de calmas. Para duraciones más largas (Figura IV.23b a Figura IV.23e) hay dos regiones particulares que evidencian un patrón: una dada en las horas de la madrugada y otra perteneciente al anochecer. En base a esto último y en relación a duraciones "largas", o sea, entre 2 y 5 horas es posible establecer que el anochecer (comienzo de las estabilidades nocturnas) y la madrugada (plenitud de las estabilidades nocturnas) constituyen dos momentos del día que son propicios para la acumulación de los contaminantes atmosféricos.

Considerando que el año está constituido por 8760 horas y que el promedio general de calmas para las cuatro estaciones del año es de 14.6%, surge que el número de horas de calmas anuales será aproximadamente 1275. De esta cifra 645 eventos corresponderán a calmas de 1 hora, 256 a calmas de dos horas, etc. Estas estimaciones deben considerarse en un sentido amplio dada la distinta calidad de los datos de los conjuntos de trabajo.

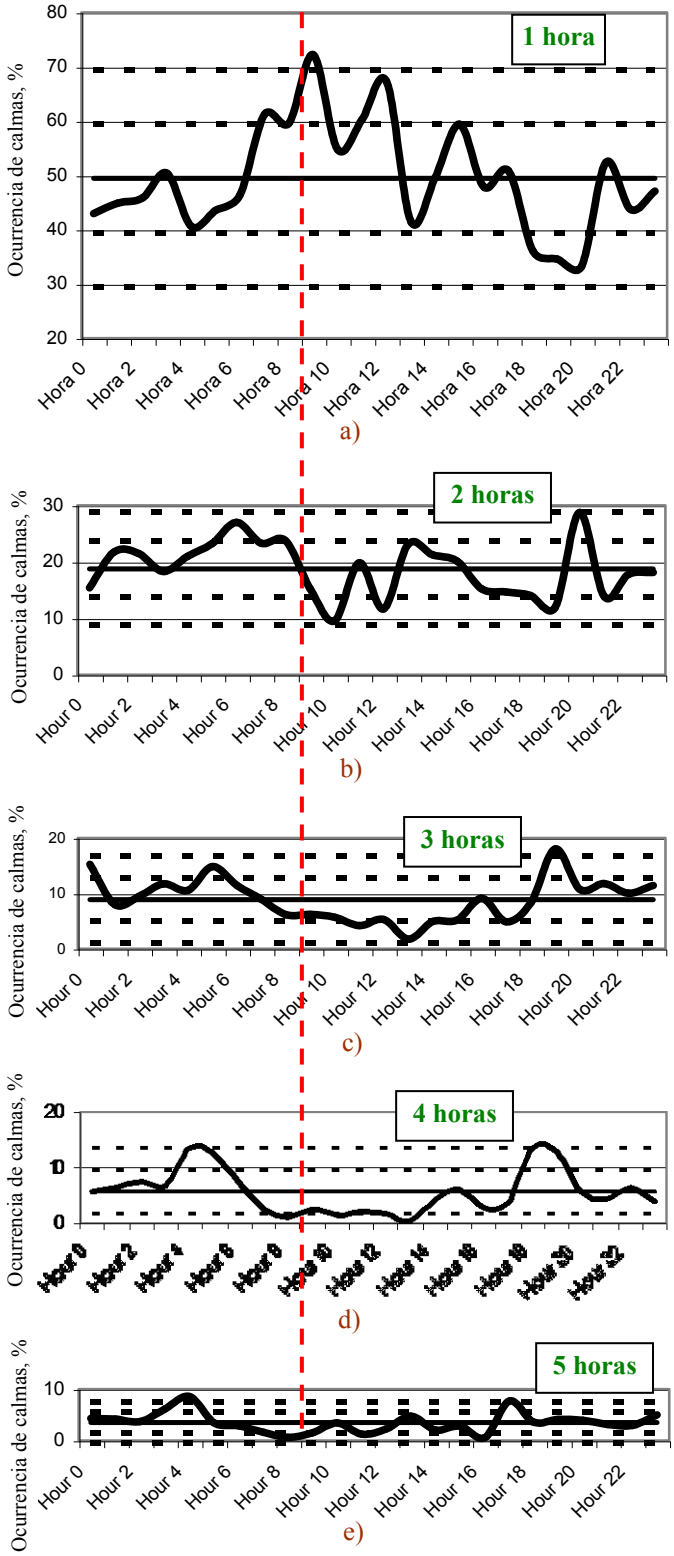


Figura IV.23: Ubicación de las calmas (%) a lo largo del día según diferentes duraciones:
a) **1 hora** de duración
b) **2 horas** de duración
c) **3 horas** de duración
d) **4 horas** de duración
e) **5 horas** de duración

Los porcentajes se hallan expresados respecto del total de duraciones (hasta 20 horas) a lo largo de una determinada hora.

La línea recta horizontal central de cada gráfica representa el promedio de ocurrencia de la duración correspondiente. Las dos líneas con guiones por encima y debajo del promedio indican 1 y 2 desvíos estándar.

La línea vertical a rayas indica el porcentaje de calmas para la Hora 9 a lo largo de las cinco duraciones.

IV.6.8 Salida de calmas

Los primeros vientos que aparecen luego de los períodos de calma son fundamentales para conocer el destino de los contaminantes que se han acumulado en torno a las fuentes de emisión. Si se computan las primeras direcciones de viento que aparecen inmediatamente después de finalizada la calma y esta cuenta se acumula para un determinado período, se estará en condiciones de construir una roseta de vientos que hemos dado en llamar (Ratto et al., 2012 a, 2012c) rosetas de vientos de salida de calmas (RVSC). La Figura IV.24

muestra la RVSC (línea de trazos) en ejes cartesianos observada en el Punto J durante el período 1998- 2007 para el verano. En la misma gráfica se halla representada (línea llena) la roseta de vientos de rango completo (aquella que contiene a todas las velocidades) correspondiente al mismo período.

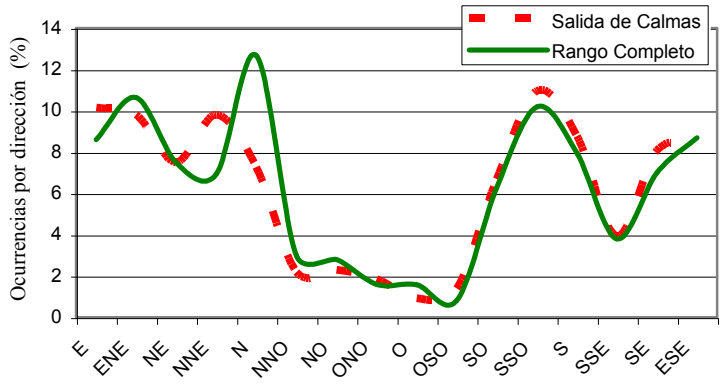


Figura IV.24: Frecuencias de ocurrencia de vientos por dirección según una roseta de vientos de rango completo y la correspondiente roseta de salida de calmas para el verano.

Es evidente la existencia de similitud entre ambas rosetas de viento. Esto implica que los primeros vientos luego de una calma siguen en promedio un patrón muy similar al de los vientos totales.

Para poder comparar el grado de similitud entre los distintos pares de rosetas correspondientes a todas las estaciones del año se recurrió al *SAD* (Sección IV.2.2).

Se eligió esta métrica (Sección V.3- Capítulo V) porque al mismo tiempo que da una idea de la disimilitud relativa entre pares, provee una idea del "error" que se produciría en utilizar la roseta de rango completo (sencilla de calcular) en lugar de la de salida de calmas (que insume mucho tiempo de cálculo). Los valores de *SAD* son para el verano (Figura IV.24) 16.9%, para el otoño 10.7%, para el invierno 27.9% y 31.7% para la primavera; el promedio de las cuatro estaciones es de 21.8%.

Un análisis más detallado que involucre el cómputo de rosetas de viento de dirección por rangos de velocidad permitirá encontrar aquellas que aproximen mejor a las RVSC siendo las candidatas más firmes aquellas de bajas velocidades (ver Ratto et al. (2012a).

Dada la importancia de los sectores 1 y 2 será conveniente cuantificar sus frecuencias relativas luego de las calmas. La Tabla IV.10 muestra tales frecuencias junto a las de los sectores 1 y 2 de las rosetas de rango completo. La máxima diferencia que se observa es para el Sector 1 en invierno (7.2%). Para el Sector 2 todas las diferencias se hallan debajo del 1.2%. Esto implica que para los sectores 1 y 2 la roseta de vientos de rango completo predice a la de salida de calmas con bajo error.

Tabla IV.10

Sector 1		
	RVSC	Rango Completo
Verano	26.8	30.1
Otoño	24.8	27.5
Invierno	23.6	30.8
Primavera	20.9	25.5
Sector 2		
Verano	28.4	27.9
Otoño	23.1	21.9
Invierno	17.8	18.3
Primavera	24.7	25.9

Tabla IV.10: frecuencias de ocurrencia (%) para los sectores 1 y 2 según las rosetas de salida de calmas (columna 2) y rango completo (columna 3).

Tabla IV.11

	Todas las direcciones	Sector 1	Sector 2
Verano	2.5	2.5	2.3
Otoño	2.6	2.6	2.7
Invierno	2.7	3.0	3.2
Primavera	2.3	2.7	2.7
Promedio	**2.5**	**2.8**	**2.6**

Tabla IV.11: Proporciones de velocidad entre la roseta de vientos de rango completo de velocidad y aquellas de salida de calmas para todas las direcciones (columna 2) y para las direcciones correspondientes a los sectores 1 y 2 (columnas 3 y 4).

Otro aspecto a considerar lo constituyen las velocidades inmediatas luego de las calmas. Es esperable que luego de una calma el viento tenga velocidades bajas. Con el objetivo de

cuantificar este hecho se construyó la Tabla IV.11. Esta tabla muestra la relación entre la velocidad de los vientos de la roseta de rango completo de cada estación con la velocidad de los vientos de la RVSC, también la relación entre las velocidades de los sectores 1 y 2 en ambas rosetas. En términos generales, las velocidades promedio de la roseta de rango completo son entre 2.5 y 3 veces superiores a las de las rosetas de salida de calmas.

IV.6.9 Velocidades de viento

Con el objetivo de darle un contexto a lo discutido (principalmente en relación a las direcciones de viento) se presenta una breve discusión sobre las velocidades de viento en los distintos sitios de monitoreo de la ciudad y alrededores. La Tabla IV.12 muestra velocidades medias observadas para los períodos 1998- 2003 en los puntos A y J y para el período 1998- 2009 en el Punto J y las correspondientes velocidades corregidas según la ecuación de la "ley de la potencia" descripta en la Sección III.6- Capítulo III.

Tabla IV.12

	Verano	Otoño	Invierno	Primavera	Promedio
Punto A 1998- 2003 observados	7.1	6.7	7.7	8.2	7.4
Punto A 1998- 2003 estimados con la ecuación (*) (p=0.25)	6.8	6.4	7.4	7.8	7.1
Punto J 1998- 2003 observados	6.6	6.4	6.3	6.8	6.5
Punto J 1998- 2003 estimados con la ecuación (*) (p=0.15)	7.3	7.1	7.0	7.5	7.2
Punto J 1998- 2009 observados	6.9	6.3	6.8	7.2	6.8
Punto J 1998- 2009 estimados con la ecuación (*) (p=0.15)	7.7	7.0	7.5	8.0	7.5

Tabla IV.12**:** Velocidades promedio de vientos (km h^{-1}) observadas en el Punto A (12 m de altura) y en el Punto J (5 m de altura).
(*): Ecuación III.1 (Capítulo III), expresa las velocidades corregidas según la Ley de la Potencia. El factor p tiene en cuenta la rugosidad del terreno y el tipo de estabilidad atmosférica según Pasquill. Los cálculos se hicieron para estabilidad neutra según recomienda Wark et al. (1998).

Los valores observados en el Punto A son mayores a los observados en el Punto J debido a que dentro de la capa límite planetaria (Sección III.6- Capítulo III) las fuerzas de fricción decrecen con la altura. Con la aplicación de la ecuación III.1, que tiene en cuenta la rugosidad del terreno, las diferencias entre observaciones se hacen muy pequeñas y las velocidades en el Punto J (zona semirural) superan, aunque levemente, las corregidas del Punto A (zona urbana). En general, la representatividad de los datos meteorológicos dependen del usuario al que están destinados (Wieringa, 1996). Los registros de vientos llevados a cabo en los aeropuertos de las ciudades tienen por principal objetivo facilitar el tránsito aéreo (Wieringa, 1980) y no resultan apropiados para realizar estudios de contaminación del aire (Holzworth, 1967). Sin embargo, se ponen en consideración observaciones mensuales llevadas a cabo en el Punto K (Aeropuerto de La Plata) durante la década 2001- 2010 (SMN, 2011) para proveer de una referencia y por tratarse de los únicos datos oficiales en la zona.

Tabla IV.13

	Verano	Otoño	Invierno	Primavera	Promedio
Sitio K 2001- 2010	14.7	12.7	13.4	15.0	14.0

Tabla IV.13: Velocidades promedio observadas a 10 m de altura sobre el terreno. El Punto K se halla ubicado en una zona de características

La Tabla IV.13 permite apreciar que los valores en el Punto K son alrededor de 2 veces más grandes que los valores corregidos en los puntos A y J. Sin embargo, los tres puntos de seguimiento (A, J y K) revelan una misma tendencia con velocidades algo más altas en verano y primavera que en otoño e invierno. El hecho de que las velocidades observadas sean más altas en el Punto K puede atribuirse, por un lado, a las diferencias que existen entre los climas rurales y los urbanos y por otro, la rugosidad de los terrenos (Sección III.5 y Sección III.6 (Capítulo III); Landsberg, 1981; Gassmann et al., 2002). Velocidades promedio de 13.0 km h^{-1} observadas en el Punto I (Figura II.6- Capítulo II) ubicado a aproximadamente 1 km al sudoeste del Punto A a una altura de 40 m cubriendo el período 1967- 1994 sustentan esta idea (el valor corregido es de 9.2 km h^{-1}). Pero también deben considerarse diferencias en la calidad de los datos (diferencias entre instrumentos, fechas, periodicidad de muestreo, etc.) según se describen en la Sección II.3.2- Capítulo II.
El promedio general de los valores corregidos de velocidades de viento de la Tabla IV.12 es de 7,3 km h^{-1}; considerando la Escala Beaufort (Sección III.4- Capítulo III) estos vientos se corresponden con "brisa suave" mientras que el promedio de vientos en la zona del aeropuerto (Tabla IV.13) se corresponden con "brisa leve".
Como se vio en la Sección IV.6.8, las calmas permiten la acumulación de los contaminantes pero también se generan condiciones propicias de acumulación cuando las velocidades de los vientos horizontales son bajas (Moore, 1969; Deadorff, 1984). Según McCormik (1968) la persistencia de vientos de superficie menores a 10 km h^{-1} tiende a acumular contaminantes. Sharan et al. (1996) y Goyal y Rama Krishna (2002) establecen que velocidades inferiores a 7.2 km h^{-1} a 10 m de altura son considerados vientos de baja velocidad. El percentil 50 de las velocidades corregidas de la Tabla IV.12 es en promedio de 7.1 km h^{-1}, esto indica que la mitad de las veces las velocidades son bajas, recién el percentil 80 supera los 10 km h^{-1}. Si además se considera el rol de las estabilidades atmosféricas (Sección III.7- Capítulo III) y las alturas de la capa de mezcla (Sección III.9- Capítulo III) se podrá tener un panorama más rico. El único trabajo con mediciones encontrado en la zona (Mazzeo et al., 1971) da a las clases D (neutra) y E (ligeramente estable) como muy frecuentes, mientras que los máximos de las alturas promedio de la capa de mezcla son de aprox. 1600 m en verano y de 700 m en invierno. Estos últimos valores son consistentes con mediciones más recientes (Gassmann, 1998) realizadas en la localidad de Ezeiza (ubicada a una distancia directa aproximada de 55 km al ONO de La Plata) que registran alturas de 1524 m en verano y 850 m en invierno. Gassmann (1998) establece que en las cercanías de la zona de estudio (Buenos Aires y su área metropolitana) las estaciones de otoño e invierno son las más pobres en capacidad de autodepuración. Gassmann y Mazzeo (2000) ubican a la zona de estudio dentro del cordón industrial-poblacional que une Rosario con La Plata (desde el noreste hacia el este) como uno de los dos sitios más pobres de la Argentina en cuanto a la capacidad para depurar el aire. Según los autores, la peor condición de autodepuración atmosférica está dada durante el invierno con frecuencias de ocurrencia que oscilan entre 23.1% y el 36,0%, mientras que la mejor se produce en verano donde las frecuencias de ocurrencia de mala autodepuración se hallan entre 8.5% y 23,0%. Considerando los máximos de los promedios mencionados (aprox. 1600 m en verano y aprox. 700 m invierno) y teniendo en cuenta que el valor crítico de ventilación de 6000 m^2 s^{-1} (Sección III.9- Capítulo III) implica un viento transporte mínimo de 4 m s^{-1} (14.4 km h^{-1}) y una altura mínima de capa de mezcla de 1500 m, surge la importancia de realizar mediciones para poder calcular el viento transporte al mismo tiempo que realizar mediciones de la altura de la capa de mezcla. De esta manera será posible caracterizar el potencial de contaminación para el caso puntual de La Plata y alrededores.
Por lo discutido en esta subsección, es posible concluir que el área de estudio reúne,

durante una parte del tiempo no pequeña, condiciones que hacen difícil la remoción de los contaminantes.

IV.6.10 Sectores 1 y 2 y selección de un sitio para observar concentraciones de fondo

En la Sección IV.6.6 se han analizado los sectores 1 y 2 cuya importancia radica en que ambos implican direcciones de viento que transportan a los contaminantes de origen industrial hacia una gran cantidad de población expuesta. Se mostró que la ocurrencia de estos vientos tiene lugar la mayor parte del tiempo.

Recurriendo a otros conjuntos de datos, Punto A (1997- 2003), Punto D (2006- 2007), Punto J (1997- 2006) y Punto K (1995- 2005) el objetivo de esta sección es, por un lado, mostrar que los patrones para los sectores 1 y 2 hallados previamente (Sección IV.6.6) son observables de forma muy similar desde otros sitios, lo cual permite generalizar el comportamiento de estos vientos a una mayor zona de influencia (ver los puntos A, D, J y K en la Figura II.6- Capítulo II). Por otro lado, y en vistas a la necesidad de la instalación de una red de monitoreo de los contaminantes del aire (Sección I.1- Capítulo I), se propone, entre los sitios de observación, seleccionar aquel que manifieste tener más ventajas para el seguimiento de la contaminación de fondo (requisito importante en el diseño de redes (EPA, 2013)). Dado que los conjuntos de datos difieren en la cantidad de años de observaciones, se trabajó utilizando el promedio ponderado (es decir, teniendo en cuenta un factor de peso proporcional a la cantidad relativa de años de medición).

La Figura IV.25 muestra las frecuencias de ocurrencia del Sector 1 en cuatro sitios de monitoreo para las estaciones de verano e invierno.

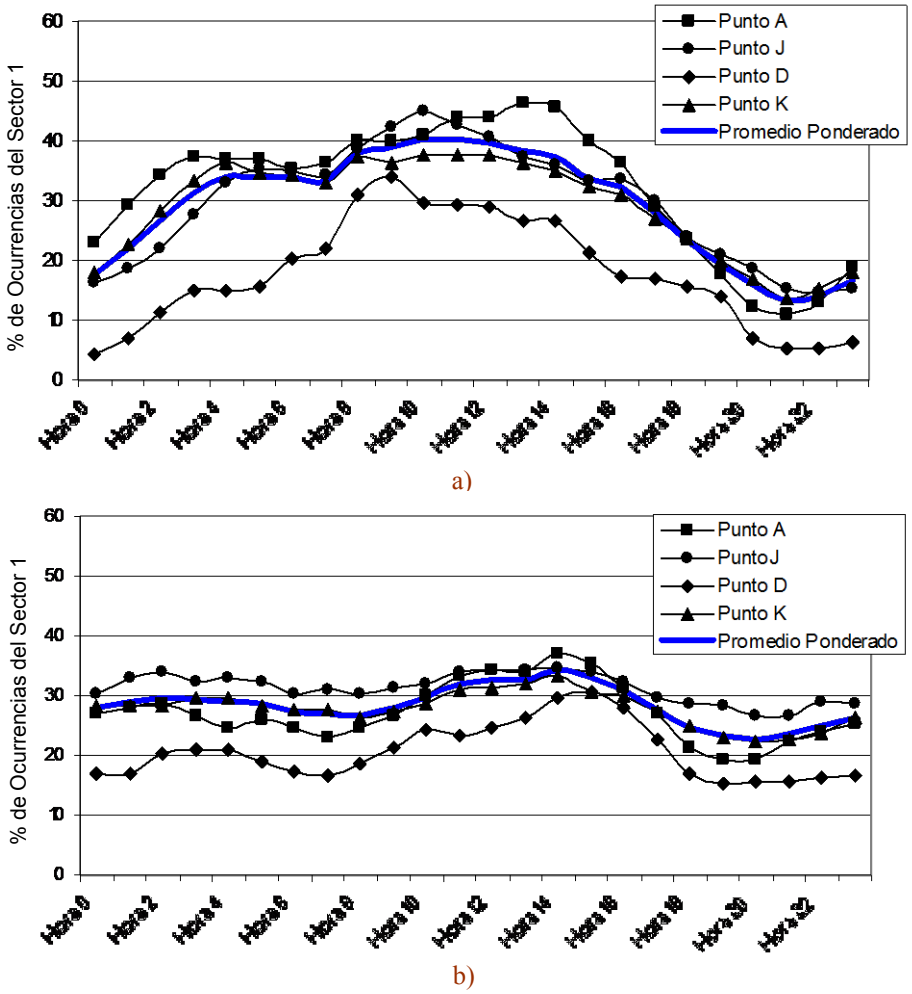


Figura IV.25: Frecuencias de ocurrencia del Sector 1 en distintos sitios y períodos de tiempo y la curva promedio. a) Verano (promedio ponderado total 29.2%) b) Invierno (promedio ponderado total 28.4%).

Las curvas muestran una buena similitud entre sitios siendo el Punto D con poco tiempo de registros el que más difiere. Todas las curvas (incluyendo las correspondientes al otoño y la primavera que no se muestran por cuestiones de espacio) poseen una franja de máximos cuyos extremos se hallan aproximadamente entre las 9 y 14 horas (como es el caso del verano con un promedio ponderado de ocurrencias de 39.1%) o entre las 11 y 16 horas (como es el caso del invierno con un promedio ponderado de ocurrencias del 32.6%).
Los mínimos se hallan para todas las estaciones entre las horas 19 y 22. Los mayores picos se observan para verano. La primavera tiene un pico menor y el del otoño se halla entre el del invierno y la primavera. El promedio ponderado general de ocurrencias del Sector 1 para los cuatro sitios y estaciones del año es de 27.3%.
Curvas análogas del Sector 1 observadas en los puntos A y J solo en el período 1998- 2003 (no mostradas por cuestiones de espacio) muestran formas y valores porcentuales muy similares; además, los máximos de ocurrencias para las franjas horarias referidas en el párrafo anterior para este sector son en promedio 42.1 % en verano y 33.2 % en invierno.
Esta comparación pone en evidencia que el Sector 1 posee un patrón generalizado en la zona de estudio (La Plata y alrededores), su curva promedio se halla representada con línea llena (azul) en la Figura IV. 25. La Figura IV.26 es análoga de la Figura IV.25 para el Sector 2. Los patrones observados en esta figura son similares entre sí.

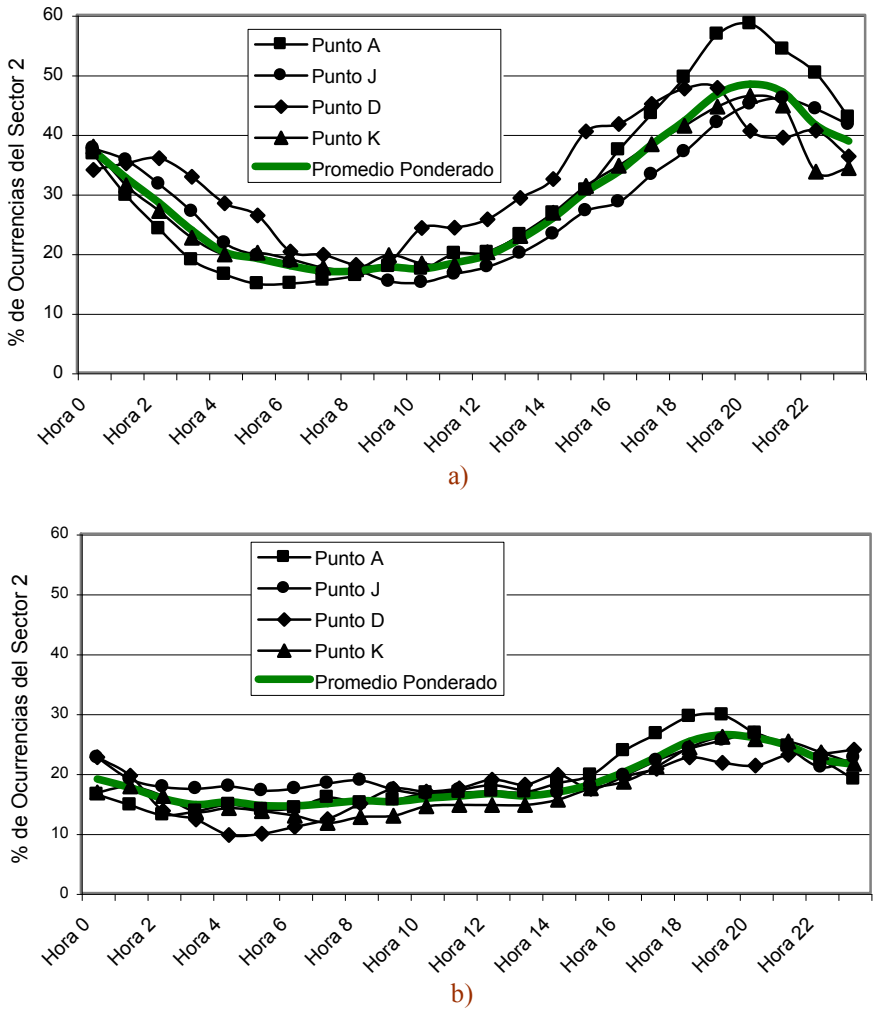


Figura IV.26: Frecuencias de ocurrencia del Sector 2 en distintos sitios y períodos de tiempo y la curva promedio. a) Verano (promedio ponderado total 29.3 %) b) Invierno (promedio ponderado total 18.6 %).

Los mismos muestran un máximo al anochecer entre las horas 18 y 21 (con un porcentaje promedio ponderado de ocurrencias del 46,3 % para el verano y un 25.1 % para el invierno), mientras que un mínimo en horas cercanas al amanecer. También aquí las estaciones cálidas tienen intensidad de picos mayores que las estaciones frías. El promedio ponderado general de ocurrencias del Sector 2 para los cuatro sitios y estaciones del año es de 24.4 %.

Curvas análogas del Sector 2 observadas en los puntos A y J solo en el período 1998- 2003 (no mostradas por cuestiones de espacio) muestran formas y valores porcentuales muy similares; además, los máximos de ocurrencias para las franjas horarias referidas en el párrafo anterior para este sector son en promedio 47.7 % en verano y 25.8 % en invierno. Esta comparación pone en evidencia que el Sector 2 posee un patrón generalizado en la zona de estudio (La Plata y alrededores), su curva promedio se halla representada con línea llena (verde) en la Figura IV. 26.

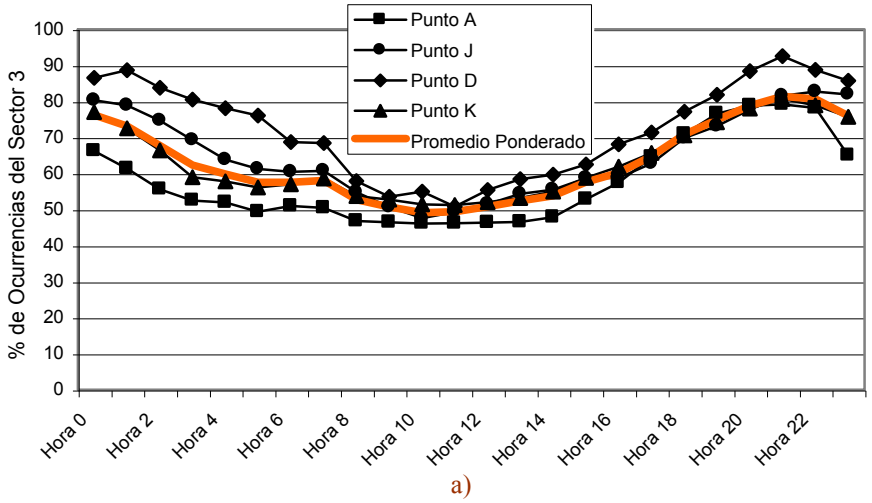
La Figura IV.25 permite apreciar que los contaminantes de origen industrial son transportados predominantemente hacia el casco urbano durante el mediodía y la temprana tarde mientras que finalizando la tarde y durante el anochecer (Figura IV.26) los mismos son transportados hacia los barrios residenciales del noroeste (Tolosa, Gonnet, City Bell, etc.). Vale decir que aquellos habitantes que tengan actividades durante el día en el Casco Urbano y hacia el atardecer se desplacen hacia dichas zonas residenciales son uno de los grupos potencialmente más afectados por la contaminación industrial (que se da en concomitancia con la urbana durante el día). Cabe agregar aquí que en el eje comprendido por las direcciones NE -SO entre el Parque Industrial de Ensenada y el Casco Urbano (Figura II.6- Capítulo II) se halla ubicado el "Paseo del Bosque" en donde tienen lugar eventos recreativos y a donde diariamente concurren muchos habitantes que realizan actividades aeróbicas.

Inspeccionando la Figura II.6 y considerando todos los sitios de observación (puntos A, D, I, J y K), los puntos A, I y J son receptores de contaminantes de origen industrial debido a las direcciones de viento del Sector 1 mientras que el Punto D lo es debido al Sector 2. El Punto K está afectado (en relación al transporte de contaminantes de origen industrial) solo por algunos vientos del Sector 1, es decir, el NNO y el N, y dada su distancia al área industrial parece ser el más adecuado para medir niveles de contaminación de fondo. Si además de las fuentes industriales se tiene en cuenta las de la ciudad (principalmente vehiculares) los sitios A, D, I y J reciben contaminantes desde un variado número de direcciones. El Punto K recibe contaminantes de la ciudad principalmente en el grupo de direcciones O- ONO- NO- NNO- N. Por otro lado, las direcciones ENE- OSO en el sentido horario son aquellas que no transportan contaminantes de origen industrial ni vehicular al Punto K, lo que hace de este sitio un lugar adecuado para adoptarse como referencia o control. A este nuevo conjunto de direcciones lo llamaremos Sector 3 (ENE-E-ESE-SE-SSE-S-SSO-SO-OSO).

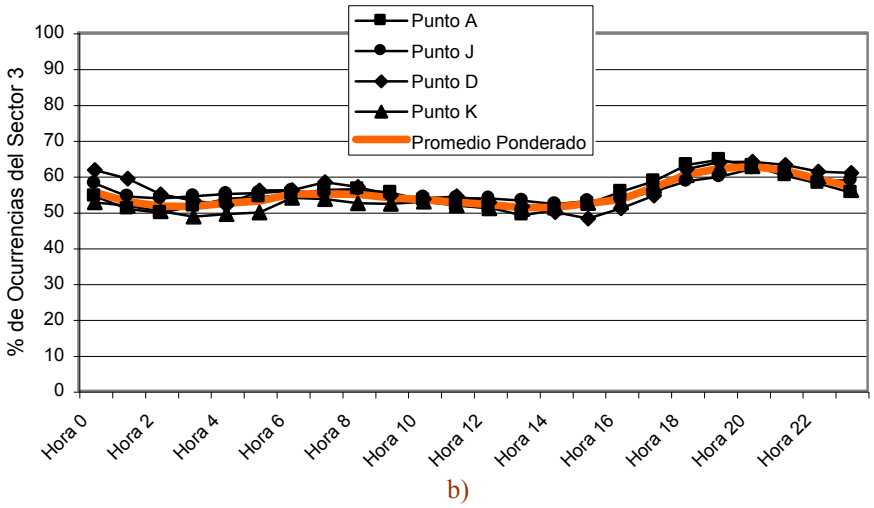
La Figura IV.27 muestra el patrón horario del Sector 3. El otoño y la primavera (no mostradas por cuestiones de espacio) muestran curvas similares e intermedias.

El invierno (Figura IV.27b) permite apreciar que la frecuencia de ocurrencias del Sector 3 es muy pareja a lo largo del día siendo el promedio general alrededor de un 10% menor que el del verano (máxima presencia del Sector 3). La forma de la curva para el verano (Figura IV.27a) deja ver la importancia de ocurrencia del Sector 2 dado que este se halla incluido en el Sector 3.

Dado que la ocurrencia del Sector 3 es en promedio alta (tomando al invierno como el caso más conservador su presencia es del 55.7 % y tomando al promedio de las estaciones -ciclo anual- es de 61.6 %) es posible concluir que el Punto K o sus cercanías son áreas recomendables para realizar el seguimiento de concentraciones de fondo.



a)



b)

Figura IV.27: Frecuencias de ocurrencia del Sector 3 en distintos sitios y períodos de tiempo y la curva promedio. a) Verano (promedio ponderado total 63.2 %) b) Invierno (promedio ponderado total 55.7 %).

Anexo IV.1
Estimador *M* de correlación

Dada una muestra de n observaciones de p dimensiones dada por el vector columna $x_i = (x_{i1},.....,x_{ip})$ con $i=1,....,n$ el objetivo es definir un vector de posición μ y una matriz p x p de covarianzas Σ que sean versiones robustas del vector de medias y de la matriz de covarianzas clásica. Con este fin se define a la distancia de Mahalanobis (Capítulo IV- Sección IV.2.2, pág. 66) como $d^2_{(x,\mu,\Sigma)} = (x-\mu)^T \Sigma^{-1}(x-\mu)$ donde el supraíndice T denota una matriz traspuesta. Si se toma una función no negativa $W(d)$ $(d \geq 0)$ entonces el estimador M queda definido implícitamente como una media pesada:

$$\mu = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \; x_i$$

y una matriz covarianzas pesada según:

$$\Sigma = \sum_{i=1}^{n} w_i (x_i - \mu)(x_i - \mu)^t \qquad \text{ec. 1}$$

en donde los pesos están dados por

$$w_i = W(d(x_i, \mu, \Sigma)) \qquad \text{ec. 2}$$

Notar que cuando $W(d)=1$ entonces μ y Σ constituyen el vector de medias clásico y la matriz de covarianzas clásica.
Para hacer que la estimación del coeficiente de correlación sea robusto se adoptó una función de peso W que tienda a cero en el infinito, o sea, $W(d) = \frac{p+1}{1+d^2}$.
El estimador M que corresponde a esta función es aquel de máxima verosimilitud para la distribución de Cauchy. Es de notarse que la distancia de Mahalanobis da una medida de "atipicidad" de los puntos p- dimensionales y por lo tanto su estima da menos peso a los valores atípicos.

La definición implícita del estimador M sugiere un proceso iterativo que puede resumirse así: dar un valor inicial para μ y Σ, calcular la distancia de Mahalanobis y luego los pesos con la ec. 2, actualizar μ y Σ con la ec. 1 y seguir así hasta lograr una convergencia.

Luego, el cálculo del coeficiente de correlación robusto, surge de considerar en la ec. IV.1 del Capítulo IV (coeficiente de Pearson- Sección IV.2.1, pág. 64), la μ y Σ obtenidas al final del cálculo.

Anexo IV.2
Una propiedad del *SAD*

Los vectores x_i *e* y_i con $i=1, p$ cumplen que

$$\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i = 100 \ (\%) \qquad \text{ec.1}$$

Sean

$$z_i = x_i - y_i \qquad \text{ec.2}$$

$$SAD = S = \sum_{i=1}^{p} |z_i|$$

$$M = \max |z_i|$$

definidos para $i=1, p$ se debe probar que $M \le \frac{S}{2}$

De (ec.1) y (ec.2) sale que $\sum_{i=1}^{p} x_i - \sum_{i=1}^{p} y_i = \sum_{i=1}^{p} x_i - y_i = \sum_{i=1}^{p} z_i = 0$

Supongamos un z_i cualquiera en particular llamado z_1 tal que $z_1=M$ (siempre es posible suponer esto).

Puesto que $\sum_{i=1}^{p} z_i = z_1 + \sum_{i=2}^{p} z_i = 0$ entonces es posible expresar $-z_1 = \sum_{i=2}^{p} z_i$

Entonces $M = |z_1| = \left| \sum_{i=2}^{p} z_i \right| \le \sum_{i=2}^{p} |z_i|$

por lo tanto

$$S = \sum_{i=1}^{p} |z_i| = |z_1| + \sum_{i=2}^{p} |z_i| \ge M + M = 2M$$

entonces queda que $M = S/2$

Esta propiedad se verifica dado que los datos (vectores) tienen expresadas sus variables en porcentaje.

Anexo IV.3
Método LOESS y tendencia de una serie

A continuación se realiza una descripción de la secuencia de pasos del método de LOESS (no paramétrico) empleado en la Sección IV.6.6 (pág. 90).

Dada una secuencia de observaciones (x_i, y_i) el procedimiento adopta para cada x dado dentro de un rango un valor de y. Se designa I a la ventana de ancho h alrededor de x: $I=[x-h, x+h]$. Para cada $x_i \in I$ se calculan los pesos según $w_i = W\left(\frac{|x_i - x|}{h}\right)$ donde W es la "función tricúbica" $W(x) = (1-|x|^3)^3$ para $|x| \leq 1$ y $W(x) = 0$ en los demás casos. Esta función se hace máxima para $x=0$ y decrece hasta cero para $x=1$. Entonces, para (x_i, y_i) con $x_i \in I$ se ajusta un polinomio de grado dos por cuadrados mínimos pesados, o sea, se encuentran los coeficientes $\beta_0, \beta_1, \beta_2$ tales que $\sum_{x_i \in I} w_i \ (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2 = \min$ (mínimo). Finalmente, se calcula $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$.

Lo usual es calcular el ajuste para cada observación obteniendo $\hat{y}_i = \hat{y}(x_i)$, pero el ajuste puede ser llevado a cabo para cualquier punto dentro del rango de las x. El procedimiento se denomina no paramétrico en cuanto a que $y = \hat{y}(x)$ no tiene una forma explícita y no pertenece a ninguna familia paramétrica de curvas. Para una introducción ver Fox (2000) y para mayores detalles y variantes Loader (1999).

Como se señaló antes, este tipo de regresión permite visualizar la tendencia de los datos pero es importante discriminar si su aplicación permite revelar aspectos de los datos o se constituye como un mero artefacto estadístico (el empleo del método produce un patrón artificial). Con esta finalidad es útil recurrir a la comparación de las medias de cada intervalo (Maronna, CP); esto se hace calculando el desvío estándar (de dichas medias). Para esto se debe tener en cuenta la falta de independencia de las observaciones consecutivas. Si se considera a $x_1, x_2, \ldots$ una secuencia estacionaria con varianza σ^2 y $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ entonces (Box et al., 2008) la varianza de las medias estará dada por $\mathrm{Var}(\bar{x}) = \frac{V\sigma^2}{n}$ donde V es un factor de "inflación" (Wilks, 2006) dado por $V = 1 + 2\sum_{k=1}^{n} \rho_k$ en donde ρ_k es el orden k de la autocorrelación de la secuencia.

El análisis de las observaciones sugirió (Maronna, CP) para el caso de aplicación de la Sección IV.6.6 (pág. 90) que la dependencia de variables quedaba bien representada por un proceso autoregresivo de primer orden, o sea, $\rho_k = \rho_1^k$ y por lo tanto $\sum_{k=1}^{\infty} \rho_k = \frac{\rho}{1-\rho}$.

Finalmente, los desvíos respecto de la media para cada ventana son obtenidos a partir de la varianza de la media así calculada.

"Cierro los ojos y veo una bandada de pájaros…¿Era definido o indefinido su número…?
Argumentum Ornithologicum, J. L. Borges (1960)

"Knowledge would be fatal, it is the uncertainty that charms one. A mist makes things beautiful"
The Picture of Dorian Gray, Oscar Wilde (1891)

"*...such knowledge has traditionally resulted from the pursuit of human curiosities...*"
Brewer (1999)

Capítulo V
Análisis por conglomerados y escalamiento multidimensional

El método de análisis por conglomerados jerárquicos ha sido aplicado al estudio de vientos en la ciudad de La Plata y alrededores (Ratto et al., 2010a; Ratto et al., 2010b) y en el Río de La Plata (Ratto et al., 2014a). En las dos primeras de estas citas fue utilizado para la detección de grupos en rosetas horarias de viento (patrones temporales) mientras que en la tercera fue utilizado para definir regionalidad (patrones espaciales). El método de análisis por escalamiento multidimensional (EMD) se empleó de manera simultánea con el análisis por conglomerados (Ratto et al., 2010b) para profundizar en la características de los patrones y la relación entre ellos.
Otros métodos de análisis exploratorio multivariado tales como Componentes Principales (CP) y las Curvas de Andrews fueron empleados para asistir a la discusión de la homogeneidad de los grupos hallados por análisis por conglomerados (Ratto et al., 2014b).
La aplicación de análisis por conglomerados utilizando restricciones, el método de las *k*-medias y el diagrama de las Siluetas se presentan, con menor grado de profundidad, para enriquecer la discusión de aspectos particulares o como enfoques alternativos.
Las secciones V.1 a V.7 presentan y discuten los principales aspectos teóricos de los métodos estadísticos utilizados en las publicaciones referidas (y brindan ejemplos) mientras que la Sección V.8 está dedicada al trabajo de campo que fue motivo de la aplicación de los distintos métodos.

V.1 Análisis por conglomerados

Muchas actividades de investigación dependen de encontrar objetos parecidos (Anderberg, 1973; Romesburg, 2004). En el pasado el agrupamiento de objetos se realizaba de manera subjetiva según el criterio del investigador y estaba limitado al estudio de objetos descriptos por hasta tres variables (o dimensiones). La necesidad de contar con herramientas más objetivas que incluyeran muchas variables impulsó el desarrollo y la proliferación de algoritmos y programas en las últimas décadas (Kaufman y Rousseeuw, 2005).
Las técnicas numéricas de clasificación que se habían originado en las ciencias naturales con el nombre de taxonomía numérica (Everitt et al., 2011) fueron adoptadas por diversas disciplinas tales como las de estudio de mercado (segmentación), psicología (Q análisis), psiquiatría, meteorología, astronomía, arqueología, bioinformática, robótica y genética. La frase en inglés "cluster analysis" es el término en común con el que se difundió y generalizó un conjunto amplio de métodos de partición o aglomeración de datos, también llamados de clasificación no supervisada (Peña, 2002), de reconocimiento no supervisado de patrones (Escudero, 1977) o de aprendizaje no supervisado (Dudoit y Fridlyand, 2002). El término "no supervisado" refiere a que no se parte de un conocimiento *a priori* de los datos (Edelstein, 1999).

El análisis por conglomerados implica investigar la existencia de estructura en los datos sin la ayuda de una variable dependiente (Tibshirani y Walther, 2005). Dadas la atracción y la utilidad del análisis por conglomerados son diversas las disciplinas que han hecho aportes metodológicos, a veces sin buena comunicación entre ellas, produciendo hipótesis o algoritmos muy similares a los existentes (Xu y Wunsch, 2009). Seber (1984) dice que la literatura sobre análisis por conglomerados constituye un cuerpo inmanejable mientras que Jain et al. (2000) señalan que la variedad de métodos y sus propias variantes proporciona riqueza a la vez que confusión. Esta diversidad de métodos no solamente dificulta la elección por parte de los usuarios, sino que hace difícil las comparaciones entre resultados de distintos autores (Gan et al., 2007). Según Mirkin (2005) este tipo de análisis se puede aplicar desde distintas perspectivas, entre ellas, una que hace énfasis en la estadística (inferencial) y otra que busca la exploración de los datos (tema presentado en la Sección I.1.4).

Desde la perspectiva *estadística* cualquier conjunto de datos se considera una muestra de una distribución cuyas propiedades necesitan ser estimadas. Se sigue el paradigma de suponer una hipótesis sobre el fenómeno de estudio y luego se chequea el grado de cumplimiento de dicha hipótesis (ajuste y testeo de un modelo estadístico), es decir, los métodos se emplean para hacer inferencias o con fines confirmatorios (Kaufmann y Rousseeuw, 2005). El problema de esta perspectiva es que, en muchos de los casos en que se requiere realizar un análisis por conglomerados, se sabe poco de los fenómenos involucrados o de las variables más relevantes y por lo tanto hacer suposiciones puede ser muy arbitrario. Además, en muchos casos el conjunto de datos a analizar es único (por ejemplo, "países de Europa") y no puede considerarse como una muestra de una población. También suele ocurrir que más de una distribución conocida "ajusta bien" por lo que será algo arbitrario elegir una u otra.

Desde la perspectiva del *análisis exploratorio* no es fundamental conocer el origen de los datos, no se busca ajustar un modelo (Jain y Holmes, 2011) sino poder encontrar aspectos singulares de los datos y resumirlos de tal manera de hacerlos entendibles y útiles para el usuario (Hand et al., 2001). La perspectiva exploratoria se apoya en descubrir "lo que dicen los datos", permitiendo encontrar patrones, revelar estructuras y proponiendo un "modelado" tentativo (Behrens, 1997). Debido a su flexibilidad es posible analizar grupos cuando la cantidad de variables es mayor que el número de datos (Everitt et al., 2011).

En el presente trabajo de tesis los métodos estadísticos de análisis multivariado (análisis por conglomerados, escalamiento multidimensional, componentes principales, etc.) han sido aplicados mayormente desde el punto de vista del análisis exploratorio, para asistir a la descripción de fenómenos ambientales. Según Tukey (1977) el análisis exploratorio es más una actitud que un conjunto de herramientas, es una manera de mirar los datos. La palabra heurístico (del griego "hallar", "inventar" etimología que es compartida por la palabra "eureka") alude a la búsqueda de la solución de un problema por métodos no rigurosos (DLE, 2003). El criterio heurístico, típico del análisis exploratorio, incluye métodos de visualización y de cálculo que permiten alcanzar el objetivo de estudio. Este enfoque comprende que dentro de las herramientas con las que se cuenta puede no haber una o algunas que sean "las mejores".

El análisis de conglomerados es el arte de encontrar grupos en un conjunto de datos (Kaufman y Rousseeuw, 2005) en el sentido de revelar la presencia de dichos grupos (Everitt et al., 2011) o de establecer si los datos originales pueden resumirse o representarse por un pequeño número de casos (Gordon, 1999).

Dado que el número de alternativas para dividir una muestra de n datos en k grupos está dada por (Rencher, 2002):

$$N(n,k) = \frac{1}{k!}\sum_{i=1}^{k}\binom{k}{i}(-1)^{k-i}\ i^n \approx \frac{k^n}{k!}$$

donde $k!$ es el factorial de k. Siendo, por ejemplo, $N(n,k)$ para n=25 y k=10 un número muy grande ($\approx 10^{18}$), el análisis por conglomerados tiene entre uno de sus objetivos, buscar una manera computacionalmente eficiente de encontrar los potenciales grupos en una muestra dada.

Es difícil dar una definición relevante de lo que constituye un grupo ("cluster") ideal dado un conjunto de datos (Gordon, 1999), pero típicamente refiere a un conjunto de objetos que tienen tal grado de cohesión que se parecen entre sí (es decir, mayor que los objetos que quedan afuera) (Mirkin, 2005). O sea, existe un grado de cohesión interna dentro de cada grupo y un grado de aislamiento de cada grupo respecto de los otros (Escudero, 1977; Gordon, 1999; Timm, 2002; Everitt et al., 2011; Ritter, 2015).

Existen una gran variedad de métodos de análisis de conglomerados (Gan et al., 2007; Everitt et al., 2011) pero desde un punto de vista práctico se pueden considerar dos grandes abordajes (Timm, 2002; Moii y Sarstedt, 2011): los métodos de partición (no jerárquicos) y los métodos jerárquicos. Una característica particular del análisis por conglomerados, en relación a otros métodos de análisis multivariado (análisis discriminante, componentes principales, regresión, correlación, etc.), es la de como opera la variación. Mientras que en los otros métodos la estimación de la variación (típicamente la varianza) viene dada por el método, en el análisis por conglomerados es el investigador quien debe especificar dicha variación (por ejemplo, eligiendo una distancia o un coeficiente de correlación –Sección V.5.3-) (Hair et al., 2010). O sea, el foco del análisis por conglomerados está en distinguir objetos (o variables) basándose en algo que cuantifique la variación, pero no en la estimación de la variación en sí misma.
Un objeto puede ser descripto en relación a sus variables (un conjunto de objetos generará una matriz de dos modos) o por su relación con otros (matriz de un modo). Los métodos de partición (por ejemplo, el popular método de las k-medias o el PAM "partición alrededor de medioides" propuesto por Rousseeuw (1987) que es más robusto) necesitan partir de la matriz de datos (matriz de dos modos) y definir *a priori* el número de grupos; dan una única solución. Los métodos jerárquicos pueden partir de una matriz de similitudes o disimilitudes (matriz de un modo), proveen una estructura de agrupamiento sucesivo y despliegan un conjunto de soluciones. Los métodos jerárquicos pueden ser aglomerativos (cuando parten de un conjunto de n objetos y de forma sucesiva se agrupan hasta que formen un solo grupo) o divisivos (cuando se parte de un grupo que contiene a todos los objetos y se van obteniendo sucesivamente subgrupos hasta llegar a identificar a cada uno de los objetos). Los métodos divisivos son poco usados (Romesburg, 2004; Wilks, 2006; Everitt et al., 2011) puesto que insumen muchos más cálculos que los aglomerativos, además, son complejos los algoritmos para hallar la primera bipartición y se debe definir un criterio para las sucesivas particiones (Gan et al., 2007). Esto hace que estén poco disponibles en los softwares comerciales. Sin embargo, Kaufman y Rousseeuw (2005) y Hastie et al. (2011) señalan que la ventaja potencial del divisivo frente al aglomerativo aparece cuando se buscan unos pocos grupos en un conjunto grande de datos (el aglomerativo "cometerá más errores" hasta llegar a pocos grupos). Por su parte, Maronna (CP) señala que esto último no es generalizable.

V.2 Conglomerados jerárquicos

Agrupar datos de forma jerárquica (proceso aglomerativo) es la forma más antigua y popular de hacerlo (Gong y Richman, 1995); tiene actualmente una gran vigencia dentro de las herramientas de procesamiento de datos (Mirkin, 2005). La palabra jerarquía implica una estructura anidada donde los objetos se van agrupando sucesivamente (asociación en cadena) y donde los niveles superiores contienen a los inferiores. Esta jerarquía puede representarse mediante un dendograma (del griego "dendro" (árbol) y "grama" (gráfico)).

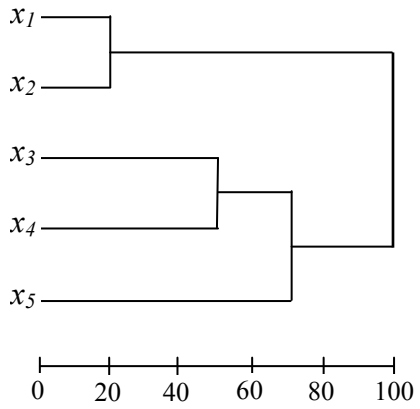


Figura V.1: Ejemplo de Dendograma

Esta forma gráfica de representación en dos dimensiones ilustra las "fusiones" entre individuos y grupos paso a paso, o sea, describe el proceso mediante el cual la jerarquía fue obtenida (Everitt et al., 2011).

La Figura V.1 muestra 5 objetos o vectores p-dimensionales iniciales x_i. El eje de las X en escala de 0 a 100 representa una medida de disimilitud entre objetos o grupos. En un primer paso de aglomeración (alrededor del 20%) se fusionaron x_1 y x_2. Luego se fusionaron x_3 y x_4. Posteriormente x_5 se unió al grupo preexistente de x_3-x_4 y finalmente, el grupo x_1-x_2 se fusionó con el grupo x_3-x_4-x_5 (100% de la escala) para dar lugar a un único grupo formado por 5 miembros.

El dendograma es, por su claridad, la forma más difundida de representación de jerarquías y según Brereton (1992) es la forma más informativa de presentarlas. Por lo tanto, este tipo de representación ha sido adoptada como la principal para mostrar y comparar los distintos procesos de aglomeración. Volviendo a la Figura V.1 se puede notar que la distancia de x_5 a x_4 en el dendograma (distancia "vía el dendograma") está dada por la "altura" de la rama correspondiente a x_5 que difiere de la "distancia directa" entre el par de objetos la cual queda velada (Legendre y Legendre, 1998). Esta "deficiencia" se subsana recurriendo a la matriz original de distancias entre pares de objetos. Este tema se ampliará al tratar el coeficiente cofenético (Sección V.5.6.2 y Anexo V.3, pág. 180). Otras vías de representación se detallan en el Capítulo 7 de Gan et al. (2007), entre ellas el esquema de aglomeración (o "icicle") que se muestra en el Anexo V.3.

El agrupamiento jerárquico tiene ventajas tales como permitir visualizar de manera integral estructuras (de objetos o variables) según las similitudes/disimilitudes. En el caso en que, por la naturaleza del fenómeno, sea esperable una cierta taxonomía, el agrupamiento jerárquico permite modelar la clasificación. En relación a los métodos no jerárquicos la principal ventaja del jerárquico se halla en no tener que definir de antemano el número de grupos (Timm, 2002). Tiene como principal desventaja su rigidez en cuanto a que, una vez que dos individuos se han agrupado, ya no pueden separarse en etapas posteriores, o sea, el método no puede "reparar" lo que hizo en pasos previos (Kaufman y Rousseeuw, 2005). Esta forma de agrupar anidando (jerarquía indexada) sin volver atrás hace que la homogeneidad dentro de un grupo vaya decreciendo a medida que se agregan nuevos individuos. En cambio en los métodos de partición –mientras se lleva a cabo el proceso de agrupamiento- un individuo puede dejar de pertenecer a un grupo para pasar a pertenecer a otro y así optimizar algún criterio de homogeneidad intragrupo hasta alcanzar el número predefinido de grupos.

V.3 Medidas de similitud y disimilitud

Es de importancia central definir una medida de proximidad (similitud o disimilitud) entre objetos como un primer paso para encontrar grupos. Los conceptos de similitud y disimilitud ya han sido tratados en el Capítulo IV. Aquí se presentan algunas consideraciones que tienen influencia en el análisis por conglomerados.

Dados tres objetos (vectores) cualesquiera x_r , x_s y x_h de un conjunto de datos en el espacio p-dimensional es importante, como punto de partida, tener en cuenta las propiedades básicas que una medida de **disimilitud** pueda cumplir:

1) $d_{rs} \geq 0$ para todos los objetos r y s (no negatividad)

2) $d_{rr} = 0$ (identidad)

3) $d_{rs} = d_{sr}$ (simetría)

4) $d_{rs} \leq d_{rh} + d_{hs}$ (desigualdad triangular)

La condición (1) implica trabajar solo con valores positivos, la distancia de r a s igual a cero no implica que el objeto sea el mismo sino que puede haber más de un objeto con las mismas coordenadas. La condición (2) implica simplemente que la distancia de un objeto a si mismo es cero. La condición (3) no permite trabajar con la matriz confusión (aunque esta podría "simetrizarse"). Cuando se cumplen solo las condiciones 2) y 4) se habla de disimilitud semimétrica. La condición (4) implica que yendo "directo" de r a s se realiza un camino más corto que pasando por otros puntos (por ejemplo, h). Estas condiciones posibilitan que haya una interpretación geométrica de las relaciones entre objetos.

Cuando se cumplen con las cuatro propiedades simultáneamente se dice que la medida de disimilitud es una métrica lo cual posibilita (en el caso de las distancias) la interpretación física (Everitt et al., 2011). Cambiando la cuarta propiedad por una condición más restrictiva $d_{rs} \leq máx(d_{rh}, d_{hs})$ se habla de magnitudes ultramétricas. Todas estas propiedades pueden ser importantes cuando se evalúan las ventajas y desventajas de la utilización de los distintos coeficientes de similitud o disimilitud, sin embargo, como señalan Seber (1984), Timm (2002) y Kaufman y Rousseeuw (2005), no son condiciones esenciales para llevar a cabo un análisis por conglomerados (se recordará esto al tratar con restricciones en la Sección V.8.4). Veltkamp y Lateecki (2006) agregan que estas propiedades no son siempre útiles, por ejemplo, cuando se busca una concordancia o encaje parcial entre formas no se requerirá el cumplimiento estricto de la desigualdad triangular. La condición de simetría puede ser de poca importancia cuando lo que se están evaluando son percepciones. Finalmente, Legendre y Legendre (1998) destacan que el proceso de aglomeración puede ser llevado a cabo sin tener como referencia las características del espacio. Otras propiedades tales como de invariancia a algún tipo de transformación, pueden ser requeridas según el tipo de objeto que se estudie (Veltkamp y Lateecki, 2006).

Análogamente, dados dos objetos (vectores) cualesquiera x_r, x_s de un conjunto de datos en el espacio p-dimensional es importante, como punto de partida, tener en cuenta las propiedades básicas que una medida de **similitud** pueda cumplir:

1) $0 \leq s_{rs} \leq 1$ para todos los objetos r y s (coeficiente acotado)

2) $s_{rr} = 1$ (la correlación consigo mismo es máxima)

3) $s_{rs} = s_{sr}$ (simetría)

Tanto las medidas d_{rs} como las s_{rs} pueden provenir de distintas fuentes y pueden ser el resultado de evaluaciones subjetivas o de combinación de variables de distinto tipo. En el caso en que se puedan definir dependerá marcadamente de lo que se necesita discriminar en la aplicación.

Es típico en los métodos de análisis por conglomerados encontrar que la medida de

disimilaridad sea una distancia Euclídea y que la medida de similitud sea un coeficiente de correlación (típicamente los coeficientes de Pearson o Spearman- Sección IV.1.1) (Romesburg, 2004). Puesto que todas las aplicaciones de métodos de análisis multivariado llevadas a cabo en la presente tesis involucran variables continuas (sus valores están dados en intervalos de números reales) se describen algunas de las medidas de similitud y disimilitud más utilizadas para operar con dichas variables.

Se define la distancia de Minkowsky como:

$$d_{rs}^{Mink} = \left(\sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|^{\lambda} \right)^{1/\lambda}$$

donde según el valor de λ se acentúan las mayores o menores distancias, para cualquier valor de λ la distancia de Minkowsky es una métrica. Cuando λ=1 se obtiene la distancia City-block (también llamada Manhattan - Figura V.2). Esta distancia es no Euclídea (Husson et al., 2011) pero suele tener aplicaciones específicas en análisis por conglomerados. Para λ=2 se obtiene la distancia Euclídea (también llamada Pitagórica o distancia directa- Figura V.2). Una galería de distancias y sus propiedades se presenta en Gan et al. (2007). Capítulo 6. El coeficiente de correlación de Pearson definido en el Capítulo IV (Sección IV.1.1) tiene el inconveniente de presentar valores negativos (entre –1 y 0).

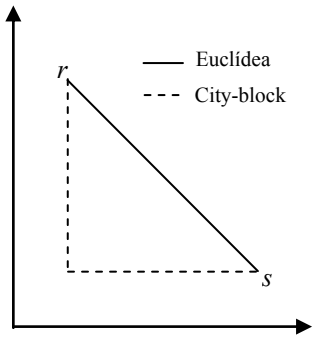


Figura V.2: Casos particulares de distancias de Minkowsky.

Para que pueda ser utilizado como medida de similitud en análisis por conglomerados debe realizarse alguna transformación. Se ha sugerido (Kaufman y Rousseeuw, 2005), $s_{rs} = (1 + \rho_{rs})/2$ en donde siempre que ρ_{rs} de –1 s_{rs} dará 0, para los casos en que los valores negativos y positivos de correlación tengan significados distintos. Para cuando tanto valores cercanos a –1 como valores cercanos a 1 tengan un significado análogo (por ejemplo, cuando se desea reducir el número de variables) se ha sugerido la transformación $s_{rs} = |\rho_{rs}|$.

La forma de operar en la mayoría de los programas de conglomerados jerárquicos es con distancias. Si las relaciones entre pares de objetos vienen dadas por correlaciones, se puede recurrir a algún tipo de transformación adecuado al caso. Una de las más difundidas es $d_{rs} = (1 - \rho_{rs})/2$ (Kaufman y Rousseeuw, 2005). Pero debe tenerse en cuenta que puede "afectarse" la eficiencia de discriminación.

A continuación se citan ejemplos de cuando conviene trabajar con distancia o correlación según el objetivo de la aplicación (tomado del Capítulo 8 de Romesburg (2004)).

Ejemplo 1: Un conjunto de objetos (vectores) representan plantas que se hallan sembradas en parcelas de tierra (un vector por parcela), las variables medidas a lo largo del tiempo dan cuenta del crecimiento de las plantas. Si se desea detectar la presencia de grupos homogéneos de plantas y diferenciar las parcelas con distinto grado de crecimiento será apropiado trabajar con distancias. La distancia es sensible a los tamaños, o sea, cuantifica el crecimiento de las plantas. Puesto que todas las plantas crecerán, algunas parcelas crecerán mucho y otras poco; esto será el factor fundamental de discriminación. De aplicarse correlación la misma sería insensible a la diferencia y no sería apta para discriminar grupos.

Ejemplo 2: Si se desea comparar tendencias en la acumulación de stock según precios de productos entre empresas similares (precios paralelos) no importará el tamaño del stock (que se halla en relación al tamaño de la empresa). La correlación indicará cuales son las empresas que siguen la misma tendencia en la acumulación de stock (aumentando o

disminuyendo el mismo) aunque los volúmenes sean muy distintos.
En general, en análisis jerárquico por conglomerados las distancias son más aplicadas cuando se buscan grupos de objetos mientras que las correlaciones se utilizan cuando se buscan grupos de variables.

V.4 Criterios de agrupamiento

En alguna fase del proceso de agrupamiento se deberá estimar una medida de disimilitud entre un objeto y un grupo o entre grupos. Basados en la medida de similitud o disimilitud adoptada entre pares de objetos debe definirse una "regla", "estrategia" "método" o "criterio" que permitan relacionar los grupos y los objetos (Anderberg, 1973).
Varios criterios se han constituido como "clásicos" en la literatura. Ellos son el Enlace Simple ("single linkage"), el Enlace Completo ("complete linkage"), el Enlace Promedio (UPGMA por sus siglas en inglés- Ver Anexo V.1, pág. 171), el Enlace Centroide y la Regla de Ward. Una breve presentación y discusión de estas alternativas, que están a disposición del investigador, se halla en el Anexo V.1. Resta agregar aquí, que no existe un criterio que sea universalmente más recomendable y que la aplicación de distintos criterios conducirá a resultados distintos.

V.5. Pasos en la implementación del análisis por conglomerados

Llevar a cabo un proceso de análisis de conglomerados involucra varias etapas. Antes de pasar a describir, discutir y dar ejemplos de aplicación de las mismas es oportuno realizar una digresión para dejar contextualizado el empleo del método de Componentes Principales (CP).

> La idea central del análisis por CP es llevar a un conjunto de datos de *p* variables (en mayor o menor grado correlacionadas entre sí) desde un sistema de coordenadas a otro, en el que las nuevas variables se hallan incorrelacionadas. Las nuevas variables son llamadas componentes principales (habrá tantas CP como variables originales) y sus características son tales que con pocas de estas nuevas variables se explica la mayor parte de la variación que presentaba el conjunto de datos en sus variables originales. De aquí que este método suela aplicarse para "reducir" dimensionalidad.
> El método de análisis por CP es uno de los más antiguos del análisis multivariado (Jolliffe, 2002) y es quizás el que haya sido más empleado en las ciencias ambientales (Wilks, 2006). EL análisis por CP es usado frecuentemente para reducir dimensionalidad como paso previo a la aplicación de otros métodos (Affifi y Clark, 1998). En el contexto del análisis por conglomerados un uso difundido (Jolliffe, 2002) es el de proveer una representación gráfica de los datos para investigar la presencia o ausencia de estructura de grupo (carácter exploratorio). Algunos autores (Lavine, 2000) recomiendan la reducción de dimensionalidad con CP como paso previo al análisis por conglomerados. Sin embargo, este uso se debe realizar con precaución, puesto que no hay garantía de que la separación entre grupos esté siempre en la dirección de las CP de mayor varianza (Jolliffe, 2002). Chang (1983) muestra un ejemplo de como algunas CP de baja varianza pueden ser importantes para la discriminación de grupos (en lugar de aquellas con alta varianza). Por su parte, Yeung y Ruzzo (2001) señalan que reducir dimensionalidad con el método de CP, previo a realizar un análisis por conglomerados, puede no solamente ser indistinto para la discriminación de grupos sino que puede degradar la calidad de los resultados.

En la presente tesis este método se aplicó con fines exploratorios como herramienta complementaria para detectar potenciales valores atípicos (Sección V.5.2.4.3) y como herramienta auxiliar para representar Curvas de Andrews (Sección V.8.3). Una breve descripción del mismo se halla en el Anexo V.2 (pág. 175), para una descripción más rigurosa ver Timm (2002) y para un tratamiento en detalle ver Jolliffe (2002).

Basándose en un trabajo de Milligan, Everitt et al. (2011) proponen un conjunto de pasos que pueden ser tenidos en cuenta al realizar el análisis por conglomerados. Los pasos que se muestran a continuación siguen otro orden al dado por los autores, pero se mantienen los contenidos en el contexto de la experiencia y de las aplicaciones llevadas a cabo.

V.5.1. Objetos a ser analizados
V.5.2. Transformación de datos
  V.5.2.1 Selección de variables
  V.5.2.2 Asignación de pesos a las variables
  V.5.2.3 Tratamiento de datos faltantes
  V.5.2.4 Detección de valores atípicos
  V.5.2.5 Estandarización
V.5.3 Criterio de aglomeración
V.5.4 Procedimiento de aglomeración
V.5.5 Determinación del número óptimo de grupos
V.5.6 Validación
V.5.7 Interpretación

V.5.1 Objetos a ser analizados

El investigador tiene un conjunto de datos o individuos (vectores *p*- dimensionales) de una muestra o población en donde necesita conocer la estructura de grupo y determinar los grupos presentes. Puesto que el análisis por conglomerados no es una herramienta inferencial, la muestra no necesita ser "representativa" de una población aunque, cuanta mayor información disponible haya, el resultado será más generalizable.

V.5.2 Transformación de datos

El proceso de transformación de datos para el abordaje de un análisis por conglomerados –y en general para cualquier análisis multivariado- es un paso que el analista no debe evitar plantearse. El mismo puede ir desde la aplicación de un simple método de homogeneización de variables (escalamiento), la estandarización de los datos iniciales o la transformación de los mismos según alguna distribución conocida hasta la aplicación de un conjunto de métodos (Legendre y Legendre, 1998). Esta amplitud de posibilidades guarda relación con la naturaleza de los datos, el método que se utilizará (jerárquico, *k*- medias, etc.) y el objetivo de la aplicación. Los procesos de transformación de datos más comunes pueden comprender la selección de variables, la asignación de pesos, el tratamiento de datos faltantes, la detección de valores atípicos y la estandarización de las variables. En el caso en que el resultado del análisis por conglomerados deba cumplir con ciertas restricciones (Sección V.8.4) pueden requerirse métodos adicionales de tratamiento de los datos iniciales.

V.5.2.1 Selección de variables

Las variables que no contienen información relevante pueden afectar de manera adversa el proceso de revelar estructura de grupo en los datos (Kaufman y Rousseeuw, 2005). La selección de las variables puede considerarse un caso de asignación de pesos (Sección

V.5.2.2) en donde se asigna peso nulo a las variables que se desea eliminar y peso unitario a las que se incluyen. En muchos casos el investigador selecciona las variables en base al conocimiento que tiene sobre el tema y a los objetivos de la investigación. Pero también existen métodos de selección de variables desde el punto de vista del análisis exploratorio (Everitt et al., 2011) que son específicos cuando el objetivo es realizar un posterior análisis por conglomerados (Gnanadesikan et al., 1995; Jolliffe, 2002). Además puede recurrirse a alguno de los métodos que permiten reducir la dimensionalidad de los datos iniciales tales como el de componentes principales (Chae y Warde, 2006) ya mencionado.
Mooi y Sarstedt (2011) consideran como regla empírica un valor tope del coeficiente de correlación entre pares de variables. Si dicho coeficiente es mayor a 0,90 es problemático dejar ambas variables (colinealidad), puesto que los aspectos que representan quedarían sobrerepresentados en la salida del análisis por conglomerados. También es posible, con carácter exploratorio, realizar un análisis por conglomerados de las variables (Khattree y Naik, 2000) para tener un panorama de cuales tienden a formar grupos. Mayor grado de detalle puede encontrarse en el Capítulo 5 de Theodoridis y Koutroumbas (2003).

V.5.2.2 Asignación de pesos a las variables

Darle peso a una variable implica darle mayor o menor importancia relativa frente a las otras. Esta asignación repercutirá en como se juzgará la similitud de los objetos analizados. Los pesos pueden asignarse según el criterio del investigador o recurriendo a la matriz de datos. Everitt et al. (2011) citan, en el Capítulo 3, una variada referencia bibliográfica dedicada a este tema. Un caso particular de asignación de pesos, que suele darse en la práctica, es cuando se cambian las magnitudes originales de algunas o todas las variables a otras nuevas. Cuanto más pequeña es la unidad (por ejemplo, pasar de pies a milímetros) se incrementará más el rango de la variable en cuestión y esto afectará a la estructura resultante. Varios ejemplos de esto, en relación al análisis por conglomerados, pueden verse en el Capítulo 1 de Kaufman y Rousseeuw (2005).

V.5.2.3 Tratamiento de datos faltantes

En textos tales como WMO (1983), Bower (1997), Allison (2001), EPA (2006) y Kondrashov y Ghil (2006) se subraya la importancia de considerar la completitud de datos que se van a procesar. Existen varias causas por las cuales se producen ausencias de algunos datos en el conjunto original (no se registró el dato, se registró de manera incompleta, se perdió, etc.). Dichas ausencias se producen, en términos generales, en los casos (objetos) o en las variables. La presencia de datos faltantes puede producir un debilitamiento en la confiabilidad los datos básicos (se pierde sistematicidad), pero también puede debilitar la validez de las conclusiones del análisis que se lleva a cabo (por ejemplo, sobre la relación entre las variables) y puede limitar la representatividad del alcance del estudio (McKnight et al., 2007). En general, los distintos autores recomiendan completar los datos, sin embargo, ningún método de relleno es inocuo; el simple reemplazo del valor ausente en una variable por la media muestral reducirá la varianza de esa variable y por lo tanto exacerbará la similitud entre los individuos (Krzanowski, 2007), lo cual hará más difícil la discriminación de grupos. Este sencillo ejemplo, muestra la relevancia que puede tener la intervención del investigador en el tratamiento inicial de los datos. Para el caso de análisis por conglomerados la mayoría de los métodos de relleno de datos lo hacen durante el pre-procesamiento (Mirkin, 2005).
En la presente tesis el tratamiento de datos faltantes fue de secundaria importancia y se pudo abordar con métodos sencillos que se discutirán en el particular. Sin embargo, debe tenerse en cuenta que es un tema complejo tal como lo demuestra la abarcativa obra de Little y Rubin (1987).

V.5.2.4 Detección de valores atípicos

Al igual que en el caso univariado, el interés en la detección de valores atípicos en un sistema multivariado, reside en su posterior análisis tendiente a calificarlos en el contexto del conjunto de datos según las características de la investigación (sentido físico, causas posibles, etc.) y finalmente en la adopción una decisión (descartarlo, corregirlo, dejarlo identificado y/o elegir un método típico o un método robusto).

Como se vio en el Capítulo IV (Sección IV.1), en una muestra de datos de una sola variable aleatoria, la noción de valor atípico puede quedar bien definida, cuando ordenados los datos de manera creciente o decreciente, se identifican al valor mayor y al menor. Pero, cuando se trata de datos multivariados no hay una manera tan unívoca de definir el orden, por lo que la noción de valor atípico se torna más compleja; a esto se le agrega el hecho de que cuanto más pequeña es la relación *n/p* (casos/variables) los valores atípicos se hacen menos evidentes (Sajesh y Srinivasan, 2013). Una observación que no es un valor atípico en ninguna de las variables puede ser un verdadero valor atípico cuando se consideran todas las variables en conjunto. Este hecho, señala Jolliffe (2002), es el mayor problema en detectar valores atípicos en casos multivariados. Aun representando las *p* variables de a pares habrá casos en que el valor atípico no se detectará porque la variación aparece en otra dirección que la que imponen los ejes cartesianos. Esto último lleva a pensar en la implementación de otras herramientas de exploración tales como el método de Componentes Principales (Sección V.5.2.4.3). Por otra parte, Barnett y Lewis (1978) presentan una manera para definir un orden en el espacio multidimensional cuyo uso se ha generalizado (Sección V.5.2.4.2).

Desde el punto de vista de la exploración de datos, Maronna (CP) recomienda poner en práctica varias herramientas simultáneamente. Las posibilidades de que un sistema multivariado se aleje de la distribución multinormal son muchas y variadas (Gnanadesikan, 1997). Una observación que se destaque en más de una de las herramientas aplicadas para la exploración se constituye en una firme candidata a valor atípico (Barnett, 2004).

Tomando como ejemplo los datos horarios anuales acumulados de rosetas de frecuencias de ocurrencia de vientos por dirección correspondientes al Punto A (1997- 2000) publicados en Ratto et al. (2010a), se mostrará una operativa posible de investigación de valores atípicos sugerida por Maronna (CP). La misma se basa en el análisis de los gráficos cuantil- cuantil, en el cálculo de distancias (Euclídea y Mahalanobis) a la media y en el método de Componentes Principales.

V.5.2.4.1 Gráficos cuantil- cuantil

Barnett y Lewis (1994) señalan que en la identificación de valores atípicos de un sistema multivariado no se debe desestimar el análisis de cada una de las variables (variables marginales). Cuando no se conoce el tipo de distribución que pueden tener las variables (por ejemplo, de las frecuencias de viento por dirección) corresponde suponer que se comportan normalmente. La condición necesaria (aunque no suficiente) para que una distribución multivariada sea normal (multinormal) es que las distribuciones marginales (o sea, de cada una de las variables componentes del sistema multivariado) sean normales (Thode, 2002). Este hecho fundamenta la estrategia de explorar la distribución multivariada a partir de las marginales al mismo tiempo que establece su alcance.

El gráfico cuantil- cuantil (o "QQ-Plot") ya fue utilizado en el Capítulo IV (Sección IV.6.4). En el presente caso de estudio (datos anuales de rosetas horarias de frecuencias de viento observadas en el Punto A durante 1997- 2000 empleados en Ratto et al. (2010a)) se cuenta con 16 variables (direcciones de viento); es posible tomar a cada una por separado y evaluar su relación con la distribución normal. En la Figura V.3a se muestra el gráfico cuantil- cuantil para la variable N (norte) que cubre las 24 horas del día (Sección V.5.2.4).

Los puntos circulares (azules) son las observaciones. La línea (roja) es una recta de regresión. La exploración mediante gráficos cuantil- cuantil permite detectar el tipo de apartamiento de la distribución teórica que se ensaya (en este caso la normal). Los apartamientos pueden presentarse fundamentalmente como desvío de las colas, en la forma que adopta la nube de puntos y/o en la presencia de puntos alejados del patrón de la nube (Thode, 2002).

Puesto que para una distribución normal los valores se estandarizan como $z = (x-\mu)/\sigma$, la ordenada en este gráfico queda expresada como $x = \mu + \sigma z$; o sea, la ordenada al origen es la media de la distribución mientras que la pendiente es el desvío estándar (recta a 45 grados en el caso de una distribución normal perfecta). Para el caso de una muestra estos valores se corresponden con $\bar{x}$ y s (media y desvío estándar muestrales) y se pueden inferir del gráfico. La recta permite también inferir el valor de los percentiles a partir del eje de las X (superior).

En la Figura V.3a la distribución de los datos (empírica) se aproxima a la normal (teórica) representada por la línea recta. No hay evidencia de valores atípicos (los mismos deberían aparecer por debajo de la recta en el extremo izquierdo (cuantiles bajos) y por encima de la recta hacia el extremo derecho (cuantiles altos) y ser puntos notablemente alejados del resto. La Figura V.3b apoya lo observado en la Figura V.3a. Cabe agregar que al comparar la distribución de las observaciones con una distribución teórica siempre se observará una variabilidad de los datos (alrededor de la línea recta). Por otra parte, y puesto que la distribución teórica (en este caso la normal estándar) involucra en realidad a una familia de curvas podrá haber algunas diferencias en la pendiente y en la ordenada al origen (respecto de "$y = x$") (Chambers et al., 1983), por lo tanto, los datos de frecuencia de la dirección N se pueden considerar como respondiendo a una distribución normal y con ausencia de atípicos.

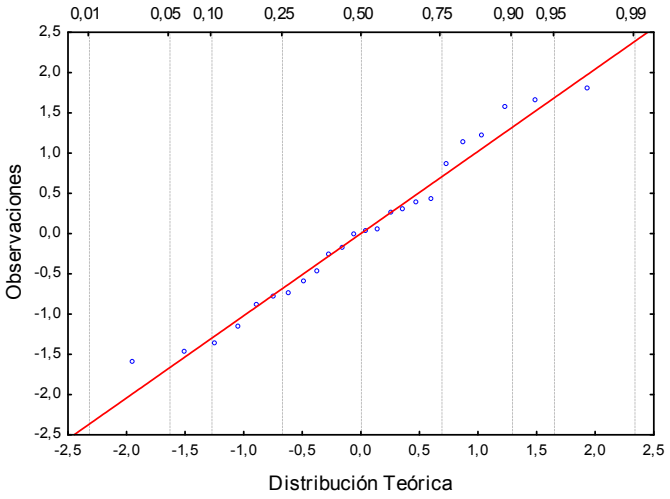


Figura V.3a: Diagrama cuantil-cuantil correspondiente a las frecuencias de ocurrencia de la dirección N (norte). Eje X inferior: valores de los percentiles de la Distribución Normal Estándar (teórica). Eje X superior: percentiles expresados como probabilidad. Eje de las Y: Valores observados (datos).

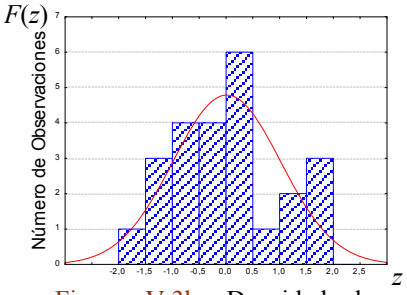


Figura V.3b: Densidad de distribución para las observaciones (barras) y para la curva teórica ajustada (forma de campana) de la Figura V.3a.

Gráficos similares al de la Figura V.3 se obtuvieron para la mayoría de las 16 variables analizadas que no se muestran por cuestiones de espacio. Los dos casos más singulares lo constituyeron la dirección ESE (este-sud-este) por alejamiento de la distribución normal (Figura V.4a) y la dirección O (oeste) (Figura V.5a) por la presencia de un potencial valor atípico.

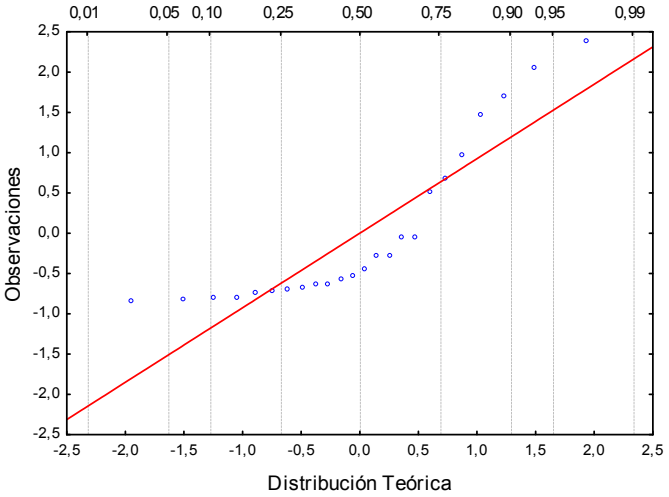


Figura V.4a: Diagrama cuantil-cuantil correspondiente a las frecuencias de ocurrencia de la dirección ESE (este-sudeste). Eje *X* inferior: valores de los percentiles de la Distribución Normal Estándar (teórica). Eje *X* superior: percentiles expresados como probabilidad. Eje de las *Y*: Valores observados (datos).

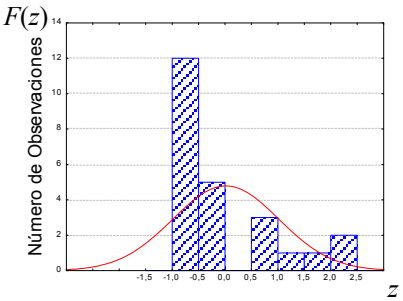


Figura V.4b: Densidad de distribución para las observaciones (barras azules) y para la curva teórica ajustada (rojo) de la Figura V.4a.

En la Figura V.4a es apreciable una cola larga a la derecha (sesgo a la derecha en Figura V.4b) indicando asimetría con valores que se dispersan desde poco antes del percentil 75 hacia los percentiles superiores. La curva presenta una forma cóncava indicando un fuerte apartamiento de la condición de normalidad. Sin embargo, separando alrededor del percentil 70 podrán apreciarse dos tramos más rectos indicativos de una potencial mezcla de dos distribuciones normales.

La Figura V.5 muestra el mismo tipo de análisis llevado a cabo para la dirección oeste. La curva se muestra moderadamente normal (no hay colas pesadas ni a la derecha ni a la izquierda), pareciera haber varios tramos como si se tratara de una mezcla de distribuciones. No hay valores atípicos a la derecha pero a la izquierda (el valor más extremo) hay un punto que merece atención dado que se halla por debajo de la recta (aunque levemente) y algo alejado del resto de los puntos. Si bien no constituye un atípico contundente es bueno tenerlo identificado en relación al método de análisis que se empleará.

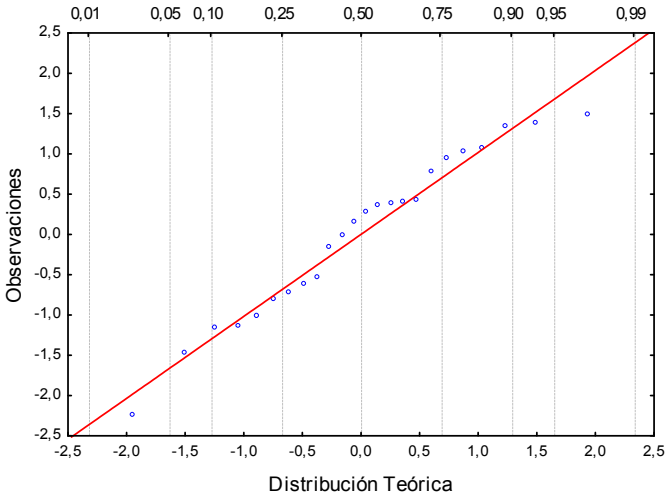


Figura V.5a: Diagrama cuantil-cuantil correspondiente a las frecuencias de ocurrencia de la dirección O (oeste). Eje *X* inferior: valores de los percentiles de la Distribución Normal Estándar (teórica). Eje *X* superior: percentiles expresados como probabilidad. Eje de las *Y*: Valores observados (datos).

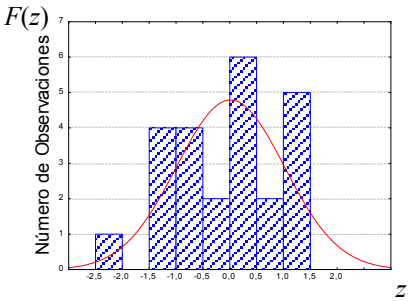


Figura V.5b: Densidad de distribución para las observaciones (barras azules) y para la curva teórica ajustada (rojo) de la Figura V.5a.

Por lo tanto, en vista de las 16 curvas exploradas es posible resumir que el sistema tiene un comportamiento bastante cercano a la normal, con muy pocas excepciones. En relación a los atípicos no hay evidencia definitiva de los mismos.

V.5.2.4.2 Cálculo de distancias a la media

Se pueden definir varios "subórdenes" (u ordenes multivariados) para un conjunto dado de datos multivariados (Barnett y Lewis, 1978). El tipo de suborden que más se ha difundido (Barnett, 2004) considera la reducción de un vector multidimensional a un escalar.
Consideremos una distancia genérica:

$$R^2(x, x_0, \Gamma) = (x - x_0)^t \Gamma^{-1} (x - x_0) \qquad \text{ec. V.1}$$

donde x_0 es alguna medida de centralidad de las variables y Γ^{-1} una matriz de peso relacionada a la dispersión de las variables.
Una manera de ordenar la muestra multivariada (suborden) es calculando $R^2(x_0,\Gamma)$ para cada objeto y ordenar los resultados de forma creciente (o decreciente) con el objetivo de detectar apartamientos o saltos sobresalientes. Dos métricas que surgen de inmediato de la ec. V.1 son la distancia Euclídea (x_0 es la media aritmética y Γ es la matriz identidad) y la distancia de Mahalanobis (x_0 es la media aritmética y Γ es la matriz de covarianzas). La simple detección de un apartamiento grande dará lugar a sospechar de la presencia de un valor atípico. La distancia Euclídea permitirá apreciar como el potencial valor atípico "infla" la escala mientras que, la distancia de Mahalanobis (que tiene en cuenta la matriz covarianza es decir, la forma de la nube multidimensional de puntos) posibilitará descubrir objetos que se hallen alejados de la nube de puntos (Barnett y Lewis, 1978). Dependiendo de las características del valor atípico (magnitud y posición relativa en la nube de puntos) podrá ser evidenciado más fácilmente con una u otra de estas distancias (Peña, 2002).

Tabla V.1

Distancia Euclídea			Distancia de Mahalanobis	
6,906	Hora 9		9,498	Hora 1
7,402	Hora 10		10,575	Hora 0
8,081	Hora 11		10,823	Hora 9
9,748	Hora 12		10,845	Hora 10
10,469	Hora 8		11,423	Hora 11
10,774	Hora 3		11,686	Hora 23
11,552	Hora 0		12,282	Hora 8
11,860	Hora 1		12,660	Hora 12
11,963	Hora 2		12,878	Hora 2
12,246	Hora 23		14,158	Hora 3
14,327	Hora 13		14,884	Hora 5
14,747	Hora 22		15,092	Hora 22
15,431	Hora 4		15,495	Hora 4
16,595	Hora 5		15,545	Hora 13
17,581	Hora 7		15,691	Hora 6
17,754	Hora 6		16,177	Hora 7
18,478	Hora 14		17,953	Hora 14
18,650	Hora 17		18,222	Hora 21
19,364	Hora 21		18,897	Hora 15
20,502	Hora 15		20,179	Hora 16
21,095	Hora 18		20,391	Hora 20
22,563	Hora 16		21,202	Hora 17
24,332	Hora 19		22,540	Hora 19
25,581	Hora 20		22,752	Hora 18

Tabla V.1: Distancias a la media; Euclídea (columna 1); Mahalanobis (columna 2).

La Tabla V.1 muestra los resultados de ambas distancias al cuadrado. Las mismas fueron ordenadas de manera creciente con el objeto de facilitar la detección de saltos abruptos. En ninguno de los dos casos es apreciable alguna singularidad. Es posible, sin recurrir a un test de hipótesis, dar un paso más (aunque no se hará por cuestiones de espacio) realizando un gráfico cuantil- cuantil para los $R^2(x_0,\Gamma)$ (ec. V.1) suponiendo algún tipo de distribución. Según Thode (2002) el cuadrado de la distancia de Mahalanobis sigue una distribución β con las posiciones (factores de forma) según Bloom. Por su parte Reinmann et al. (2008) señalan que cuando se puede suponer que las observaciones multidimensionales no atípicas siguen una distribución normal la distancia de Mahalanobis robusta (es decir, aquella que contiene estimadores robustos de posición y escala) sigue una distribución *Chi cuadrado*.

V.5.2.4.3 Componentes principales

Como se señaló anteriormente (Sección V.5), el análisis por componentes principales (CP) posibilita operar con menos variables que las originales. En la presente sección se aplica este método para explorar valores atípicos. La aplicación en muestras multivariadas se halla bien documentada con muchos ejemplos y herramientas relacionadas en Gnanadesikan y Kettenring (1972) quienes también proponen la alternativa del uso de CP no lineales y robustas.

Barnett y Lewis (1978) proponen explorar el conjunto de datos a partir de las primeras y de las últimas componentes principales dado que hay buenas razones para ello (Jolliffe 2002).

Los valores atípicos detectables a partir de las primeras componentes (usualmente las primeras dos) son aquellos que "inflan" las varianzas (en cuyo caso se podrán apreciar valores extremos en las gráficas de las variables marginales) y/o las covarianzas o correlaciones (en cuyo caso se podrán apreciar valores extremos al graficar las variables marginales de a pares). O sea, el método de CP proveerá una panorámica rápida para detectar potenciales valores atípicos.

Los valores atípicos detectables con las últimas componentes (por ejemplo, las últimas dos) son aquellos cuya presencia no se apreciaría en las gráficas de las variables originales. Como en las últimas CP las varianzas son muy pequeñas, la detección de un punto alejado del resto de puntos no "inflará" las varianzas o covarianzas. Sin embargo, si se considera el grado de correlación de la nube de puntos de las últimas CP puede detectarse algún caso que se aleje significativamente de ella. Este tipo de valor atípico se hará detectable siempre que algún extremo no se corresponda con el grado de correlación general de la nube multidimensional de puntos (puede haber atípicos sin influencia). Si se da que hay un número importante de valores atípicos en relación a la cantidad de datos (o muy pocos datos y un atípico) no será fácil distinguirlos (enmascaramiento). En este caso se puede aplicar el procedimiento de dejar afuera uno por vez y comparar todas las corridas posibles determinando si se observa mucho cambio. Si bien costoso este método es efectivo (Jolliffe, 2002).

Volviendo a los datos de trabajo de esta sección se llevó a cabo un análisis por CP utilizando el software *Statistica 8.0*. Las componentes principales se obtuvieron a partir de la matriz de covarianzas. Luego, se graficaron las dos primeras componentes (que explican más del 90% de la variación total- ver segunda columna de la Tabla V.2)). La Figura V.6 muestra el aporte individual a la varianza total de cada una de las primeras cuatro componentes principales.

Tabla V.2

Número de Autovalor	% Varianza Matriz Cov.
1	62,72
2	91,15
3	95,31
4	97,40

Tabla V.2: Varianzas (%) acumuladas para los primeros cuatro autovalores según la matriz de covarianzas del conjunto original de datos.

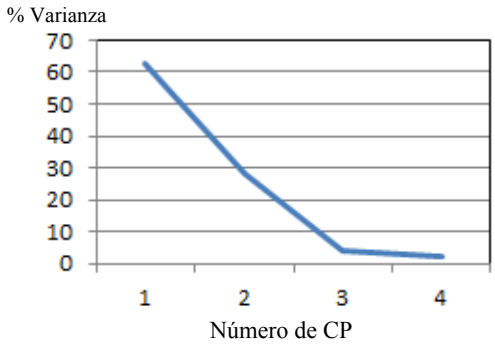


Figura V.6: Aporte a la varianza total de cada una de las primeras cuatro componentes principales.

La Figura V.7 muestra la configuración de puntos para las dos primeras CP. En ella puede notarse la ausencia de valores atípicos.

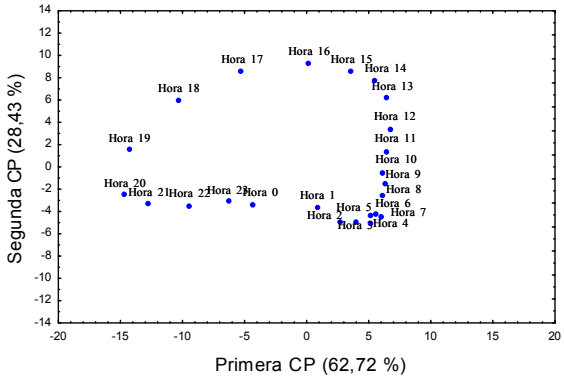


Figura V.7: Rosetas horarias expresadas en función de las dos primeras componentes principales.

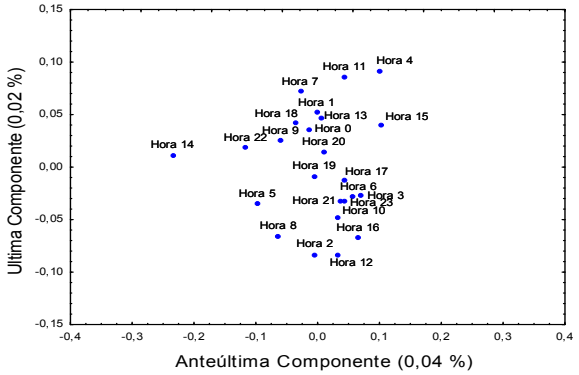


Figura V.8: Rosetas horarias expresadas en función de las dos últimas componentes principales.

La Figura V.8 muestra las dos últimas componentes principales. En ella no se observan valores que sean claramente extremos. La Hora 14 podría investigarse con otras herramientas pero a la luz de los ejemplos mostrados en la literatura no constituye un valor que verdaderamente se "despegue" de la nube de puntos (Barnett, 2004).

Varmuza y Filzmoser (2009) recomiendan el uso de PC robustas para la detección de atípicos dado que permiten detectar extremos que suelen quedar enmascarados en el caso clásico. Utilizando una opción robusta (basada en el MCD) dada en el software *Scout 1.0* (EPA, 2009) se llevó a cabo una exploración análoga a la mostrada arriba. Con esta variante (la misma no se muestra por cuestiones de espacio), tanto las primeras como las últimas componentes principales robustas no revelaron la presencia de valores atípicos.
Existen otros enfoques que involucran a todas las componentes principales simultáneamente (Jolliffe, 2002). Si bien son trabajosos y no hay firme evidencia de cual es la distribución que debería seguir la muestra (Maronna, CP) pueden resultar útiles cuando se tenga que calificar (junto a otras vías de detección de valores atípicos) a los valores atípicos como tales.
De lo analizado en esta sección, es posible concluir que a partir de las CP no se observan situaciones que hagan sospechar la existencia de valores atípicos. En concomitancia con la exploración realizada con las otras herramientas descriptas (Sección V.5.2.4.1 y Sección V.5.2.4.2) es posible concluir que para el ejemplo citado no hay valores atípicos contundentes.

Perspectiva
Es frecuente encontrar valores atípicos en grandes conjuntos de datos multivariados. El enfoque que se ha adoptado a lo largo de la tesis ha sido, en primer término, explorar los datos básicos con distintas herramientas y determinar así la importancia de los potenciales valores atípicos para cada conjunto de trabajo. Cuando sucede, como en el ejemplo analizado, que no hay evidencia contundente de la presencia de valores atípicos, se ha adoptado por elegir enfoques (métodos) que tengan un determinado grado de robustez (a través del algoritmo que emplean, criterios de aglomeración, etc.). Aquí subyace la idea de trabajar con un enfoque lo más tradicional posible para hacer que las comparaciones con otros estudios sean más directas. Sin embargo, no se debe dejar de lado la importancia de las alternativas robustas, en particular cuando estas tienen una alta eficiencia.

V.5.2.5 Estandarización

Esta transformación es relevante puesto que implica definir la importancia relativa entre las variables o los casos. De manera general, la transformación de datos por estandarización (o normalización) implica convertir a los mismos en adimensionales (Gan et al., 2007). Es necesaria la mayor parte de las veces y, debería ser tal, que ayude a revelar la estructura de los datos. Sin embargo, no hay una respuesta universal al dilema de si se debe o no estandarizar (Kaufman y Rousseeuw, 2005) o de qué manera es mejor hacerlo (Mirkin, 2011). Podría no necesitarse si las variables son lo suficientemente homogéneas (Peña, 2002) o, si por conocimiento del investigador, se sabe que hay variables que son intrínsecamente más importantes que otras para una determinada aplicación (Kaufman y Rousseeuw 2005). Desde un punto de vista práctico puede también evitarse cuando el método empleado en el análisis por conglomerados utiliza una medida de similitud o disimilitud invariante al cambio de escala (Sección V.4 y Anexo V.1, pág. 171) (Friedman y Rubin, 1967). Más allá de estos casos, distintos autores recomiendan la estandarización. Jajuga y Walesiak (2000) proveen una buena síntesis de criterios de estandarización teniendo en cuenta distintos tipos de variables.

El proceso de estandarización de los datos iniciales puede llevarse a cabo a) por variable (para hacer comparables las magnitudes de las mismas) cuando el objetivo es explorar grupos de individuos dentro del conjunto de datos o b) por casos (haciendo que los mismos sean más comparables entre sí) cuando el objetivo es ver si las variables forman grupos. La estandarización por variable es lo más frecuente (implica transformar cada variable "barriendo" todo el conjunto de objetos) y cubre el conjunto de aplicaciones de esta tesis.

Un principio que subyace en la estandarización de los datos iniciales para su posterior empleo en cualquier método de análisis multivariado es el *principio de igual importancia o de equivalencia de las variables*. El mismo establece que, cuando no se pueden asignar pesos a las variables, debe considerarse que las mismas contribuyen de igual manera al resultado (Mirkin, 2005). En consonancia, Jajuga y Walesiak (2000) hablan sobre la importancia de "ecualizar" las variables. De esta manera se pretende controlar algunos aspectos de los datos (variabilidad de las variables) para poner en evidencia otros (tales como la discriminación de grupos).

La elección del método de estandarización depende las características de los datos originales. En el caso en que todos los datos tengan variables con las mismas magnitudes, la estandarización producirá homogeneización de las variables evitando los efectos de distorsión (por el predominio de alguna de ellas) que se transmitirían al análisis por conglomerados. En el caso en que las variables posean distintas magnitudes la estandarización se hace inevitable. La elección del método de estandarización puede basarse también en el contexto de aplicación. Muchas veces es necesario realizar comparaciones con resultados previamente obtenidos que habían sido estandarizados de una determinada manera. En otros casos, cuando a los datos se le aplican otros tratamientos (o pre- procesamientos), como cuando se busca determinar si la muestra posee valores atípicos, la decisión de la adopción de un método de estandarización u otro puede cambiar. Jajuga y Walesiak (2000) dan cuenta de como quedan caracterizadas las variables cuando se estandarizan con uno u otro criterio. Por ejemplo, si se estandariza con media y desvío estándar cada variable tendrá $\bar{x} = 0$ y $s = 1$ mientras que si se estandarizan con media y rango (r) cada variable tendrá $\bar{x} = 0$, $r = 1$ y el desvío será s/r. Otro efecto a considerar, en la medida en que la estandarización implique una transformación lineal de

cada variable, es que se mantendrán el sesgo y la curtosis de la distribución de cada variable y si se toman las variables de a pares el coeficiente de correlación no cambiará.

Milligan y Cooper (1988) presentan un estudio muy importante de varios métodos de estandarización de variables numéricas en el contexto del análisis jerárquico de conglomerados utilizando distancia Euclídea. Los datos de trabajo fueron obtenidos por simulación bajo distintas configuraciones en donde siempre la cantidad de grupos era conocida. El estudio utiliza los datos sin estandarizar y siete formas de estandarización evaluando la eficiencia de recuperación (del número de grupos) para distintos efectos tales como estrategias de aglomeración, valores atípicos, etc. Los autores concluyen que es conveniente estandarizar los datos y que aquellas modalidades donde se divide por el rango son las que tienen mejor desempeño. Milligan y Cooper muestran que la estandarización tradicional es, en general, menos eficiente ante varios efectos. El trabajo también pone en evidencia la importancia de la acción conjunta entre alternativas de estandarización y las distintas estrategias utilizadas para calcular distancias entre grupos. Por ejemplo, la estandarización con media y desvío estándar junto a la estrategia de la distancia promedio (UPGMA- Sección V.4 y Anexo V.1, pág. 171) tienen buen desempeño, solamente superado por la regla de Ward. Si bien, es destacable la relevancia del estudio, los resultados no son plenamente generalizables a cualquier caso: los datos de trabajo pueden tener una estructura distinta a la estructura de los datos simulados.
Por último, y dependiendo de los objetivos del trabajo que se quiere realizar, puede ser importante tener en cuenta las convenciones adoptadas en el campo de estudio (Gan et al., 2007). A este respecto cabe destacar que Wilks (2006), en su libro dedicado al estudio de métodos estadísticos de aplicación en Ciencias Ambientales, da la estandarización típica con media y desvío como la manera convencional de transformar datos.

Una recomendación general de varios autores (Escudero, 1977; Everitt et al., 2011) es que no se debería estandarizar en base a todos los casos y las variables (o sea con gran media y desvío total). Fleiss y Zubin (1969) muestran que esto distorsiona los datos diluyendo las diferencias entre grupos.

Existen opciones de estandarización, la siguiente ecuación muestra una fórmula general de estandarización lineal por variables donde se efectúan dos operaciones, una es la de llevar al origen y la otra es la de cambiar la escala:

$$z_i = (x_i - a)/\, b \qquad \text{ec. V.2}$$

Se han propuesto maneras muy diversas de elegir *a* y *b* (Mirkin, 2005). Si ambos son cero los datos quedan sin estandarizar. *a* puede ser por ejemplo, el mínimo del intervalo de las x_i, el máximo, el rango medio ((max+min)/2), la media, la mediana. *b* puede ser por ejemplo, el rango, la mitad de rango ((max-min)/2), el desvío estándar.
La adopción simultánea de *a* como la media y de *b* como el desvío estándar, conocido como "escore z" ("marcador z"), ha sido la piedra angular de la estandarización y es todavía hoy la opción más popular (Mirkin, 2011). Esta "tradición" es una herencia de la estadística clásica en donde se asume que los datos tienen distribución normal y, por lo tanto, los datos transformados pertenecen a una distribución "libre" de parámetros. Pero, en los casos de aplicación los datos raramente responden de manera completa a una distribución normal (u otra distribución "con nombre") pudiendo a veces responder a más de una distribución conocida. Por otra parte, las variables no son necesariamente independientes.

La función de centrar o **llevar al origen** es posicionar los datos respecto de una referencia de tal manera que la medida de tendencia central sea cero.

El **escalamiento** por el desvío estándar pareciera, en principio, muy bueno puesto que satisface el principio de equivalencia de las variables. Sin embargo, esto constituye una visión muy simplificada del problema. Existen dos factores independientes que contribuyen al desvío estándar: a) la escala (o rango) y b) la forma de la distribución. El desvío estándar incluye a los dos de manera compacta y, por lo tanto, no puede discriminarlos. Si dos distribuciones tienen el mismo rango pero formas distintas, por ejemplo la primera es unimodal y la segunda multimodal (típica de estructuras donde hay grupos) el rango será el mismo pero el desvío será mayor en la segunda. Al estandarizar con el desvío se distorsionarán los datos en favor de una distribución unimodal. Esto implica que la separación entre las dos partes de la distribución multimodal se encogerá y al método que se aplique para discriminar grupos le "costará" más. Pero de tratarse de una verdadera distribución unimodal el desvío tenderá a separar los datos. Ambos efectos son no deseables puesto que desvirtúan la estructura de los datos más que ayudar a revelarlos. En este sentido la estandarización con el rango ayudará, en principio, a detectar la presencia de grupos. Este análisis tiene soporte en los trabajos de Milligan y Cooper (1986) y Steinley (2004). La desventaja de dividir por el rango es que es muy vulnerable a la presencia de valores atípicos, una opción es utilizar un rango intercuantil (por ejemplo, dejando afuera un porcentaje de los datos que se hallan en los extremos de rango). Kaufman y Rousseeuw (2005) señalan otra opción robusta pero de fácil aplicación que es utilizando el desvío absoluto medio.

En todos los casos, es importante tener en cuenta que la estandarización de los datos de trabajo guarda una estrecha relación con el método de análisis multivariado que se empleará. El centrado y el escalado de los datos deben considerarse como dos efectos por separado (Maronna, CP). En el caso de aplicar componentes principales es fundamental que los datos estén centrados puesto que de lo contrario variarán las proporciones explicadas de varianza de cada componente (Varmuza y Filzmoser, 2009). El escalado también influirá, principalmente si este se ha realizado o no. En análisis por conglomerados el centrado tiene menor importancia dado que las distancias o las correlaciones son insensibles al mismo. Para cualquier método mulitvariado que se quiera aplicar existen alternativas robustas (Maronna et al., 2006; Varmuza y Filzmoser, 2009).

V.5.3 Criterio de aglomeración

Este paso en la implementación del análisis por conglomerados jerárquicos involucra decisiones basadas en lo presentado en las secciones V.3 y V.4.

V.5.4 Procedimiento de aglomeración

En esta sección se pretende ampliar lo esbozado en la Sección V.2. Si el punto de partida del estudio es una matriz de dos vías (*n* objetos en *p* dimensiones) se debe calcular primero la matriz de disimilitud (distancias) $D_{n\,x\,n}$. Si el punto de partida es ya una matriz de una vía ($D_{n\,x\,n}$) entonces se procede a identificar la menor de las distancias entre dos objetos del conjunto (por ejemplo, *r* y *s*). Estos objetos formarán el primer grupo.
Luego (y según un criterio adoptado) se calculan las distancias entre el grupo recientemente formado y todos los individuos remanentes (o grupos) dando lugar a una nueva matriz de distancias $D_{(n-1)\,x\,(n-1)}$.

Siguiendo la lógica de los pasos anteriores se continúa hasta que todos los objetos terminan formando un solo grupo (de *n* individuos). En algún paso del proceso existirá un número de grupos que el investigador adoptará (ver Sección V.5.5) para la aplicación que lleva a cabo.

La construcción de un dendograma puede seguir determinadas pautas que están asociadas a las características de la medida y el criterio de disimilitud elegidos (relacionados íntimamente con la matriz de similitud/disimilitud que se obtiene). Una jerarquía indexada implica que un objeto no puede pertenecer a dos grupos al mismo tiempo y que cada grupo es la unión de los objetos y/o grupos que contiene. Es posible demostrar (Cuadras, 2012) que en toda jerarquía indexada se puede definir una distancia ultramétrica (Sección V.3) y que todo espacio ultramétrico define una jerarquía indexada.

Una consecuencia de no cumplir con la propiedad ultramétrica lo constituyen las reversiones que ocurren cuando los sucesivos niveles de aglomeración (o fusión) no siguen una secuencia monotónica.

Un método jerárquico en que las reversiones no pueden ocurrir se llama monotónico porque la distancia en cada paso es mayor que la distancia en el paso anterior. Si la distancia o el criterio de aglomeración son monotónicos entonces se dice que son ultramétricos (Rencher, 2002).

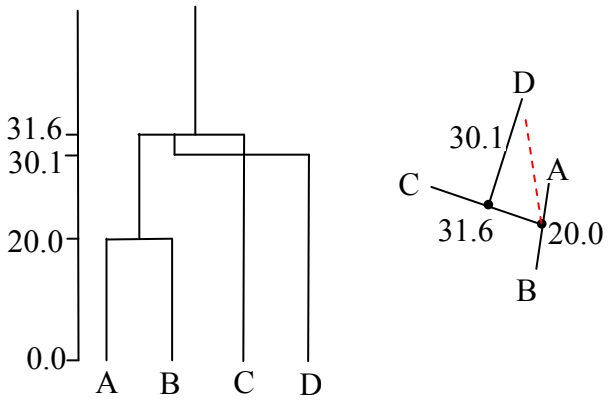


Figura V.9: La línea de rayas (roja) es la distancia del centroide de A-B hasta C trasladado para mostrar que no llega a D. Observar que no se mantiene la estructura anidada (jerarquía indexada).

En la Figura V.9 se muestra un ejemplo adaptado del Capítulo 4 de Everitt et al. (2011) para el Enlace Centroide (Sección V.4 y Anexo V.1, pág. 171). La distancia del centroide del grupo A-B es menor a C que a D, por lo tanto se forma A-B-C. Luego se calcula la distancia del centroide A-B-C a D y resulta menor (30.1) que de A-B a C del paso anterior (31.6).

Desde un punto de vista práctico también pueden darse reversiones en los casos de tricotomía (tres individuos tienen idénticas distancias de a pares). El programa fusionará dos y al fusionar el tercer punto con los dos primeros dará cuenta de una reversión (Legendre y Legendre, 1998).
Cabe agregar, que las reversiones (también llamadas inversiones) no constituyen necesariamente un problema para el investigador cuando, con fines exploratorios se busca alguna partición particular de la jerarquía (Everitt et al., 2011), sin embargo, se debe tener en cuenta que su presencia frecuente dificulta la interpretación.

V.5.5 Determinación del número óptimo de grupos

La estimación del número de óptimo de grupos es una etapa fundamental del análisis por conglomerados puesto que está muy ligado al objetivo de sintetizar información de manera representativa. Es frecuente encontrar en publicaciones de distintas disciplinas que la adopción de un determinado número de grupos se realiza en base al análisis del fenómeno en estudio. Sin embargo, se debe tener en cuenta que existen alternativas netamente estadísticas que pueden ayudar en la decisión. Desde esta perspectiva la determinación del número óptimo de grupos constituye un problema difícil de resolver (Tibshirani et al., 2001) principalmente por dos motivos.

Por una parte, como se comentó en la Sección V.1, el concepto de grupo (o "cluster") no tiene una definición lo suficientemente precisa ni generalizable (Tibshirani et al., 2001; Mirkin, 2005) y por lo tanto no se puede especificar *a priori* que es un grupo en un conjunto de datos de trabajo (Gordon 1999). Como se ha citado (Sección V.1) muchos autores toman a la cohesión interna (homogeneidad del grupo) y al aislamiento externo (separación entre grupos) como criterios "matematizables" para definir la existencia de grupos y se basan en el cálculo de la suma de cuadrados dentro del grupo (Sección V.5.5.1). Sin embargo, las distintas configuraciones que pueden adoptar los datos dan lugar a una gran variedad de definiciones de dichos criterios (Everitt et al., 2011). Es muy distinto tratar con grupos "esferoides" que con grupos "elongados" (Sección V.4 y Anexo V.1, pág. 171) y en general, el investigador desconoce que configuración pueden adoptar los grupos que intenta analizar. Bonner (1964) ha sugerido que el concepto de grupo es, en último término, aquello que le de al investigador respuesta sobre lo que está buscando o como dice Rencher (2002) le de un sentido a la investigación; en la misma dirección Baxter (1994) señala que el criterio subjetivo basado en la experticia sigue siendo el método prevalente mientras que Mirkin (2005) remarca la importancia del marco de trabajo en el que se opera. Hair et al. (2010) señala que todas las decisiones adoptadas por el investigador (sobre las características adoptadas, los criterios de aglomeración elegidos, etc.) son del mismo peso que cualquier test empírico.

Por otra parte, los métodos que se han ido diseñando durante décadas se han basado o bien en los casos particulares de análisis o bien en datos generados por simulación (por ejemplo utilizando Monte Carlo). En relación a esto último, y de manera análoga a lo que ocurre con el proceso de estandarización, los resultados no son del todo generalizables puesto que el desempeño del método que se ensaya depende tanto de la estructura de los datos como del tipo de algoritmo que estuvo involucrado en su elaboración. Es decir, distintos métodos aplicados a los mismos datos tendrán distinto grado de recuperación pero no se sabrá en que medida dependen de la estructura o del algoritmo. De todas maneras estos estudios son valiosos (Everitt et al., 2011) porque sirven para identificar métodos que muestran con sistematicidad bajo nivel de desempeño, aún con conjuntos de datos donde se conoce *a priori* cuan bien definidos están los grupos.

Dos trabajos muy importantes desde el punto de vista de la compilación y comparación de métodos para determinar el número óptimo de grupos son el de Milligan y Cooper (1985) y el de Dimitriadou et al. (2002), este último solo para datos binarios. Gordon (1999) recomienda no depender solamente de un método para determinar el número de grupos.

A continuación se presentan y discuten los principales aspectos de los métodos utilizados para determinar el número óptimo de grupos. Se trabaja con un método gráfico para detectar el fenómeno del "hombro" (basado en la estimación de W_k y en el diagrama de

sedimentos- Sección V.5.5.1). Además se calculan los índices de Calinkski y Harabasz (1974), Hartigan (1975) y Krzanowski y Lai (1988) por estar todos estos entre los índices que tienen buen desempeño en el estudio de Milligan y Cooper (1985) y por haber sido los más utilizados en los trabajos donde se proponen nuevos estimadores (Tibshirani et al., 2001; Tibshirani y Walther, 2005). El método de las Siluetas (Sección V.8.5) propuesto por Rousseeuw (1987) incluye la determinación del número óptimo de grupos.

V.5.5.1 Suma de cuadrados (W_k)

Se define W_k para indicar el grado de dispersión interna de los grupos del conjunto original de individuos (vectores p- dimensionales) que se van aglomerado en las distintas instancias para formar k grupos. W_k es la suma total de las distancias al cuadrado entre cada miembro de un grupo y su centroide para todos los grupos cuando se han formado k grupos. A medida que k aumenta según los pasos de aglomeración W_k (también llamada suma combinada de cuadrados dentro de los grupos) tiende a decrecer monótonamente hasta que para algún k el decrecimiento se aplana notablemente. La figura resultante, conocida como gráfico de sedimentación (en inglés como "scree plot"), constituye una manera sencilla de visualizar la estructura de grupos de los datos. En la Figura V.10a se ha representado un conjunto de puntos en el plano que están distribuidos de forma bastante homogénea (no mostrando evidencia de la presencia de subgrupos). En la Figura V.10b se puede apreciar la evolución del W_k obtenido para los diez primeros grupos: un decrecimiento gradual a medida que k aumenta.

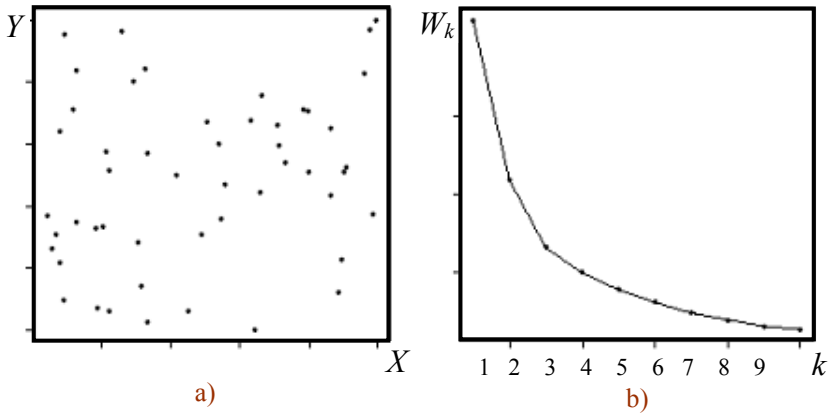


Figura V.10: ejemplo tomado de Suggar et al. (1999).
a) Datos al azar en el plano
b) Curva del W_k en función del número de grupos (gráfico de sedimentación).

En la Figura V.11a se observa un conjunto de datos en donde se evidencian marcadamente dos subgrupos. El gráfico de sedimentación correspondiente (curva de la derecha) acusa un decrecimiento abrupto ("hombro") al pasar de k= 1 a k= 2 y luego se aplana suavemente a medida que k crece.

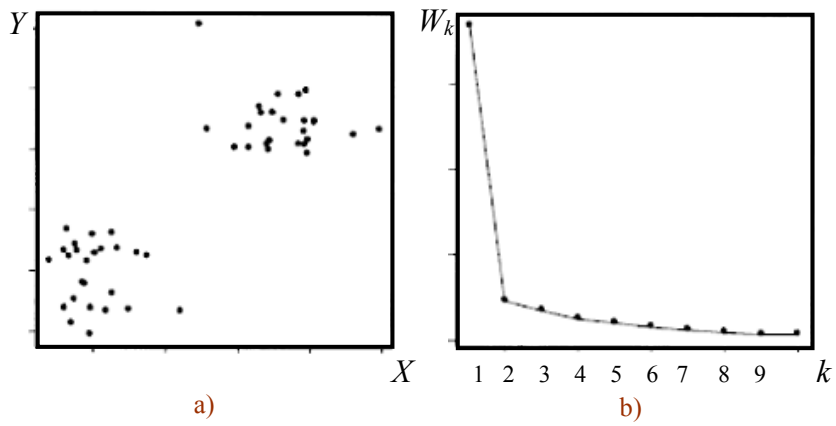


Figura V.11: Ejemplo tomado de Tibshirani et al. (2001)
a) Datos con estructura de grupo en el plano
b) Curva del W_k en función del número de grupos.

La principal desventaja de este enfoque (basado en la suma de cuadrados) es que pueden presentarse varios "hombros" y se hace difícil distinguir el más significativo conduciendo a una elección subjetiva por parte del analista.

V.5.5.2 Indice de Calinski y Harabasz (*CH(k)*)

El índice de Calinski y Harabasz se define como:

$$CH_{(k)} = \frac{B_{(k)}/(k-1)}{W_{(k)}/(n-k)}$$

donde $B_{(k)}$ (suma de cuadrados entre grupos) indica el grado de dispersión que existe entre los grupos que se van formando en el proceso de aglomeración para formar *k*- grupos. $B_{(k)}$ es la suma total de las distancias al cuadrado entre el centroide de un grupo y el centroide de los datos originales (centroide general). $W_{(k)}$ es el W_k cuya notación se ha cambiado para homogeneizar nomenclatura con los autores e indica el grado de dispersión que existe dentro de cada grupo sumado para todos los grupos. *n* es el número total de individuos (vectores *p*- dimensionales). $B_{(k)}$ se halla dividido por el número de grupos menos 1 para "escalar" su valor; algo análogo sucede con $W_{(k)}$ que es escalado con el número de datos menos el de grupos formados.

Por lo tanto, este índice expresa la relación entre la dispersión entre grupos (pesada por el número de grupos) y la dispersión dentro de los grupos (pesada por el número de individuos menos el de grupos). Se espera que *CH(k)* sea máximo, o sea que se maximice *B(k)* y se minimice ($W_{(k)}$), de esta manera se cuantifican los grados de cohesión interna y aislamiento. *CH(k)* no está definido para *k*=1 (lo está para *k*>1) lo cual implica que el método no especificará si el conjunto de individuos originales forman un solo grupo (de estar definido debería presentar el máximo para *k*=1). Cuando existe una fuerte estructura de grupo *CH(k)* da un máximo único pero cuando esto no es así, (la presencia de máximos locales indican que la estructura de grupo de los datos es moderada), los autores (Calinski y Harabasz, 1974) recomiendan tomar como número óptimo el primero de los máximos, o sea, el de menor *k*. Cuando los valores de *CH(k)* crecen monótonamente con *k* implicaría que no es posible una partición razonable de los individuos originales (o sea, que no hay estructura de grupo presente en los datos).

V.5.5.3 Indice de Hartigan (*H(k)*)

El índice que propone Hartigan (1975) se define como:

$$H(k) = \left[\frac{W(k)}{W(k+1)} - 1\right](n-k-1)$$

donde $W_{(k)}$, *n* y *k* conservan las definiciones dadas en las secciones anteriores. *H(k)* se estima para conocer cuando es justificable particionar los datos en un grupo más, a diferencia del método anterior está definido para *k*=1. Un valor alto de *H(k)* indica que la adición de un nuevo grupo es viable. Hartigan (1975) sugiere (Capítulo 4), como norma práctica, que un valor superior a 10 justifica incrementar el número de grupos pasando de *k* a *k+1* para seguir investigando. Por lo tanto, se buscará el número óptimo de grupos con el menor *k* posible tal que *H(k)* sea menor o igual a 10.

V.5.5.4 Indice de Krzanowski y Lai ($KL_{(k)}$)

El índice que proponen Krzanowski y Lai (1988) se define como:

$$KL_{(k)} = \left| \frac{DIFF(k)}{DIFF(k+1)} \right|$$

con

$$DIFF(k) = (k-1)^{2/p}\ W_{(k-1)} - \ k^{2/p}\ W_{(k)}$$

donde p designa el número de variables involucradas en los datos.

Se elige el k que maximice $KL_{(k)}$ que será el $k_{óptimo}$. Al igual que el índice propuesto por Calinski y Harabasz $KL(k)$ no se halla definido para k=1.

Cuando $k< k_{óptimo}$ $DIFF(k)$ y $DIFF(k+1)$ tienden a ser grandes ($W_{(k)}$ y $W_{(k-1)}$ tienden a ser grandes cuando los k son bajos: sus diferencias se hacen notorias) y por lo tanto el cociente se hace pequeño. Cuando $k> k_{óptimo}$ $DIFF(k)$ y $DIFF(k+1)$ tienden a ser pequeñas y los cocientes vuelven a hacerse pequeños. Para $k= k_{óptimo}$ nos encontramos en la inflexión en que $DIFF(k)$ es grande mientras que $DIFF(k+1)$ tiende a ser pequeño produciendo un máximo en el $KL_{(k)}$ o sea el $KL(k_{óptimo})$.

Como puede apreciarse los tres índices descriptos están basados en la minimización o maximización de una función objetivo basada en la suma de cuadrados.

V.5.5.5 Ejemplos de determinación del número óptimo de grupos

Tomando como material de trabajo las rosetas de vientos observadas durante el Verano en el Punto J publicado en Ratto et al. (2010b) se obtiene el dendograma de la Figura V.12

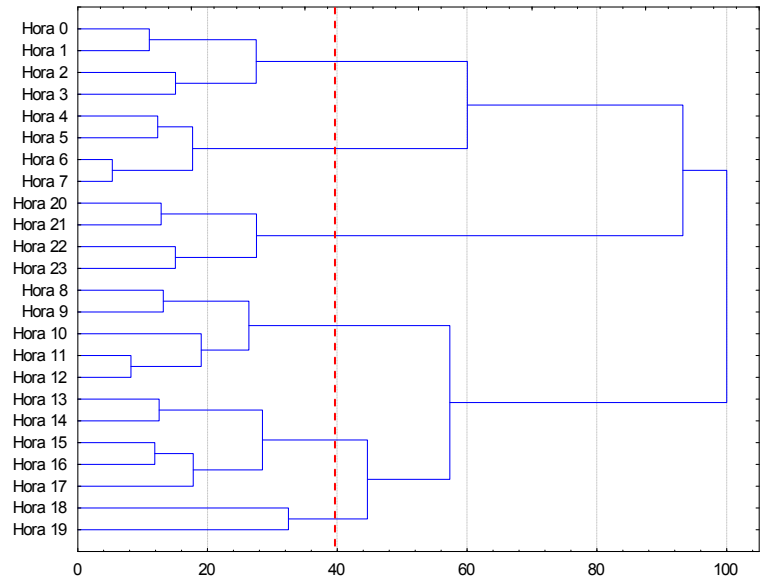


Figura V.12

Figura V.12: Dendograma de 24 rosetas horarias promedio de vientos correspondiente al verano en el Punto J para el período 1998- 2003.
En el eje de las X se halla representada la distancia Euclídea al cuadrado reescalada en % (para facilitar comparaciones con otros dendogramas).
En el eje de las Y cada "Hora" representa un vector de 16 direcciones de frecuencia de vientos. La línea de trazos vertical cercana a una distancia de corte del 40% indica la solución dada por la mayoría de los criterios aplicados para la determinación del número óptimo de grupos.

El correspondiente diagrama de sedimentación se muestra en la Figura V.13.

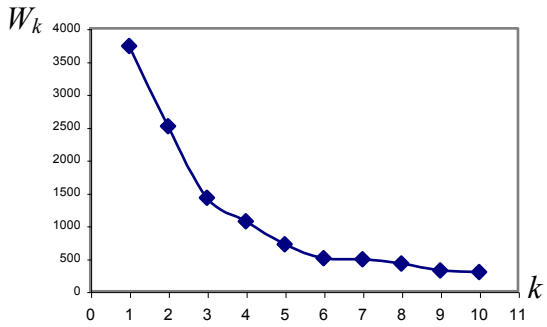


Figura V.13: Diagrama de sedimentación para el dendograma de la Figura V.12

La curva muestra una pendiente abrupta al principio indicando la separación entre grupos para valores bajos de *k* entre 1 y 3. Luego se suaviza para valores de *k* entre 3 y 5. Para *k*= 6 podría considerarse que se aplana indicando que más subdivisiones no reflejarían la realidad del conjunto.
Por lo tanto, analizando esta curva se puede concluir que 6 sería un número óptimo de grupos en que se pueden configurar los 24 datos iniciales.

En la Tabla V.3 se presentan los valores de los distintos índices para el dendograma de la Figura V.12.

Tabla V.3

k	$CH_{(k)}$	$H_{(k)}$	$KL_{(k)}$
1	--	10,7	--
2	0,9	16,1	0,9
3	2,9	**6,7**	3,0
4	3,5	9,4	0,9
5	4,6	7,6	1,6
6	**5,9**	0,5	**38,2**
7	5,5	2,5	0,1
8	5,7	4,9	0,6
9	6,9	1,4	4,0

Tabla V.3: Indices de Calinski y Harabasz ($CH_{(k)}$), Hartigan ($H_{(k)}$) y C y Lai ($KL_{(k)}$) para el dendograma de la Figura V.12.

CH(k) presenta valores crecientes hasta llegar a un máximo para *k*= 6 luego decrece y vuelve a crecer. En primera instancia el índice indica que el número óptimo de grupos es 6. Si para *k* mayores a 9 los valores siguieran incrementándose monótonamente esto sería un indicador general de que los datos iniciales no tienen una estructura fuerte de grupo (Calinski y Harabasz, 1974). En cambio si la serie fuera oscilante indicaría una estructura de grupos algo más fuerte.

H(k) presenta para *k*=1 un valor superior a diez indicando, en primer término, que el conjunto de datos originales tiene estructura como para ser particionado (al menos en dos). Luego el valor de *k* más bajo tal que *H(k)* < 10 se da para *k*= 3 indicando que para este estimador el número óptimo de grupos es tres. Notar que para *k*= 6 presenta otro mínimo local. *KL(k)* presenta claramente el máximo para *k*= 6.
De los cuatro abordajes ensayados para el dendograma de la Figura V.12 se puede concluir, en primer término, que los datos originales (individuos *p*-dimensionales) tienen una estructura que tiende a formar subgrupos. Solo el índice de Hartigan indica la presencia de tres grupos predominantes (observables en el dendograma para una amplia gama de distancias de corte- entre 60 y 90%) mientras que los otros métodos evidencian la presencia de seis grupos. Según Gordon (1999) la decisión del analista deberá basarse en lo que indican la mayoría de los métodos. Esto significa que el conjunto de 24 rosetas horarias de frecuencias de viento (obtenidas a partir de promedios acumulados) quedarán bien representadas por solo 6 rosetas de viento promedio (una por cada grupo indicado en la Figura V.12 para una distancia de corte de alrededor del 40%). En este caso, dado que las rosetas de viento que forman cada uno de los seis grupos son consecutivas, la solución dada por los indicadores tiene interpretabilidad; de no ser así se puede recurrir a adoptar otro número de grupos (Ratto et al., 2010b) o proceder a aplicar algún tipo de restricción (Sección V.8.4).

Métodos más recientes que los mostrados en este ejemplo, basados también en W_k, han sido propuestos (Suggar et al., 1999; Tibshirani et al., 2001) pero por ser más difíciles de interpretar no se ha difundido su aplicación. Enfoques no basados en el W_k, tales como los propuestos por Dudoit y Fridlyand (2002) o Tibshirani y Walther (2005) se prestan a una mejor interpretación pero resultan difíciles de abordar por su tratamiento matemático.

V.5.6. Validación

Tanto los métodos jerárquicos como los de partición conducen siempre a la obtención de grupos pudiendo imponer estructuras "no garantizadas" (Gordon, 1999). Por este motivo, cabe preguntarse si los grupos hallados, al aplicar un determinado método, son "reales" o si son simplemente "artefactos" generados por el método (Legendre y Legendre, 1998; Ritter, 2014) llamados también "artefactos estadísticos".

En el contexto general del análisis exploratorio y del concepto de métodos no supervisados, no se requiere de una validación como requisito excluyente, puesto que no hay "clases" predefinidas y puede no haber ejemplos sobre las relaciones que deberían tener los datos entre sí (Haldiki, 2002a,b). O sea, la "efectividad" de un método es cuestión opinable y no tiene verificación directa (Hastie et al., 2011). Sin embargo, es posible contar con alguna medida que provea de confianza sobre el grado en que el método aplicado resume información (Gordon, 1999), es decir, una medida que permita evaluar los resultados (Legendre y Legendre, 1998; Haldiki et al. 2001).

Si bien, la validación de estructuras de grupo resulta ser "la tarea más difícil y frustrante" del análisis de conglomerados (Jain y Dubes, 1988) o "frecuentemente frustrante" (Ritter, 2015), existe un consenso general de los autores provenientes de distintas disciplinas de que es conveniente realizar algún tipo de validación. La misma puede llevarse a cabo con distintos niveles de formalidad, los cuales guardan relación con la naturaleza de lo que se investiga, el objetivo de la investigación y las características de los datos (Gordon, 1999). Cabe agregar (Theodoridis y Koutroumbas, 2003) que la aplicación de métodos de validación debe considerarse "solo como una herramienta" a disposición del investigador.

Previo a la determinación de los grupos por la aplicación de algún método (jerárquico o de partición), el investigador podría comenzar por realizar una indagación sobre la ausencia/presencia de estructura de grupo en los datos disponibles. La existencia de grupos será detectada cuando se observe algo "inusual" en los datos (Jain y Dubes, 1988), o sea, cuando los datos no estén distribuidos completamente al azar (no poseen aleatoriedad total). Por ejemplo, podría aplicarse un test cuya H_0 (hipótesis nula) fuera la ausencia de estructura y en caso de ser rechazada proceder con un método para encontrar grupos. La búsqueda de estructura de grupos sin pretender identificarlos explícitamente es conocida como "análisis de tendencia" (Theodoridis y Koutroumbas, 2003) pero es raramente aplicada (Gordon, 1999) porque el investigador puede tener razones fundadas acerca de que los datos formen grupos, o detectando que no hay *a priori* una clara discriminación entre grupos puede simplemente estar buscando una disección de los datos para continuar con el análisis (Jolliffe, 2002) y/o considerar a los test irrelevantes, o como citan Everitt et al. (2011) puede no ser práctico.

Una manera directa y menos formal de identificar grupos en datos p-dimensionales (p>3) es aplicar algún método para reducir dimensionalidad tal como el de Componentes Principales (CP) (por ejemplo, Figura V.24 -Sección V.8.1) o Escalamiento Multidimensional (EMD) (por ejemplo, Figura V.30 -Sección V.8.3). La inspección visual

de puntos en el plano permitirá hacer una primera aproximación sobre el número posible de grupos. Estos métodos no son concluyentes; puede ocurrir que las primeras CP no expliquen un porcentaje suficientemente alto de la varianza, volviéndose dificultosa la interpretación; en el caso del EMD el factor de STRESS (indicador de la bondad del proceso de reducción de dimensionalidad) puede dar alto (ver Sección V.6). Sin embargo, cuando mediante estas herramientas se visualiza estructura es porque los datos la tienen.

Para llevar a cabo la validación existen tres categorías "típicas" de criterios de validación que pueden adoptarse: externo, interno y relativo. A continuación se describirán brevemente estos criterios para el caso de métodos jerárquicos con variables continuas.

V.5.6.1 Criterio externo

Los criterios externos se enfocan en evaluar cuan bien ajusta la estructura jerárquica hallada (con un método dado) a una estructura esperada. Esto puede implicar desde comparar directamente dos estructuras jerárquicas mediante un test (que defina si hay relación entre ambas) hasta comparar la estructura jerárquica dada por un método con una idea sobre como tienen que estar agrupados los datos (enfoque menos formal). Una ampliación de este tema se halla en Jain y Dubes (1988) y en el Capítulo 17 de Gan et al. (2007). Pero cabe agregar que esta categoría no ha recibido mucha atención puesto que, en la mayoría de los casos, no se cuenta con una jerarquía esperada, y de contarse con ella implica (en versión formal) la implementación de cálculos complejos (Xu y Wensch, 2009).

V.5.6.2 Criterio interno

Los criterios internos evalúan el grado de ajuste entre los datos iniciales y la estructura encontrada (a partir de dichos datos) como consecuencia de la aplicación de un método. Es decir, se considera solo la información del conjunto de datos que se analiza y se excluye información externa (Pande et al., 2012).

En resumen, mientras que el criterio externo compara la estructura hallada, con una estructura conocida *a priori* o que se adopta como referencia, el criterio interno se ocupa de determinar si la estructura encontrada es intrínsecamente apropiada (Jain et al., 2000).

Así como los criterios externos, los internos (Haldiki et al., 2002a) se utilizan con mayor o menor grado de formalidad. El empleo de test estadísticos implica procedimientos complejos puesto que suele requerirse de simulaciones (por ejemplo, por Monte Carlo). Detalles de este enfoque se dan en el Capítulo 4 de Jain y Dubes (1988) pero todavía estos métodos no cuentan con tests potentes y específicamente diseñados para evaluar cohesión y aislamiento (Ritter, 2015).

Los coeficientes (o índices) más aplicados para validar jerarquías con criterio interno están basados en la correlación: el coeficiente cofenético basado en el ρ de Pearson o los basados en el de Spearman (Sr) (Anexo V.3, pág. 180) o de Kendal (τ).
La distancia (o similitud) cofenética entre dos objetos se define como la distancia (o similitud) en el nivel de aglomeración del dendograma en el que ambos objetos se convierten en miembros de un mismo grupo (Legendre y Legendre, 1998), o sea la distancia "vía el dendograma" (Sección V.2). La matriz de distancias (o similitudes) que reúne a todas las distancias entre objetos (o grupos) vía el dendograma se denomina matriz cofenética de distancias (Sokal y Rohlf, 1962). Por lo tanto, el coeficiente cofenético relaciona la matriz original de distancias entre pares de objetos con la matriz cofenética de

distancias; da una idea del grado en que la "salida" del proceso de aglomeración representa a la matriz original de distancias entre pares de individuos (Romesburg, 2004).

Una ampliación sobre el concepto de coeficiente cofenético se da en el Anexo V.3 (pág. 180), en donde se provee un ejemplo de cálculo utilizando distintos criterios de aglomeración y se muestra el esquema de aglomeración. Cabe agregar aquí, que este coeficiente depende de varias de las características del problema de estudio, tales como el tamaño de la muestra -dimensión de la matriz de proximidades- (Farris, 1969), la medida de similitud/disimilitud adoptada y el criterio de agrupamiento (Theodoridis y Koutroumbas, 2003).

Ya se ha discutido (Sección V.5.3) que una matriz cofenética ultramétrica implica agrupamientos anidados (monotonicidad) y que a esta condición la cumplen solo algunos de los criterios de aglomeración. Desde un punto de vista general Seber (1984) indica que valores de ρ (coeficiente cofenético evaluado con el ρ de Pearson) menores a 0,6- 0,7 no serían aceptables. Para Chagoyen et al. (2006) este límite se halla entre 0,4 y 0,5 pero Macedo et al. (2006) sugieren que serían recién aceptables valores mayores a 0,8. Estas últimas dos citas están dentro del campo de las ciencias biológicas, en otros campos (Takahashi et al., 2007) se aceptan valores algo inferiores a 0,8 como buenos. Para el material de estudio de la presente tesis (rosetas de direcciones de vientos) no se han encontrado referencias.

V.5.6.3 Criterio relativo

Los criterios relativos evalúan como dos estructuras (cada una de ellas obtenidas con métodos distintos) se ajustan a los datos (Jain y Dubes, 1988) y permiten definir la que se ajusta mejor. Este criterio implica también la comparación de estructuras a partir de comparar los resultados que dan los distintos índices (Haldiki et al., 2002b).
La determinación del número óptimo de grupos para estructuras jerárquicas (Sección V. 5.5) utilizando distintos índices puede ser vista como un caso de validación con criterio relativo (por ejemplo empleando el coeficiente de Calinski y Harabasz), donde para cada nivel de aglomeración el coeficiente brinda una idea del grado de homogeneidad dentro de los grupos así como del grado de aislamiento entre ellos.
Existen otros índices de validación relativa que son al mismo tiempo indicadores potenciales del número óptimo de grupos (Haldiki et al., 2002b). El *RMSSTD* ("root mean squared standard deviation") es la raíz cuadrada del promedio de los cuadrados de los desvíos estándar de cada variable, el *SPR* ("semi partial R- squared") es una suma de diferencias de cuadrados relativa al nivel final de aglomeración, el *RS* ("R- squared") es una relación de suma de cuadrados y la *CD* es la distancia entre grupos. El término *CD* refiere a "centroid distance" pero como explica Sharma (1996) en el Capítulo 7, la distancia entre grupos a considerar es aquella que se haya empleado en el proceso de aglomeración. El término "R- squared" se utiliza por la analogía que tienen el *SPR* y el *RS* con el coeficiente de determinación de una regresión.
El empleo de estos cuatro índices de forma simultánea se debe a su carácter complementario; se calculan para cada nivel de aglomeración y en conjunto dan un panorama de la cantidad posible de grupos, validando determinados puntos de corte del dendograma.

RMSSTD

El *RMSSTD* se calcula en cada nivel para el grupo nuevo que se forma en dicho nivel. Este índice da una idea de la homogeneidad de cada grupo que se forma. Cuanto menor es el

incremento del *RMSSTD* al pasar de un nivel de aglomeración a otro mayor es la homogeneidad entre los grupos que se fusionan. Puesto que todos los niveles de aglomeración adoptarán un valor de *RMSSTD* será posible apreciar niveles de aglomeración con homogeneidad similar separadas de otras por saltos. Esto dará cuenta de la presencia de grupos. Ejemplos de aplicación de este índice se halla en Khattree y Naik (2000).

SPR y RS

En cualquier conjunto de datos dados por una matriz de dos modos (n_{objetos} x $p_{\text{variables}}$) es posible calcular para todo el conjunto de objetos la suma total de cuadrados (Peña, 2002).

$$S_T = \sum_{m=1}^{g} \sum_{l=1}^{n} (x_{ml} - \bar{x})^2 ,$$

donde,

x_{ml} es el vector p- dimensional del objeto l= en el grupo m

$l=1,n$ donde n es el número total de objetos

$\bar{x}$ es el vector medio de toda la muestra para cada variable

$m=1,g$ donde g es el número total de grupos

O sea, se suman los cuadrados de cada vector menos el vector medio para cada grupo y luego se suman todos todos los grupos. El resultado da una idea de la dispersión total de los datos. Esta operativa puede realizarse para cada etapa o nivel de aglomeración tal como ocurre en un proceso aglomerativo jerárquico. Será posible distinguir una dispersión dentro de un grupo definido S_W y una dispersión entre los grupos formados S_B (Everitt et al., 2011). Luego, $S_T = S_W + S_B$

El *SPR* mide la diferencia relativa de homogeneidad en cada nivel de aglomeración. Este índice da una idea de la pérdida de homogeneidad que se produce cuando se fusionan grupos en cada nivel de aglomeración. Se parte de calcular la diferencia entre el *Sw* del grupo formado en el nivel actual de aglomeración y el S_W de niveles anteriores inmediatos que componen el nivel actual. Luego, se divide por la S_T (cuando todos los individuos forman un solo grupo), o sea, se calcula $SPR = \frac{S_W}{S_T}$. En el último paso S_T coincide con S_W (puesto que S_B=0).

El *RS* expresa el grado en que los grupos son distintos entre sí (cuan aislados se hallan unos de otros) en relación a la dispersión total.

$$R_S = \frac{S_B}{S_T} = \frac{S_T - S_W}{S_T}$$

El *SPR* y el *RS* varían entre 0 y 1. El *SPR* crece a medida que se incrementan los niveles de aglomeración dando una idea de la pérdida de homogeneidad. El *RS* disminuye a medida que se avanza en los pasos de aglomeración, puesto que se pierde discriminación a medida que todos los individuos tienden a formar un solo grupo (S_B=0).

CD

Finalmente, la Distancia entre grupos *CD* puede mostrar saltos que indiquen la presencia de subgrupos en el conjunto original. Cuando se producen saltos bruscos es indicativo de que existen grupos discriminables de otros, esta característica es común a los cuatro índices.

La Figura V.14 muestra el dendograma correspondiente a las rosetas horarias de ocurrencias de viento observadas en invierno en el Punto J. El eje de las *X* no se halla reescalado con la finalidad de guardar correspondencia con las distancias mostradas en el esquema de aglomeración.

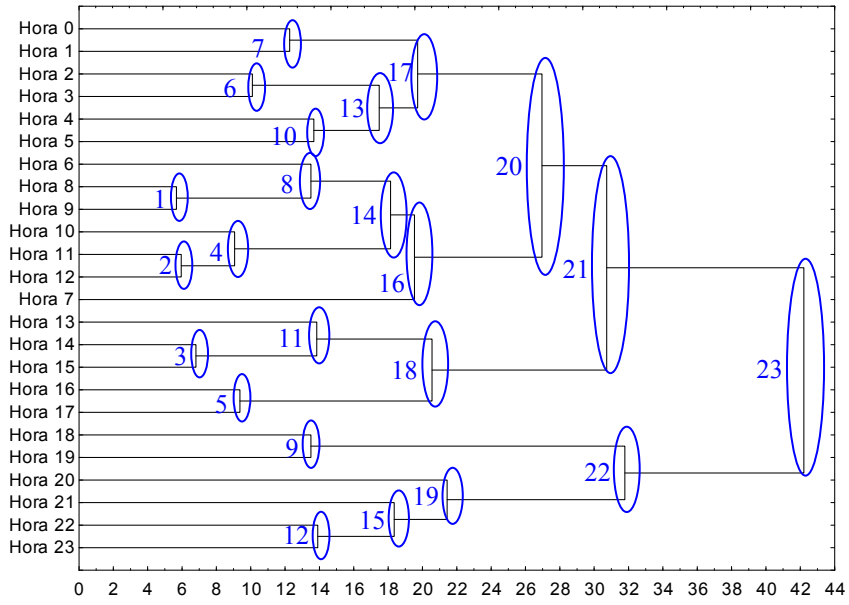


Figura V.14: Dendograma de 24 rosetas horarias promedio de vientos correspondiente al invierno en el Punto J para el período 1998- 2003.
En el eje de las *Y* cada "Hora" representa un vector de 16 direcciones de frecuencia de vientos.
En el eje de las *X* se halla representada la distancia Euclídea al cuadrado.
Los óvalos y sus números indican los sucesivos pasos de aglomeración.

Figura V.14

La Figura V.15 muestra los cuatro índices según los pasos de aglomeración del dendograma de la Figura V.14.

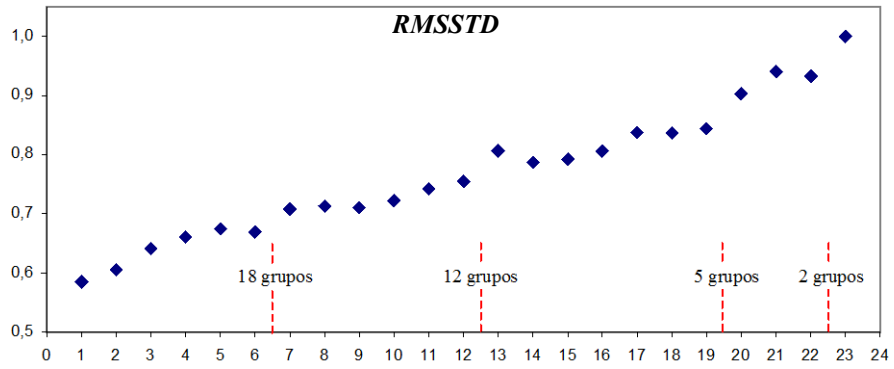


Figura V.15a: el eje de las *Y* es el *RMSSTD* y el eje de las X son los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14.

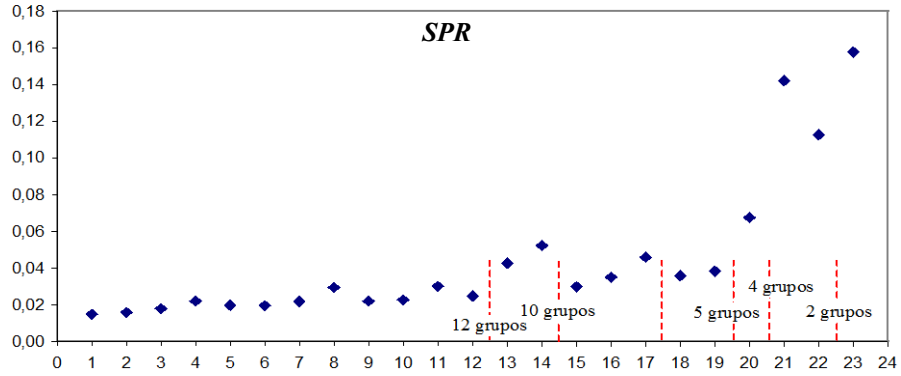


Figura V.15b: el eje de las *Y* es el *SPR* y el eje de las *X* son los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14.

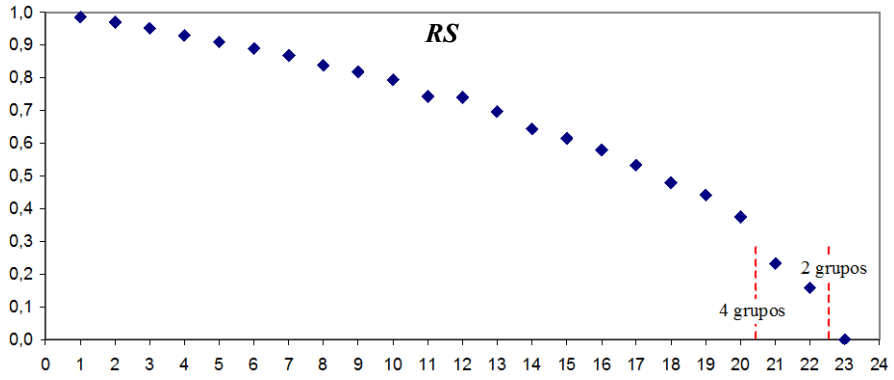


Figura V.15c: el eje de las *Y* es el *RS* y el eje de las *X* son los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14.

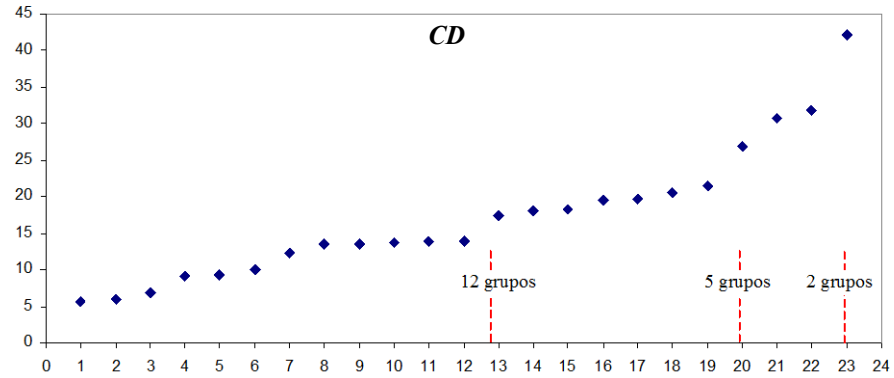


Figura V.15d: el eje de las *Y* están representadas las *CD* y en el eje de las *X* los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14.

En la Figura V.15a pueden apreciarse algunas discontinuidades a medida que el *RMSSTD* crece con el nivel de aglomeración; principalmente entre los niveles 6 y 7, entre 12 y 13, entre 19 y 20 y entre 22 y 23. Las proyecciones de las líneas de punto verticales en cada figura ayudan a distinguir posibles grupos en los datos.

La Figura V.15b muestra valores bajos de *SPR* hasta los niveles 8 u 11. Podría considerarse un primer salto brusco entre los niveles 12 y el 13 y luego entre los niveles 14 y 15 y entre 17 y 18. A partir del nivel 19 de aglomeración todos los saltos son grandes. Este resultado da idea de la existencia de pocos grupos.

La Figura V.15c muestra el *RS*. Este índice marca que la estructura de los datos bajo análisis no es fuerte. Da cuenta de la presencia de grupos entre los niveles 20 y 21 y entre el 22 y el 23 y, al igual que el anterior indica pocos grupos.

La Figura V.15d muestra la *Distancia* entre grupos (distancia Euclídea al cuadrado) siguiendo el esquema de aglomeración (criterio del enlace promedio). Esta distancia permite observar la presencia de grupos desde niveles tempranos de aglomeración siendo los saltos más importantes para 5 y 2 grupos.

Resumiendo, todos los índices distinguen grupos. *RMSSTD* y *Distancia* distinguen grupos en niveles tempranos de aglomeración, *SPR* tiende a distinguir pocos grupos mientras que *RS* muestra una estructura débil de grupos. Dado que se conocen los fenómenos físicos

asociados a la existencia de grupos de rosetas de viento (ciclo diario de la capa límite) se esperan pocos grupos. Puesto que estos indicadores operan en conjunto y que desde el punto de vista de los fenómenos meteorológicos asociados se esperan pocos grupos, es posible concluir que los datos presentan alrededor de cuatro o cinco grupos distinguibles.
Cabe agregar que existen métodos específicos cuando se requiere la validación particular de uno de los grupos obtenidos (Theodoridis y Koutroumbas, 2003). Por otra parte, existen en la literatura otros abordajes en relación a la confianza de los resultados (Gordon, 1999; Jolliffe et al., 1986). Por ejemplo, se modifica el conjunto de datos originales quitando de a un dato ("leave one out") y llevando a cabo el proceso de aglomeración. Este proceso se repite hasta agotar las posibilidades y se comparan la similitud entre la estructura que dio el resultado inicial y todas aquellas obtenidas en los subconjuntos. Sin embargo, la aplicación de este tipo de métodos es infrecuente.
En la etapa de validación, al igual que en otras etapas de implementación del análisis por conglomerados, no puede hablarse de criterios y/o coeficientes mejores que otros para todos los casos que se planteen (Xu y Wunsch, 2009).

V.5.7 Interpretación

Además de lo discutido en la Sección V.1, es pertinente resumir lo comentado por Romesburg (2004) sobre los diferentes usos que se le da a los métodos de clasificación en las distintas disciplinas: posibilitan sintetizar información favoreciendo el análisis del tema que se estudia (Anderberg, 1973) y sus objetivos; ayudan a organizar grandes cantidades de datos y poner en evidencia información importante, hacen posible encontrar objetos o variables destacables, permiten realizar generalizaciones sobre el tema abordado y mejorar la planificación. En la Sección V.8 se hace posible apreciar los beneficios resultantes de la aplicación del análisis por conglomerados en los distintos casos de aplicación.

V.6 Análisis por escalamiento multidimensional

Con el nombre de escalamiento multidimensional (EMD) –en inglés Multidimensional Scaling (MDS)- se conocen un conjunto de métodos de análisis multivariado que frecuentemente se utilizan con fines exploratorios (Timm, 2002), aunque algunos de ellos puedan ser empleados con fines inferenciales (Borg et al., 2013).
Dado un conjunto de objetos en un espacio altamente dimensional, de los cuales se conocen las proximidades entre ellos (similitudes o disimilitudes), el objetivo del EMD es representar dichas proximidades mediante distancias en un espacio de pocas dimensiones (típicamente el plano) dando lugar a una configuración de puntos. Este método permite reducir la dimensionalidad de los datos originales y poner así en evidencia las relaciones subyacentes entre las observaciones (Rencher, 2002). Cada punto en el hiperespacio tiene su correspondiente punto en el plano. Los puntos en el plano forman un arreglo tal que sus distancias se corresponden lo mejor posible con las proximidades de los objetos originales (Everitt et al., 2011); grandes disimilitudes en el hiperespacio estarán representadas por grandes distancias en el plano y viceversa. O sea, la configuración (plano) pondrá en evidencia la estructura que se halla oculta en los datos originales (Kruskal y Wish, 1978). El método puede emplearse también cuando no se conocen los datos originales sino la relación entre ellos (por ejemplo, las similitudes entre pares de objetos surgidas de una encuesta); en este caso será posible visualizar en un espacio de pocas dimensiones un conjunto de puntos que guardan la misma relación que los datos desconocidos.
El EMD nace en el campo de la psicología (Young, 1987), entre los primeros antecedentes pueden mencionarse el trabajo de M. W. Richardson en 1938 y el de G. Young y A. S. Householder de 1941. W. S. Torgerson introduce por primera vez el término en 1958 (Jain y Dubes, 1988). Por su parte, R. Shepard en 1962 demostró empíricamente que,

conociendo una ordenación de distancias en el hiperespacio, es posible hallar una configuración en espacios de bajas dimensiones que mantengan dicha ordenación (Linares, 2001). Luego Kruskal refinó el método de Shepard (Kruskal, 1964a,b) e introdujo el índice de "STRESS" como criterio de bondad de ajuste entre los puntos en la configuración y los datos originales utilizando un método de regresión isotónica (Gordon, 1999). Carroll y Arabie (1980) presentan un estudio panorámico mientras que muchas variantes de los métodos de EMD y aplicaciones recientes pueden encontrarse en la obra de Borg et al. (2013).

Es común hallar en los textos de análisis multivariado o en la bibliografía específica dos enfoques históricos que abordan el tratamiento del EMD.

El enfoque del EMD métrico o "clásico" asume que las proximidades en el hiperespacio cumplen una relación lineal (explícita) con las distancias en el espacio de la configuración (Wish y Carroll, 1982). Gower (1966) lo llamó "coordenadas principales" y señaló aspectos en común con el método de las Componentes Principales. Este enfoque, cuya aplicación fundamental es la reconstrucción de mapas a partir de matrices de distancias (Timm, 2002) no fue aplicado en la tesis (ver párrafos sucesivos) pero el lector interesado puede encontrar una muy buena presentación del tema en el Capítulo 9 de Timm (2002).

EL enfoque del EMD no métrico no impone (como el métrico) que las magnitudes de las disimilitudes sean proporcionales a las magnitudes de las distancias. Tal restricción se abandona en virtud de que en muchos casos las proximidades exactas no se conocen (son percepciones o las medidas tienen error), por lo que solo se considera el orden de las proximidades en relación al orden de las distancias en la configuración. Por lo tanto, la problemática queda abordada desde una perspectiva más amplia (McCune y Grace, 2002) tolerando mejor la presencia de no linealidades en los datos (Kenkel y Orlóci, 1986). Si por ejemplo, la configuración obtenida se utilizara para detectar patrones, el EMD no métrico tolerará mejor algo de distorsión, puesto que es el ordenamiento de rangos lo que se pone en juego y no las magnitudes (Seber, 1984). Dado que el EMD no métrico solo requiere del ordenamiento de rangos esto posibilita el uso de una gran variedad de medidas de disimilitud. El término "no métrico" solo refiere al hecho de que lo central son los rangos entre disimilitudes o distancias, o sea; que las distancias entre objetos en la configuración aumenten (o disminuyan) con el mismo orden que las disimilitudes aumenten (o disminuyan) en el hiperespacio sin importar la magnitud con que lo hacen.

El enfoque no métrico recomendado por Maronna (CP) para las aplicaciones de esta tesis ha mostrado muy buen desempeño en estudios comparados con otros métodos de ordenación (Kenkel y Orlóci, 1986; Roux, 2008). Si bien es un método fundadamente recomendado en otros campos de aplicación (Clarke, 1993; McCune y Grace, 2002; Borg y Groenen, 2005) no se encontraron aplicaciones específicas en relación al estudio de vientos.

V.6.1 EMD no métrico

El investigador parte de una matriz de n datos en p dimensiones, $\boldsymbol{X}_{n\,x\,p}$ o de una matriz de proximidades $\Delta_{n\,x\,n}$ cuyos elementos son δ_{rs} ($r,\ s$ =1, n). En el caso de que esta última matriz esté dada por similitudes las mismas se convertirán a disimilitudes; por otra parte cualquiera sea la disimilitud, el método no requiere satisfacer la desigualdad triangular.

El objetivo es hallar una configuración de n puntos en k dimensiones, $\boldsymbol{Y}_{n\,x\,k}$ ($k << p$, en general k=2) tal que satisfaga que las distancias Euclídeas entre dichos elementos $d_{rs} = \sqrt{\sum_{r<s}^{n}(y_r - y_s)^2}$ constituyan una buena representación de las disimilitudes δ_{rs}. Puesto que no existirá un encaje perfecto debido a apartamientos de la monotonicidad y a que

habrá que partir de una configuración inicial, tendrá lugar un proceso iterativo. Por lo tanto, se busca que la escala ordinal de distancias en la configuración (d_{rs}) encaje lo mejor posible con la escala ordinal de las disimilitudes en los datos originales δ_{rs}. O sea, se busca una relación lo más monótona posible entre ambas escalas. Para que esto sea posible se define una función que posibilite minimizar los apartamientos de la monotonicidad.

La Figura V.16 muestra cualitativamente la relación entre las distancias en la configuración (eje *X*) y las disimilitudes en el espacio de los datos (eje *Y*). Las letras minúsculas designan puntos en los distintos espacios dimensionales.
Los datos originales ordenados cumplen que:

$$\delta_{ab} \leq \delta_{cd} \leq \delta_{ef} \leq \delta_{gh} \leq \delta_{ij} \leq \delta_{kl}$$

Para que satisfaga el requisito de monotonicidad deberá cumplirse que:

$$d_{ab} \leq d_{cd} \leq d_{ef} \leq d_{gh} \leq d_{ij} \leq d_{kl}$$

En el ejemplo esta relación no se cumple para el punto 4 puesto que entre los puntos 3 a 4 δ crece mientras que d decrece.

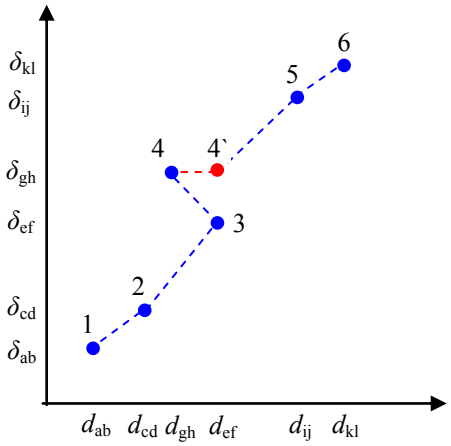


Figura V.16: Disimilitudes vs. distancias en la configuración.

Por lo tanto, se deberá recurrir a una función que haga mínimo los apartamientos posibles entre las δ_{rs} y las d_{rs}. Puesto que son las d_{rs} las que romperán la relación de monotonicidad son ellas pues las que deberán ser "corregidas".

Puede definirse una función error $S^* = \sqrt{\sum_{r<s}^{n}(d_{rs} - \hat{d}_{rs})^2}$ (llamado a veces "estrés bruto") donde $\hat{d}_{rs}$ define a un valor llamado frecuentemente disparidad que permite que se minimicen diferencias entre disimilitudes y distancias. Las $\hat{d}_{rs}$ se obtienen aplicando una regresión isotónica (Kruskal, 1964b). Si S^* es cero entonces la relación entre las δ y las d se hace monotónicamente perfecta.
Puesto que la configuración de puntos que minimizan S^* no es única (cualquier transformación rígida tal como una rotación o traslación dará lugar al mismo S^* y lo mismo ocurre afectando la configuración por un escalar) es conveniente estandarizar S^*.

$$\frac{S^*}{\sqrt{\sum_{r<s}^{n} d_{rs}^{\,2}}} = S = \sqrt{\frac{\sum_{r<s}^{n}(d_{rs} - \hat{d}_{rs})^2}{\sum_{r<s}^{n} d_{rs}^{\,2}}} \qquad \text{ec. V.3}$$

El S en la ecuación V.3 es llamado "estrés" con la sigla STRESS que refiere a "standarized residual sum of square". La fórmula presentada es llamada frecuentemente Stress-1 debido al denominador empleado (Kruskal y Wish, 1978). S informará sobre el grado de apartamiento de la monotonicidad y, por lo tanto, reflejará cuan bien la configuración encaja con las observaciones. Dada una matriz de entrada de disimilitudes la minimización de S se realiza -según Kruskal (1964 a,b)- aplicando una regresión isotónica por mínimos cuadrados y un algoritmo iterativo llamado de "descenso más empinado" ("steepest descent"). Dada la complejidad matemática del algoritmo y el alcance de la presente tesis se refiere al lector al trabajo original de Kruskal (Kruskal, 1964 a,b) que trata el tema en detalle o al Capítulo 3 de Cox y Cox (2001) donde se muestran variantes.

El coeficiente *S* definido con la ec. V.3 está acotado entre 0 y 1 y puede ser fácilmente expresado en porcentaje. Kruskal (Kruskal, 1964a) sugirió una escala basada en la experiencia con observaciones y datos simulados tal que: *S*(%)= 0 implica un ajuste perfecto, *S*(%)=2.5 excelente, *S*(%)=5 bueno, *S*(%)= 10 regular y *S*(%)≥20 pobre. Puesto que estos valores deben ser considerados como una guía para el investigador, es conveniente tener en cuenta una clasificación similar hecha por Clarke (1993) dentro del campo de la ecología en la cual *S*(%)<5 se considera excelente, *S*(%)=5- 10 buena sin riesgos de mala interpretación, *S*(%)=10- 20 si el valor está cerca del límite inferior es útil, si está cerca del superior es poco confiable, *S*(%)>20 la interpretación se vuelve riesgosa (tener en cuenta que valores entre 35- 40 se asocian a una configuración al azar). Según el mismo autor, en las aplicaciones de comunidades ecológicas, valores de *S*(%) entre 10 y 20 son bastante usuales. En cualquiera de los casos, estos "valores guía" deben emplearse con precaución (por ejemplo, si no se investiga la presencia de atípicos no se sabrá si un *S* alto será debido a todo el conjunto de datos o a un solo punto influyente).
Debe tenerse en cuenta que el estrés depende del número de objetos en estudio (Kruskal y Wish, 1978), *S* tenderá a aumentar para muestras grandes comparadas con otras pequeñas. Si $n > 4\,k$ (n = número de objetos y k = número de dimensiones) la interpretación de *S*(%) no será afectada pero si n se acerca a k, por ejemplo, 7 objetos en 3 dimensiones aún obteniendo un *S*(%)=2 no implicará que haya buen encaje entre los datos y la configuración. También afectarán el valor de *S*(%) la presencia de muchos valores atados y el número de variables originales (McCune y Grace, 2002).
Aquí es pertinente agregar que, tal como lo visto para análisis por conglomerados, la forma de estandarizar los datos influirá en los resultados de la configuración (Kenkel y Orlóci, 1986).
Existen desventajas "tradicionales" asociadas a la aplicación de EMD no métrico: a) la presencia de mínimos locales y b) la lentitud de cálculo. Sin embargo, los avances en los programas de cómputo y de hardware han eliminado ambas problemáticas (McCune y Grace, 2002).
Los cálculos de EMD presentados en este capítulo fueron llevados a cabo por el Dr. Ricardo Maronna (coautor de varias de las publicaciones asociadas a la presente tesis) y recalculados con el software *Statistica 8.0* que utiliza un algoritmo equivalente, los resultados fueron en todos los casos de alta coincidencia.

V.6.2 Ejemplo de aplicación

La Tabla V.4 muestra coeficientes de correlación robustos utilizados para comparar observaciones de ocurrencia de calmas en distintos sitios de monitoreo (Punto A, Punto J y Punto K- Figura II.6- Capítulo II) según las estaciones del año (Ratto et al., 2012c). La última columna muestra los promedios de a pares entre sitios de monitoreo, la última fila muestra los promedios estacionales. El promedio general del coeficiente de correlación se halla en la última fila de la última columna.

Tabla V.4

Sitios	**Verano**	**Otoño**	**Invierno**	**Primavera**	Promedio
A,J	0,9846	0,9567	0,9293	0,9742	0,9612
A,K	0,9891	0,9766	0,9112	0,9925	0,9674
J,K	0,9520	0,7415	0,8020	0,9752	0,8677
Promedio	0.9752	0.8916	0.8808	0.9806	**0.9321**

Tabla V.4: Valores del coeficiente de correlación MCD (Sección IV.2.1- Capítulo IV) referidos a las curvas de calmas observadas en distintos sitios de monitoreo para las distintas estaciones del año.

Si bien la aplicación de EMD será de más utilidad para tablas grandes (dado que en las pequeñas es más fácil visualizar el arreglo de números) se tomó, por simplicidad, la Tabla V.4 como ejemplo. La misma permite apreciar de forma "pictórica" los parecidos y las diferencias (Borg y Groenen, 2005) entre sitios y estaciones, o sea, una vista general de la

estructura que adoptan los coeficientes. Para simplificar se han evitado graficar los promedios correspondientes a los sitios y a las estaciones (salvo el promedio general). La Figura V.17 muestra la configuración obtenida en el plano utilizando el enfoque no métrico de Kruskal; el coeficiente de estrés *S* correspondiente es 13%. Los puntos graficados tales como "J ot" surgen del cálculo de una matriz intermedia de distancias, de 13 x 13 ($4_{(\text{estaciones del año})}$ x $3_{(\text{sitios})}$ + $1_{(\text{promedio})}$). "J ot" indica la ubicación en el plano para el otoño en el Punto J de monitoreo que guarda una relación de distancias con el resto de los puntos de la configuración incluyendo el promedio general (centro de la figura). A título general, la secuencia característica de cálculos se describe en el Anexo V.4 (pág. 184).

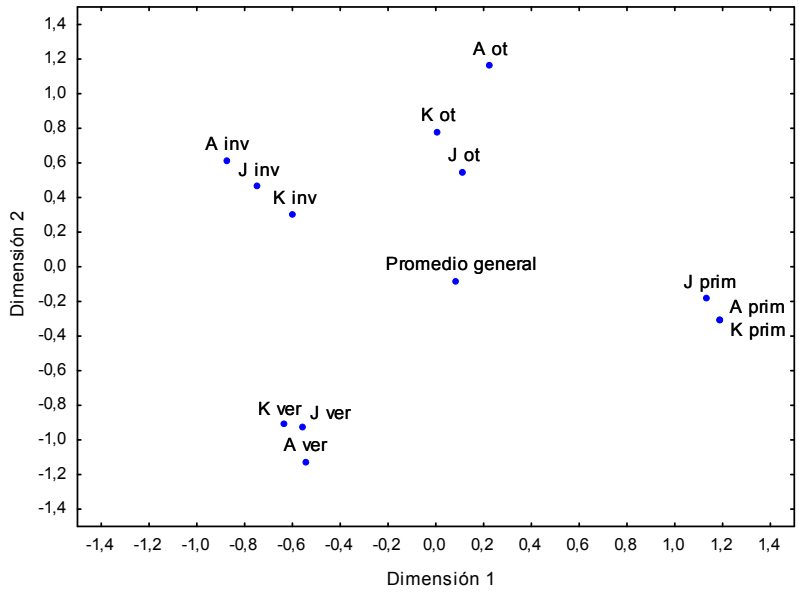


Figura V.17

Figura V.17: Configuración en dos dimensiones. La misma fue obtenida a partir de los coeficientes de correlación de la Tabla V.4 excepto los valores promedios de sitios y estaciones del año. Los ejes (dimensión 1 y dimensión 2) no tienen un significado absoluto sino que reflejan distancias relativas entre los puntos del plano (configuración hallada).

La gráfica permite apreciar una clara separación entre estaciones del año. El promedio general de las correlaciones de la tabla se indica a modo de referencia. Dos puntos cercanos indican que tienen correlaciones similares. Durante las estaciones cálidas (verano y primavera) las calmas forman grupos bastante cohesionados mientras que en las más frías hay más diferencia entre los distintos sitios de monitoreo. Un eje horizontal imaginario que pase por el punto del promedio general dividiría a la configuración en puntos de mayor correlación (por debajo) y de menor correlación (por encima) –ver Tabla V.4. Notar que las primaveras en A y K tienen un punto coincidente, esto refleja que ambos sitios están altamente correlacionados (Tabla V.4).

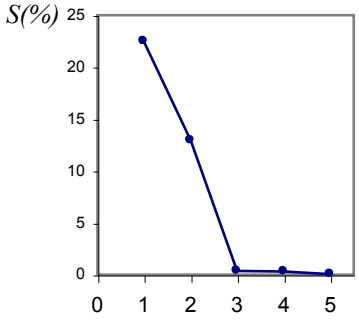


Figura V.18: STRESS (eje *Y*) versus número de dimensión (eje *X*).

Puesto que el valor del estrés puede no conformar al investigador (recordar valores propuestos por Kruskal y Clarke– Sección V.6.1) puede realizarse el gráfico de dimensionalidad versus *S* (Kruskal, 1964a). La Figura V.18 muestra una curva típica realizada con los datos de trabajo operando con el software *Statistica 8.0* hasta 5 dimensiones. Según la figura el codo define que la dimensión ideal es tres a la cual pertenece un *S*(%)= 0.37 que es casi perfecto.

La Figura V.19 permite apreciar de manera aún más definida la estructura de grupo que muestran las correlaciones para cada una de las estaciones del año.

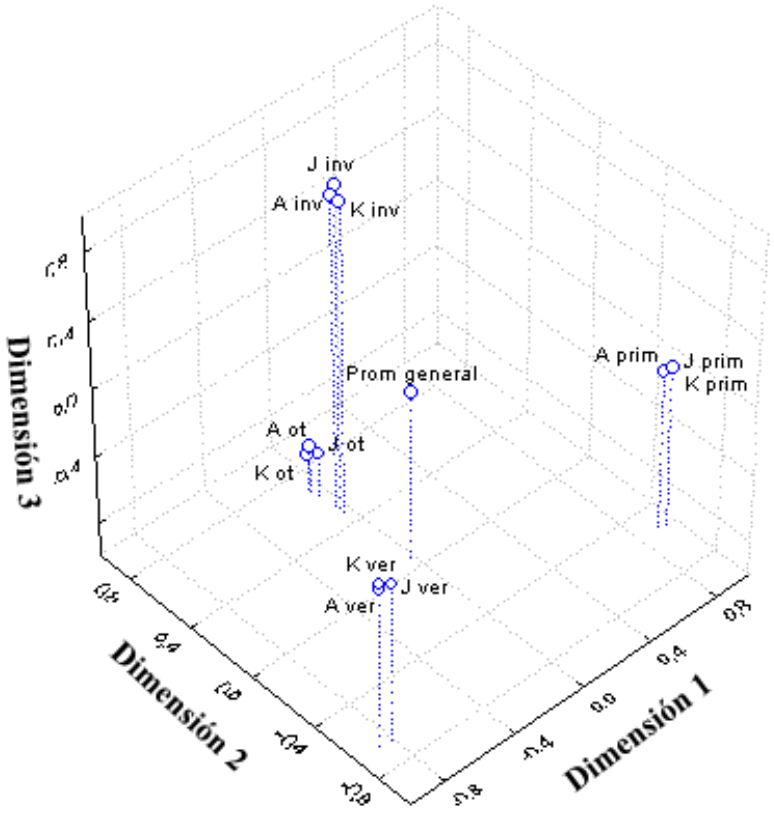


Figura V.19: Configuración en tres dimensiones.

V.7 Misceláneas

En las secciones hasta aquí presentadas de este capítulo el análisis por conglomerados jerárquicos ha tenido un rol protagónico. Además de otros métodos no supervisados tales como el de las *k*-medias (partición) y el método de las Siluetas (uno de los métodos de agrupamiento difuso -"fuzzy clustering"- que se presenta más adelante) existen métodos con mayor nivel de formalidad que también permiten encontrar grupos (Everitt et al. 2011). Estos métodos, dejan atrás en gran parte la perspectiva heurística (en el sentido de no hacer suposiciones explícitas acerca de la estructura de los datos) y suelen referirse como basados en modelos estadísticos (Ritter, 2015), en donde se considera que los datos en los que se desea encontrar grupos (muestra) provienen de una población que contiene subpoblaciones (grupos o "clusters") cada una caracterizada por una función densidad de distribución multivariada particular (por ejemplo, normales con distintas medias y varianzas, *t*-Student, *Chi*-cuadrado, etc.).

Cada una de estas "categorías clásicas" (jerárquicos, de partición y aquellos basados en un modelo) ofrece una gran variedad de métodos y nuevos enfoques, por ejemplo, basados en la densidad espacial, el de agrupamiento con restricciones o el de aglomeración en dos pasos ("two-step clustering") que van enriqueciendo las aplicaciones. Everitt et al. (2011) reflejan de manera abarcativa y sintética este universo de métodos donde, como se señaló en la Sección V.1, convergen aportes de diversas disciplinas. Cabe agregar que el tipo de variable que se utiliza (continuas, binarias, categóricas, etc.) y el objetivo de la investigación permiten circunscribir el método a seleccionar.

El análisis por escalamiento multidimensional (EMD), utilizado para visualizar datos y detectar patrones es muy empleado en algunas ciencias sociales (principalmente en la psicología) pero está poco difundido en las disciplinas ambientales. Los métodos involucrados (métrico o no métrico) permiten reducir dimensionalidad en datos con

muchas variables sin necesidad de conocer la matriz de datos (requisito necesario en el método de las CP). Por ejemplo, en una muestra de automóviles los encuestados expresan que autos se parecen más entre si pero las consideraciones (variables) tenidas en cuenta por la persona no se saben. El "mapa" dará finalmente cuales son las preferencias de determinado grupo. Como en cualquier disciplina el reconocimiento de parecidos permitirá investigar acerca de las causas (variables puestas en juego) que le dan lugar.
Variantes menos usadas del EMD son el EMD con restricciones y el de desenrollamiento ("unfolding"). Reuniendo variantes de métodos de conglomerados y de EMD Shepard (1980) hace una interesante discusión sobre la aplicación de las distintas herramientas. Borg et al. (2013) presentan una reciente síntesis de EMD con variantes nuevas y sus aplicaciones.

V.8 Aplicaciones

V.8.1 Patrones horarios de vientos en La Plata y alrededores

En la Figura V.20 se muestran 24 rosetas de frecuencias de ocurrencias de vientos por dirección para cada hora del día (promedio horario de observaciones realizadas entre 1998 y 2003) para el verano en el Punto J. Este conjunto de datos se muestra como ejemplo y será analizado junto a los tres conjuntos restantes correspondientes a las otras estaciones del año y a los cuatro conjuntos análogos de datos correspondientes al Punto A para el mismo período.
Con el objeto de proveer fundamentos para los resultados obtenidos de esta sección (que está dedicada al estudio de las observaciones de los puntos A y J), se utilizaron como referencia dos conjuntos de datos: los del Punto K (Aeropuerto de La Plata- Figura II.6- Capítulo II) y los de 5 estaciones meteorológicas en la zona del Río de La Plata pertenecientes a la red de estaciones meteorológicas del Servicio Meteorológico Nacional (Figura II.4a- Capítulo II). El lector puede encontrar una síntesis comparativa de los vientos en la Sección II.1.2- Capítulo II de esta tesis y un desarrollo complementario en la Sección 4 de Ratto et al. (2010b).

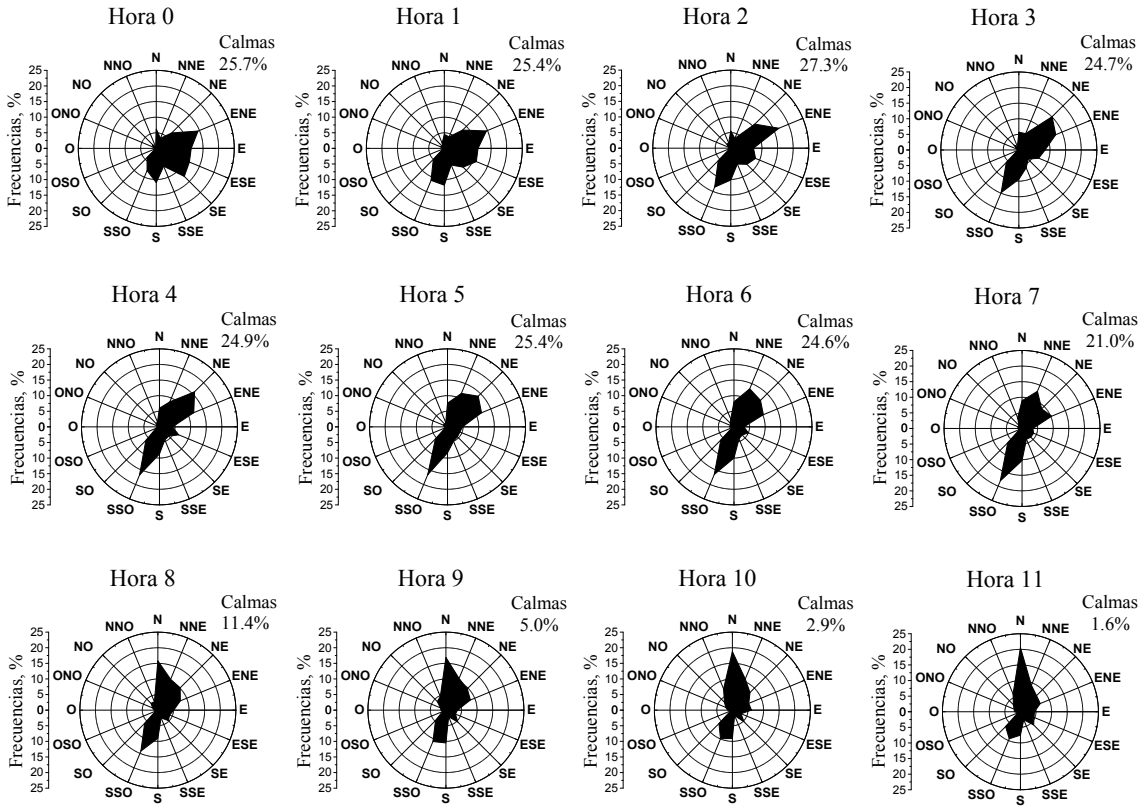


Figura V.20 (continúa en la página siguiente)

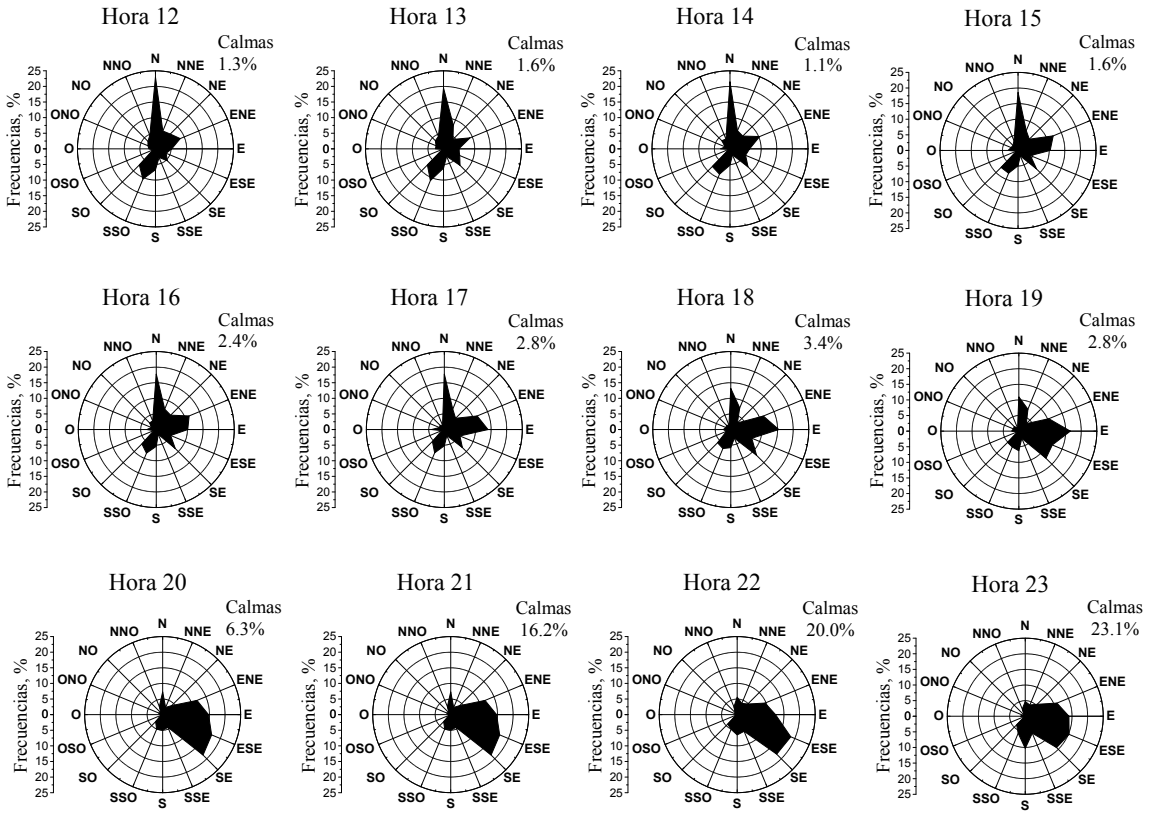


Figura V.20: Rosetas horarias promedio de frecuencias de viento por dirección observadas para la estación verano durante el período 1998- 2003 en el Punto J (Figura II.6). Los bloques horarios refieren a la Hora Local, por ejemplo, Hora 0 equivale a 00:00- 00:59 Hora Local. Las calmas están expresadas como la cantidad de observaciones menores a 1.6 km h^{-1} respecto del total de observaciones. La velocidad media observada para esta estación durante el período es de 6.6 km h^{-1}.

Con el objetivo general de asistir a la comparación de este gran conjunto de datos (192 rosetas horarias de 16 direcciones: $24_{(horas)}$ x $4_{(estaciones\ del\ año)}$ x $2_{(sitios)}$) y develar su estructura intrínseca, se aplicaron los métodos de análisis por conglomerados y escalamiento multidimensional. Ambos métodos son complementarios, en cuanto a que permiten sintetizar información y así describir con mayor claridad la ocurrencia de patrones de vientos; al mismo tiempo se facilita la interpretación física de los fenómenos involucrados.

Como antecedentes de aplicación de análisis por conglomerados para evaluar distintos tipos de patrones climáticos cabe citar a Kalkstein et al. (1987), Wolter (1987), Gong y Richman (1995), Fovell y Fovell (1993), Huth et al. (1993), Jackson y Weinard (1995), Unal et al. (2003), mientras que para patrones más específicos relacionados a los vientos se pueden citar Kaufmann y Whiteman (1999), Darby (2005), Beaver y Palazoglu (2006) y Jiménez et al. (2008). No se hallaron antecedentes de aplicaciones de escalamiento multidimensional a patrones de viento.

Volviendo a la Figura V.20, la forma típica de agrupar rosetas de viento es definiendo un intervalo horario fijo a partir de determinada hora del día (por ejemplo, de a tres horas tomando como punto de partida la Hora 0 y calculando los promedios de los bloques horarios (Alvarez Escudero y Alvarez Morales, 2001)). Pero esta modalidad, depende de la elección subjetiva del investigador y puede enmascarar particularidades. El análisis por conglomerados provee de una herramienta más objetiva y flexible para agrupar a los

individuos en grupos según su similitud/disimilitud. La flexibilidad viene dada porque los grupos formados pueden contener números distintos de individuos. El análisis por conglomerados se realizó siguiendo los pasos delineados en la Sección V.5.
Al comenzar la exploración de datos se supone que los datos poseen estructura intrínseca de grupo. Dados los vientos dominantes a escala sinóptica y el ciclo diario que tiene lugar en la CLP (capa límite planetaria) junto a otros fenómenos que caracterizan al viento, tales como los involucrados en el ciclo de brisa de mar y tierra (Sección III.10- Capítulo III), se espera que los vientos diarios posean una estructura de grupo reconocible.

En la Figura V.20 (dado el tipo de representación en estrella (Jacoby, 1998) que facilita enormemente la visualización de las dimensiones vectoriales presentes en cada individuo) es fácil suponer que dicha estructura tiene carácter cíclico. De la simple inspección de la figura puede observarse que rosetas de viento consecutivas tienden a parecerse mientras aquellas que forman opuestos (por ejemplo, Hora 0 y Hora 12) tienden a diferir mucho. Por otra parte, si comenzamos nuestra inspección por ejemplo, por la Hora 0 y avanzamos en sentido horario, a medida que nos alejamos vamos observando más diferencia entre la Hora 0 y las horas más alejadas. Este alejamiento se da hasta cierto punto a partir del cual las horas subsiguientes comienzan a parecerse cada vez más hasta llegar a la Hora 23.

Como se expresó en la Sección V.1, existen diversos métodos para analizar grupos. Por las ventajas ya descriptas, se recurrió a un **método aglomerativo** y **jerárquico**; este último porque brinda un rango coherente de posibles agrupaciones dejando al investigador la determinación del número óptimo de grupos. Si bien se sabe que, en principio, hay cuatro momentos del día bien diferenciados (amanecer, día, anochecer y noche) en las que pueden identificarse distintos patrones de viento (cabe suponer que existirán pocos grupos) no se sabe con mayor detalle cuantos grupos pueden ser los que mejor representen a los 24 patrones horarios ni como estarán conformados los grupos.
Otros métodos, como el popular método de las *k*-medias (Anexo V.6, pág. 188) o el de las siluetas (Sección V.8.5), requieren de un conocimiento previo para definir el número de grupos. En ellos, las distintas soluciones posibles (y que son recomendables de ensayar) pueden dar lugar a grupos con individuos muy distintos ya que los agrupamientos no están anidados. En contraste, el análisis por conglomerados jerárquicos con una sola corrida permite visualizar distintos agrupamientos (soluciones posibles) según se elija la distancia de corte.
Las direcciones de los vientos de una zona de estudio siguen en promedio un patrón diario "típico". Dicho patrón implica la rotación de los vientos con características (direcciones involucradas, velocidades, etc.) que pueden determinarse. Por lo tanto, cabe suponer que las variables involucradas en la roseta de ocurrencia de vientos por dirección se hallen correlacionadas. Por lo visto en la Sección V.2.1, no es deseable que la correlación entre pares de variables sea muy alta ya que esto distorsiona los resultados. Con la finalidad de explorar la necesidad de **seleccionar variables** se recurrió a calcular las matrices de correlación para los 8 conjuntos de datos. Los resultados mostraron que no había correlaciones lo suficientemente fuertes como para descartar variables.
La **completitud de datos** para el Punto A es en promedio 84.1 % mientras que para el Punto J es de 94.2%. La inspección de los datos permitió identificar al invierno de 2000 en el Punto J como un bloque con alto grado de incompletitud. Mientras que el resto de los datos faltantes se hallaban distribuidos al azar a lo largo de las estaciones y los años; esto hizo que, por cuestiones prácticas, solo se procediera a aplicar un **método de relleno** para el caso del invierno de 2000 (promedios horarios porcentuales de cada frecuencia por dirección).

Por simplicidad se adoptó un algoritmo basado en la minimización de sumas al cuadrado. Para realizar la estimación se recurrió a los conjuntos de datos completos, o sea, los correspondientes a los veranos, otoños y primaveras en el Punto J. Se tomaron como vectores de referencia a los dos vecinos más próximos anteriores y posteriores al 2000.

Sea $M_{n\ x\ p}^{Verano\ J} = \begin{pmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1p} \\ x_{21} & & \cdots\cdots & \\ x_{31} & & \cdots\cdots & x_{3p} \\ x_{41} & & \cdots\cdots & \\ x_{n1} & & \cdots\cdots & x_{np} \end{pmatrix} \begin{matrix} \leftarrow 1998 \\ \leftarrow 1999 \\ \leftarrow 2000 \\ \leftarrow 2001 \\ \leftarrow 2002 \end{matrix}$ la matriz de datos de los veranos entre

los años 1998 y 2002. n =5 (datos) y p =16 (variables involucradas). Para cada x_{ij}, i indica el objeto o vector y j la variable. En general, es aconsejable realizar cálculos con varios "métodos" para determinar cual es el que mejor realiza el "relleno", o sea, aquel que minimice la suma de cuadrados. Para ello primero se elimina la fila 3 (año 2000) y se calcula: a) promedio, b) mediana, c) regresión lineal por cuadrados mínimos y d) un polinomio de grado dos por cuadrados mínimos.

Luego se calcula la suma de cuadrados: $S_{\mathrm{Pr}omedio}^{Verano\ J} = \sum_{\substack{i\neq 3 \\ j=1,16}} (x_{ij} - P_{ij})^2$ ec. V. 4

donde P_{ij} son los elementos del vector fila "promedio" (caso a)). Esto mismo se hace para las tres estaciones de datos completos (obteniendo además $S_{\mathrm{Pr}omedio}^{Otoño\ J}$ y $S_{\mathrm{Pr}omedio}^{\mathrm{Pr}imavera\ J}$). Luego se calcula $S_{\mathrm{Pr}omedio}^{TOTAL} = S_{\mathrm{Pr}omedio}^{Verano\ J} + S_{\mathrm{Pr}omedio}^{Otoño\ J} + S_{\mathrm{Pr}omedio}^{\mathrm{Pr}imavera\ J}$.

Se procede análogamente con los otros métodos propuestos (casos b) c) y d)) y se determina la menor de las sumas al cuadrado. En el caso que se ejemplifica, la menor suma dio para $S_{Mediana}^{TOTAL}$. En consecuencia, se adoptó la mediana como reemplazo de los datos faltantes de la rosa de vientos del invierno de 2000.

Para explorar la presencia de potenciales **valores atípicos** se llevaron a cabo algunas de las herramientas descriptas en la Sección V.2.4, no detectándose la presencia de los mismos.

Si bien los valores de todas las variables están dados en la misma unidad (frecuencias de dirección de vientos) se recurrió al proceso de **estandarización** con el fin de poder comparar los resultados de los distintos conjuntos de datos (dados por las estaciones del año y los sitios de monitoreo). Demostrada la ausencia de valores atípicos y la necesidad de comparar los resultados con otros hallados en publicaciones previas (Ratto et al., 2010a) se estandarizó con media aritmética y desvío estándar. Teniendo en cuenta lo discutido en las secciones V.3 y V.4 y según algunos autores (Kalkstein et al., 1987; Fovell y Fovell, 1993; Huth et al., 1993) se adoptó a la distancia Euclídea al cuadrado como medida de disimilitud y a la distancia promedio ("average linkage- UPGMA"- Sección V.4 y Anexo V.1, pág. 171) como criterio de disimilitud entre grupos.

En la Figura V.21 se muestra el dendograma correspondiente a las rosetas de viento de la Figura V.20. En esta figura se indican (a modo de ejemplo) tres distancias de corte que dan lugar a distintos agrupamientos: alrededor de la distancia 35% se forman 6 grupos, alrededor de la distancia 50% se forman 5 grupos y alrededor de la distancia 75% se forman 3 grupos. La distancia de corte para 3 grupos muestra ser muy estable dado que moviéndose según el eje de las X entre 60% y 90% el número de grupos obtenidos es siempre 3. Como se señaló anteriormente y debido a la física de los fenómenos involucrados (ciclo diario de la CLP y brisa de mar- tierra) se esperan al menos 4 grupos,

por lo tanto, esta distancia de corte no parece ser del todo apropiada al caso de estudio (aunque el lector puede apreciar que divide al día en tres grupos horarios bien definidos: día, anochecer, noche).

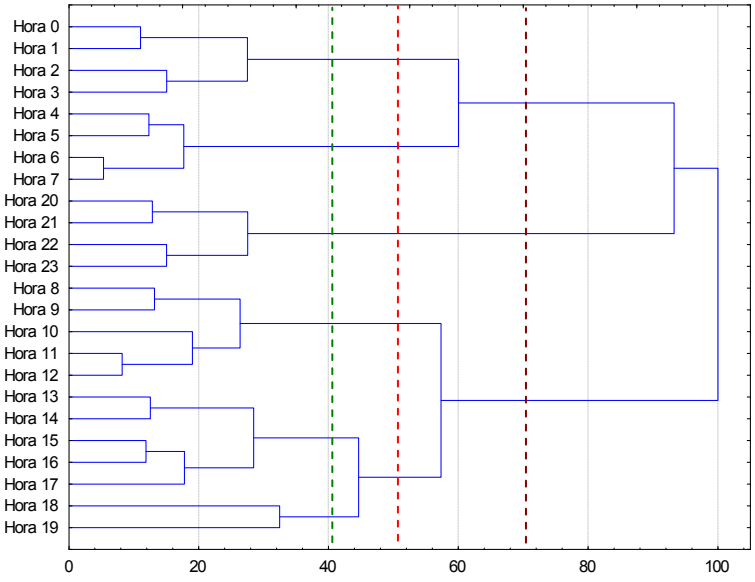


Figura V.21

Figura V.21: Dendograma de 24 rosetas horarias de viento correspondiente al verano en el Punto J para el período 1998- 2003.
En el eje de las *X* se halla representada la distancia Euclídea al cuadrado reescalada en % (para facilitar comparaciones con otros dendogramas).
En el eje de las *Y* cada "Hora" representa un vector de 16 direcciones de frecuencias de ocurrencias de vientos (Rosetas de la Figura V.20). Las líneas de trazos verticales indican posibles distancias de corte.

Las distancias de corte para 5 y 6 grupos muestran también una buena estabilidad. En el contexto del presente estudio (en donde se consideran no solo el dendograma de la Figura V.21 sino los otros siete correspondientes al resto de las estaciones del año y al Punto A) se debe buscar en cada dendograma una distancia de corte tal que forme la misma cantidad de grupos en todos los dendogramas. Otro requisito que surge de la aplicación práctica es que los miembros de cada grupo deben tener un carácter correlativo en el tiempo (no puede haber grupos con "horas" discontinuadas) para que la interpretación sea sencilla. Estas dos condiciones se agregan a las que se deben tener en cuenta al buscar definir el número óptimo de grupos. Teniendo en cuenta lo antedicho se determinó, por prueba y error, que 5 era un número óptimo en todos los casos. Tomar como referencia el verano resulta muy útil dado que es la estación que presenta mayor amplitud (variación) en las direcciones de viento, el resto de las estaciones quedan "incluidas". Cinco grupos implica, por un lado, una buena reducción respecto de los miembros originales (24) y por otro, parecen ser representativos de los fenómenos meteorológicos implicados.
La Figura V.22 muestra las rosetas de viento resultantes del análisis por conglomerados que se le aplicó a las rosetas de la Figura V.20.

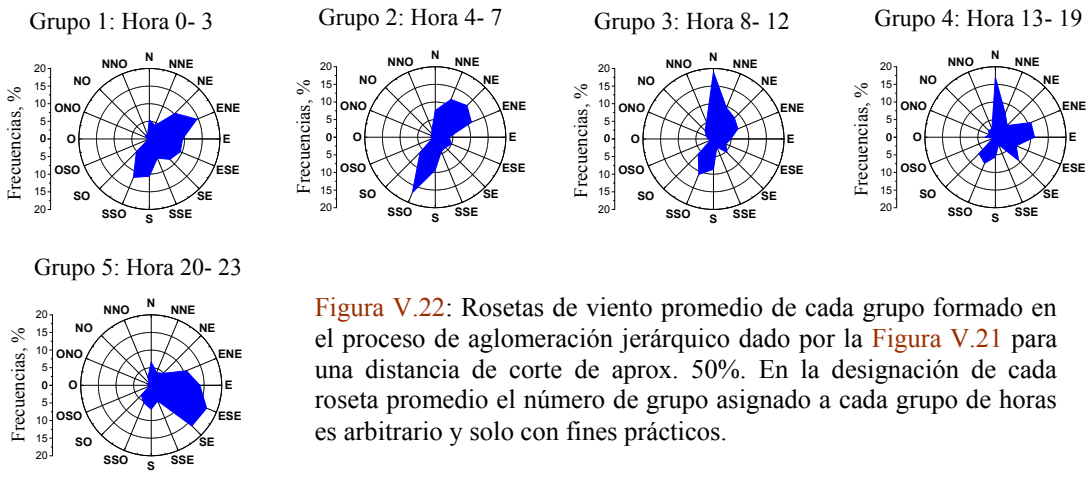


Figura V.22: Rosetas de viento promedio de cada grupo formado en el proceso de aglomeración jerárquico dado por la Figura V.21 para una distancia de corte de aprox. 50%. En la designación de cada roseta promedio el número de grupo asignado a cada grupo de horas es arbitrario y solo con fines prácticos.

Notar que los grupos obtenidos están formados por un número variable de miembros (flexibilidad); esto sucede con los 8 dendogramas realizados obteniéndose grupos de entre 2 y 7 miembros. Este fenómeno no sería apreciable al formar grupos preestablecidos como en el análisis tradicional. La variabilidad que es posible apreciar en el número de miembros de cada grupo se debe a la naturaleza de los datos pero, cabe recordar, que la adopción de otros criterios de aglomeración (tal como la regla de Ward- Sección V.4 y Anexo V.1, pág. 171) tenderán a dar grupos con números muy similares de miembros.

Si se calcula el vector resultante de cada roseta de vientos de la Figura V.22 y todos aquellos correspondientes al resto de las estaciones y sitios de observación se obtiene la Figura V.23. El módulo de cada vector resultante indica la cantidad de tiempo que las direcciones de viento no estuvieron compensadas por su opuesto. Un módulo pequeño indica que la rosa de vientos tiene sus direcciones bastante compensadas en todas las direcciones de la brújula. Un módulo grande indica que hay direcciones no compensadas y en caso de que las direcciones compensadas sean poco frecuentes puede indicar la presencia de vientos dominantes.

A pesar de que la reducción realizada es "drástica", la Figura V.23 permite inferir información relevante.

Las resultantes se hallan predominantemente en los cuadrantes primero y cuarto, lo cual se halla en coincidencia con los vientos dominantes observados en la zona para las cinco últimas décadas (1961- 2010) en el Punto K (Figura II.4c- Capítulo II). Una vista panorámica de la Figura V.23 muestra que en promedio todas las estaciones y sitios tienen un desarrollo similar durante el día. A lo largo de las "cinco etapas del día" (definidas en el análisis por conglomerados) las estaciones de verano y primavera presentan distribuciones menos compensadas que las estaciones de otoño e invierno, en donde se obtienen resultantes con módulos pequeños. A lo largo de todas las estaciones (ambos sitios) una franja horaria comprendida aproximadamente entre la Hora 20 y la Hora 23 es la que presenta mayores módulos en coincidencia con los vientos dominantes y en concordancia con lo hallado por Berri et al. (2010) para la Hora 21.

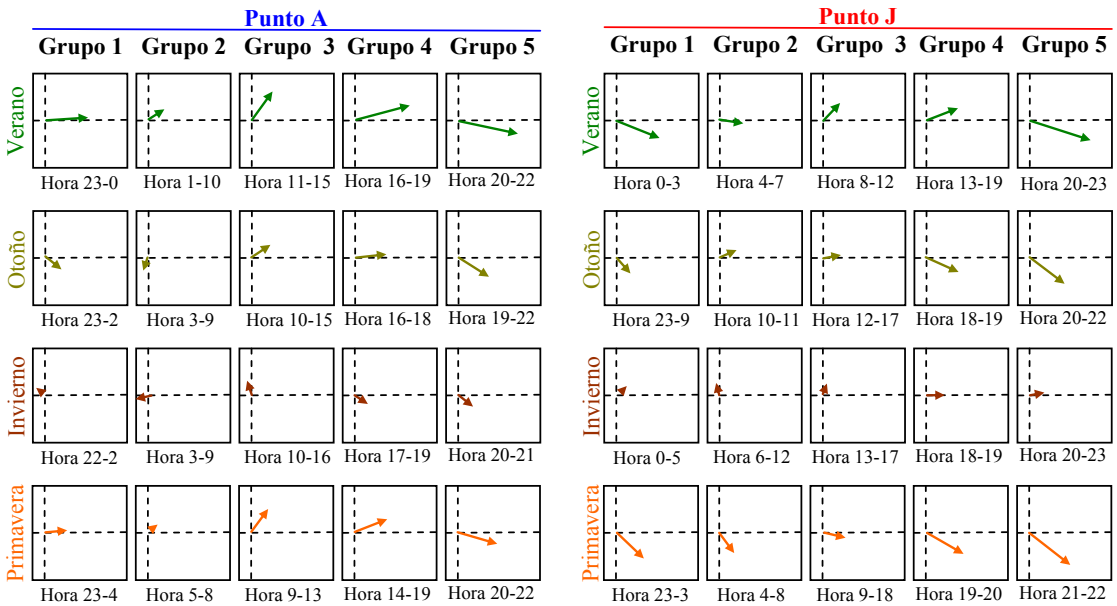


Figura V.23: Vectores resultantes (de las rosetas de frecuencias de viento promedio por dirección) de grupo para cada estación y sitio de monitoreo. La flecha indica la dirección desde donde sopla el viento. El verano en Punto J se corresponde con las rosetas de la Figura V.22. Los números naturales del 1 al 5 (por ejemplo en "Grupo 1") señalan las cinco etapas en que ha quedado dividido el día a partir de los cinco conglomerados establecidos para cada estación y sitio. Los ejes en línea punteada indican la separación en cuadrantes con un predominio de los vientos en el primero y el cuarto (derecha arriba y abajo respectivamente).

En todas las estaciones las últimas tres etapas del día (grupos 3 a 5 en Figura V.23) muestran una rotación de vientos desde el N hacia el SE (en sentido horario). A partir del mediodía las resultantes comienzan adquirir módulos más grandes mientras que durante la noche los módulos se hacen muy pequeños y con direcciones variadas. Durante la mañana las resultantes se hallan de acuerdo a lo hallado por Berri et al. (2010) para la Hora 9. A la Hora 15 vientos del E se agregan a los vientos del N y NE (dominantes durante la mañana) (Berri et al., 2010) tendencia que es observable en las etapas 3 y 4. Estas comparaciones permiten decir que, las direcciones principales de los vientos observadas en la zona de estudio (La Plata y alrededores), siguen un patrón similar a las de una zona mucho más amplia del Río de La Plata en donde los vientos dominantes se hallan influenciados principalmente por el anticiclón del Atlántico Sur y por la circulación de brisa de mar y tierra, ambos centrales para definir el Tiempo y el Clima de la zona (Sección III.2-Capítulo III).

Con el análisis por conglomerados se ha podido sintetizar información sin perder dimensionalidad, el número de rosetas originales (192) se redujo a 40 ($5_{(\text{promedios de grupo})}$ x $4_{(\text{estaciones del año})}$ x $2_{(\text{sitios})}$) y esto permitió describir de manera sencilla algunas características del desarrollo de los vientos durante el día. Con el fin de obtener más información a partir del mismo conjunto de datos se recurrió al método de EMD no métrico (descripto en la Sección V.6). Según Seber (1984) y Jain y Dubes (1988) estos dos métodos pueden operar de forma complementaria. Reduciendo la dimensionalidad de los patrones originales es posible visualizar todos los datos al mismo tiempo. La Figura V.24 muestra, como ejemplo, las configuraciones de puntos (salidas del EMD en el plano) obtenidas para el verano y el invierno en ambos sitios.

Como se ha descripto en la Sección V.6.1, para que la reducción de dimensionalidad (en este caso de 16 a 2) tenga respaldo se debe tener en cuenta el coeficiente de STRESS (ec. V.2). La Tabla V.5 indica que, a excepción del invierno en el Punto J, los coeficientes según Kruskal y Clarke son buenos.

Tabla V.5

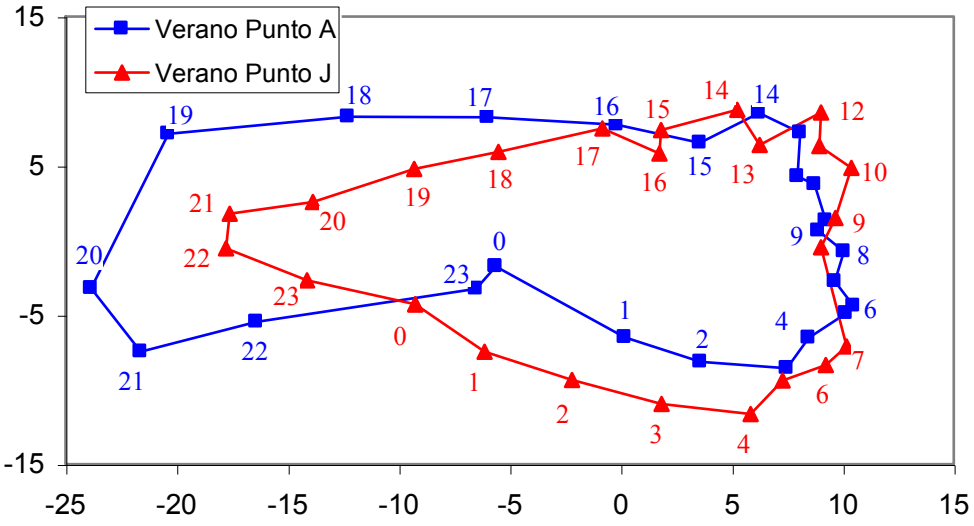
	Punto A	Punto J
Verano	3,64	2,11
Otoño	3,27	6,14
Invierno	6,87	15,43
Primavera	4,06	4,4

Tabla V.5: Coeficientes de STRESS (%) correspondientes a la reducción de dimensionalidad de 16 a 2 para todas las estaciones del año en ambos sitios de monitoreo.

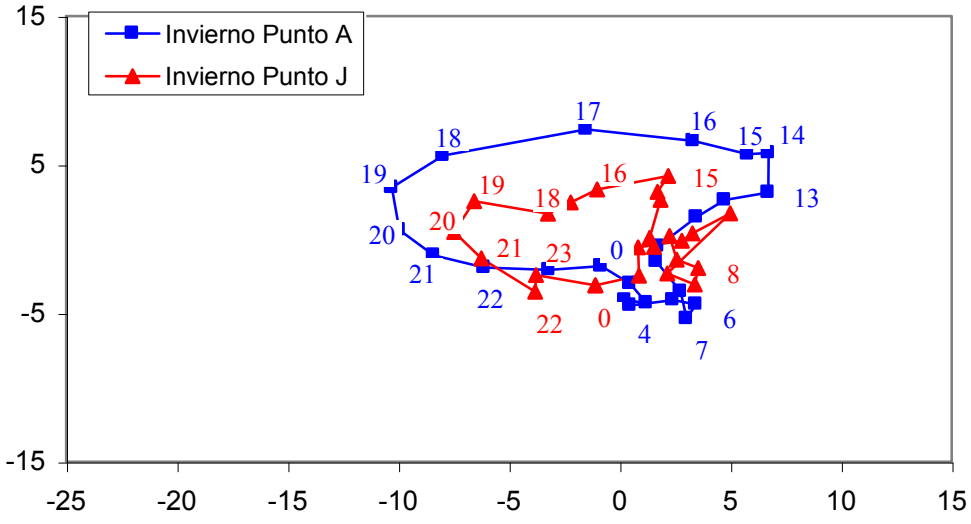
La información más importante que proveen los "mapas" de la Figura V.24 son las distancias relativas entre puntos, a mayor cercanía mayor similitud entre los puntos y, por lo tanto, entre los vectores originales (rosetas de viento de 16 direcciones). La configuración de puntos también permite apreciar tanto el carácter cíclico de los datos (aspecto discutido en relación a la Figura V.20) como su estructura de grupo.

El EMD (Figura V.24) hace muy visible como en invierno (en contraste con el verano que es la estación de mayor amplitud (Ratto et al., 2014a)) las rosetas de viento tienden a parecerse entre sí, mostrando un patrón más contraído.

Verano y primavera (esta última no mostrada) presentan más dispersión durante el día que el otoño (no mostrado) y el invierno. Estas diferencias se deben a la intensidad de la brisa de mar y tierra que es mayor en las estaciones cálidas que en las frías. Este mismo fenómeno meteorológico explica la diferencia entre sitios de observación: el Punto A (más cercano a la costa del río) es más sensible al mecanismo de brisa de mar y tierra dando rosetas horarias de direcciones de viento más dispersas entre sí (más distintas entre sí) mientras que el Punto J (ubicado más tierra adentro) muestra una distribución algo más compacta.



a)



b)

Figura V.24: Salida de EMD. Cada punto del gráfico representa una roseta horaria de vientos de 16 direcciones (correspondiente a una estación del año y un sitio de monitoreo) que ha sido reducida a un punto en el plano aplicando EMD. Los ejes *X* e *Y* están dados en unidades arbitrarias. El número cercano a cada cuadrado o triángulo refiere a la hora del día de la roseta original, algunas etiquetas han sido omitidas por cuestiones de claridad. Las líneas que unen puntos (azules) para el Punto A y (rojas) para el Punto J han sido dibujadas como ayuda para la visualización.
a) Veranos 1998- 2003 en los puntos A y J. b) Inviernos 1998- 2003 en los puntos A y J.

Introduciendo nuevamente la primavera y el otoño (salidas de EMD no mostradas) y comparando las franjas horarias entre estaciones cálidas (verano y primavera), se observa que las mismas presentan patrones similares entre la Hora 0 y la Hora 16 (sentido horario), mientras que entre las estaciones frías (otoño e invierno) las mayores similitudes se hallan entre las horas 22 y 12 (sentido horario).

V.8.2 Definiendo regionalidad en una zona amplia del Río de La Plata

Berri et al. (2010) presentan un modelo de capa límite para mesoescala que simula los vientos en capas bajas de la atmósfera con alta resolución espacial. Este modelo fue diseñado a partir de datos de la red nacional de estaciones meteorológicas del Servicio Meteorológico Nacional y su principal aplicación ha tenido lugar en la región del Río de La Plata que se indica con un rectángulo (rojo) en la Figura V.25.

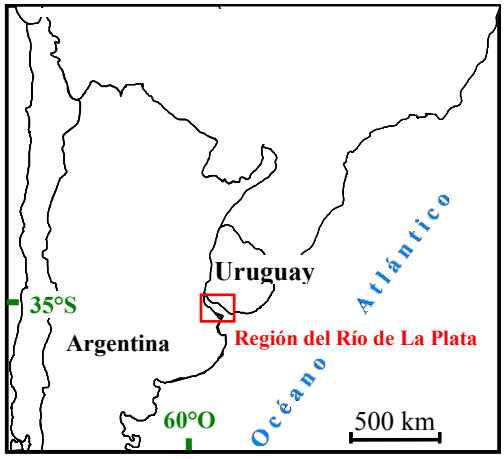


Figura V.25: Mapa parcial de la Argentina y países limítrofes. El rectángulo (rojo), que cubre aproximadamente 390 km en longitud y 285 km en latitud, es la zona de la cuenca del Río de La Plata en donde tiene alcance el modelo de predicción de vientos.

El modelo permite trabajar con una resolución horizontal de hasta 5 km pero para reducir los datos de salida se optó por una de aprox. 22 km. La salida del modelo consiste en vectores de viento de 17 dimensiones (8 para las frecuencias por dirección, una para la calma y 8 para las velocidades medias por dirección) dispuestos en una grilla de 180 puntos (18 en el eje este- oeste y 10 en el eje norte- sur).

La salida de trabajo del modelo que se utilizó es la de vientos a 10 m de altura.

La calma queda definida como todo valor de velocidad menor a 1 km h^{-1} (~0.28 m s^{-1}).

Para conocer los detalles del modelo de mesoescala, así como para obtener una descripción de la climatología de los vientos en la zona de estudio, el lector puede referirse a Berri et al. (2010).

En esta sección se discute la aplicación de análisis por conglomerados a la salida provista por el modelo, con un primer objetivo de sintetizar información "espacial" sobre la ocurrencia de vientos y de esta forma asistir a la discusión de los fenómenos meteorológicos involucrados.

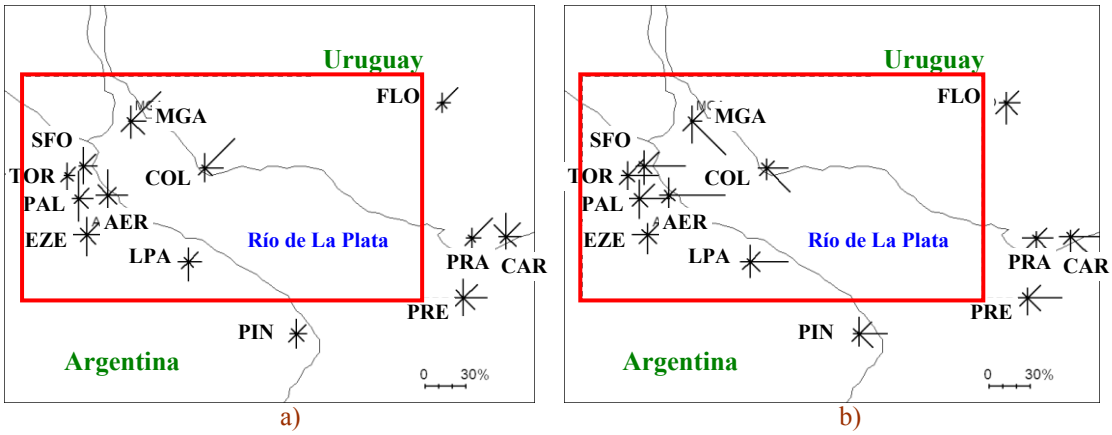


Figura V.26: Frecuencias promedio de direcciones de viento observadas entre 1994 y 2008 expresadas en porcentaje: a) Hora 6 y b) Hora 18. El rectángulo interior (rojo) indica la región en que se llevó a cabo el estudio de análisis por conglomerados. Las estaciones meteorológicas, en orden alfabético son: Aeroparque (AER), Carrasco (CAR), Colonia (COL), Ezeiza (EZE), Florida (FLO), La Plata Aero (LPA o Punto K en la Figura II.6- Capítulo II), Martín García (MGA), El Palomar (PAL), Punta Indio (PIN), Prado (PRA), Pontón Recalada (PRE), San Fernando (SFO) y Don Torcuato (TOR). La dirección Norte en los mapas se halla hacia arriba. La velocidad promedio total observada en el rectángulo en estudio para la estación verano fue de 16.2 km h^{-1} (4.5 m s^{-1}) que en la escala Beaufort (Sección III.4- Capítulo III) corresponde a Brisa leve.

Un segundo objetivo, busca identificar subáreas cuyos vientos puedan ser considerados homogéneos (base para definir regionalidad) y poder sugerir la instalación de estaciones meteorológicas en zonas que estén deficientemente representadas por las estaciones actuales.

Con el objeto de respaldar la aplicación del análisis por conglomerados se procedió, previamente, a comparar las salidas del modelo (predicciones) con las observaciones de las estaciones meteorológicas del Servicio Meteorológico Nacional que se hallan en la zona de estudio. La Figura V.26 muestra la zona del Río de La Plata en donde se aplicó el modelo (el rectángulo interior es el mismo que el de la Figura V.25). Solo con fines ilustrativos se muestran las rosetas de viento observadas en las estaciones meteorológicas de la red en el período 1994- 2008.

El *SAD* (suma de los valores absolutos de la diferencia) definido en la Sección IV.2.2- Capítulo IV, fue empleado para evaluar la disimilitud entre vectores. Esta distancia tiene la ventaja de ser más tangible que la distancia Euclídea y, por lo tanto, es más fácil asignarle un valor límite de tal modo de poder decidir cuando dos vectores pueden ser considerados como iguales.
Se recurre al *SAD* tanto para evaluar diferencias entre vectores observados y calculados como entre vectores calculados pertenecientes a distintos grupos.
Debido a las distintas magnitudes involucradas, el cálculo del *SAD* se llevó a cabo para las primeras 9 variables del vector (direcciones y calma) por un lado y para las últimas 8 variables (velocidades) por otro. Dado que (luego de varias pruebas) se determinó que la diferencia entre las velocidades (a lo largo de la grilla de 180 vectores) era mucho menor que la diferencia entre las frecuencias por dirección (y las calmas), el valor umbral para el *SAD* se definió solo en base a las direcciones. Puesto que la suma de las frecuencias para cada dirección y la calma es de 100% se adoptó un *SAD* = 10% como límite, es decir, se considera que dos rosetas de viento no difieren entre sí cuando el *SAD* en las direcciones y calma entre ambas es ≤ 10%.

Los vectores a comparar, para ilustrar en qué medida el modelo da cuenta de las observaciones, son aquellos pertenecientes a la grilla del modelo que más cerca están de las coordenadas geográficas correspondientes a las estaciones de la red (la máxima distancia para las 9 estaciones correspondientes (Figura V.26) fue de 3.54 km). La Tabla V.6 muestra los valores de *SAD* obtenidos.
El lector notará que, en la comparación, se ha incluido a la estación PRE que está afuera del rectángulo de estudio, por ser esta muy representativa de la desembocadura del río. Los valores de la tabla están en concordancia con las estimaciones del "error relativo" del modelo definidas en Berri et al. (2010) como bondad de predicción del modelo para el período 1959- 1984. Según los autores, los "errores" grandes (en nuestro caso valores altos de *SAD*) en algunos puntos (tales como AER y MGA), se dan en puntos cercanos a la costa en donde la resolución utilizada se vuelve un factor limitante. El alto valor de *SAD* para la velocidad en PRE se atribuye a que el instrumento de observación se halla ubicado a 22 m de altura y no a 10 m como en el caso del resto de las estaciones.

Para comenzar el análisis por conglomerados se realizaron las mismas consideraciones que en la sección anterior. Se optó por un método de aglomeración jerárquico, se estandarizó con media y desvío estándar (dado que las variables tienen distintas magnitudes), se adoptó la distancia Euclídea al cuadrado como medida de disimilitud y el criterio de distancia promedio (UPGMA).

Tabla V.6		
Sitio	***SAD* (%)**	***SAD* (m s^{-1})**
AER	30,1	12,0
COL	20,4	18,0
EZE	25,1	7,5
LPA	28,5	9,2
MGA	30,5	2,9
PAL	29,9	6,2
PRE	9,2	28,8
SFO	20,1	10,0
TOR	24,0	8,3

Tabla V.6: Valores de *SAD* correspondientes a las estaciones meteorológicas de la región de estudio (rectángulo interior de la Figura V.25) incluyendo a PRE.
SAD (%) expresa la suma de los valores absolutos de las diferencias entre los vectores observados y los obtenidos con el modelo (grilla en coordenadas gaussianas). Dado que estos últimos son provistos por el modelo en coordenadas preestablecidas se han buscado los puntos de la grilla que más cerca se hallen de las estaciones meteorológicas indicadas. El *SAD* promedio (sin PRE) es de 26.1% mientras que incluido PRE es de 24.2%.
El *SAD* (m s^{-1}) es análogo para velocidades de viento; 1 m s^{-1} equivale a 3.6 km h^{-1}.
Las siglas de la primera columna indican los nombres de las estaciones descriptas en la Figura V.26.

La estación verano fue considerada como referente, dado que mostró tener la mayor variación (la varianza para las velocidades es bastante pareja a lo largo de las estaciones pero para las frecuencias por dirección el verano presentó una varianza de 30.9 frente a 16.9 de la primavera, 14.8 del otoño y 9.5 del invierno). Esta característica se da también al comparar las observaciones de las estaciones meteorológicas involucradas. Por lo tanto, en el verano quedan "incluidas" el resto de las estaciones.

La Figura V.27 muestra, a modo de ejemplo, el proceso de aglomeración de los 180 vectores originales correspondientes a la salida del modelo de mesoescala para la estación verano. Una comparación con los dendogramas correspondientes a las otras estaciones del año (no mostrados por cuestiones de espacio) permite observar que para todas las estaciones, las posibles distancias de corte se hacen más estables a partir de 30%. Esto indica que los 180 vectores originales quedarán representados por un número relativamente pequeño de grupos. Con el objeto de explorar la estructura de grupo de los datos, se trabajó con distintos resultados posibles del proceso de aglomeración; se adoptaron tres distancias de corte (elegidas arbitrariamente): 48, 30 y 23, las que implican reducir el número original de 180 vectores a 6, 12 y 18 grupos respectivamente. Estas tres soluciones posibles tienen asociadas una configuración espacial dentro del área de estudio (ésta última representada en las figuras V.25 y V.26) lo cual, implica una división de dicha área en subáreas de alta homogeneidad de vientos (Figura V.28). Es decir, cada rectángulo coloreado ("píxel"), que está representado por un par de rosetas de vientos (una de frecuencias por dirección y calmas y otra de velocidades por dirección), se une a otros de similares características (para formar una subárea) según lo impone la solución elegida del dendograma. Los promedios de cada grupo (entendidos como representantes de grupo) según las tres soluciones establecidas, se muestran en la Figura V.29.

Similitud entre los grupos correspondientes a la solución de 18 grupos
Por inspección visual de la Figura V.29 es posible detectar pares de rosetas de viento muy similares, por ejemplo, las que están caracterizadas por el color cian y verde manzana o entre la amarilla y la gris oscuro (Figura V.29a). Para cuantificar estas similitudes se recurrió al *SAD* (Sección IV.2.2- Capítulo IV) obteniéndose para ambos casos valores de *SAD* menores al 10% entre las rosetas de frecuencias por dirección y calmas mientras que diferencias insignificantes ente rosetas de velocidad. Realizando el cálculo del *SAD* entre todos los pares posibles de los promedios de grupo restantes de la Figura V.29a se obtuvieron en todos los casos valores superiores al 10% para las frecuencias mientras que el máximo *SAD* para velocidades fue de 3.6 m s^{-1}.

Similitud entre grupos pertenecientes a la solución de 18 grupos y a la de 12 grupos
También aquí, por inspección visual (figuras V.29a y V. 29b) se revela una fuerte similitud entre las rosetas de viento, por ejemplo, las caracterizadas por el color violeta en ambas soluciones. Los valores de *SAD* entre las rosetas caracterizadas por violeta, rojo y amarillo oscuro entre la solución de 18 grupos y la de 12 grupos así como las caracterizadas por gris oscuro y amarillo están todos debajo del 10% para las frecuencias y debajo de 5 m s^{-1} para las velocidades. El resto de los cálculos entre todos los posibles pares dados por las dos soluciones dan valores de *SAD* superiores al 10%.
Estos resultados sugieren, a primera vista, que la solución de 18 grupos puede ser algo redundante. Para validar esta suposición se llevó a cabo el mismo tipo de análisis con las soluciones de 12 y 6 grupos y se compararon entre sí los grupos que forman la solución de 6 grupos.

Similitud entre grupos pertenecientes a la solución de 12 grupos y a la de 6 grupos
Solo dos grupos entre la solución de 12 y 6 (figuras V.29b y V.29c) grupos tuvieron *SAD* apenas menores al 10% en las frecuencias y diferencias poco significativas en las velocidades. La comparación entre el resto de las rosetas dio valores de *SAD* mayores al 10% y rápidamente crecientes.

Similitud entre los grupos correspondientes a la solución de 6 grupos
El menor de los *SAD* obtenidos para frecuencias fue de 17.2% mientras que de 1.9 m s^{-1} para las velocidades.

Por lo tanto, es posible concluir que las soluciones de 12 y 6 grupos son ambas apropiadas cuando un valor límite de *SAD* del 10% es puesto como referencia.

Similitud entre rosetas de viento observadas
Para apoyar los resultados obtenidos a partir del análisis por conglomerados basado en las salidas del modelo, se recurrió al cálculo del *SAD* entre rosetas observadas en las estaciones meteorológicas de la zona de estudio incluida PRE (Figura V.26). La Tabla V.7 muestra los valores de *SAD* obtenidos entre las estaciones ubicadas en la zona de Argentina continental.

Tabla V.7

Site	AER	EZE	LPA	PAL	TOR	SFO
AER		36,0	25,8	20,9	22,1	20,5
EZE			15,5	17,9	17,2	17,3
LPA				11,3	8,6	14,4
PAL					13,4	14,6
TOR						10,3

Tabla V.7: Valores de *SAD* para verano entre rosetas de direcciones de viento observadas en las distintas estaciones meteorológicas.

El *SAD* promedio entre estas estaciones es de 17.7%. La primera fila de la Tabla V.7 muestra que los mayores valores de *SAD* corresponden a la diferencia entre AER y el resto de las estaciones. El promedio de esta fila es del 25.1%. El máximo valor es con EZE. Excluyendo AER el *SAD* promedio baja al 14.1% bastante cercano al 10% impuesto como valor umbral para comparar entre los potenciales grupos provistos por el análisis por conglomerados a partir del modelo. Por lo tanto, excepto AER, el resto de las estaciones en la zona continental argentina muestra patrones de viento similares lo cual sustenta lo hallado mediante el análisis por conglomerados para las distintas soluciones que indica a

toda esta zona como una subárea. Dado que AER es una estación que está a solo algunos cientos de metros tierra adentro respecto de la costa del río y por lo tanto muy expuesta al contraste térmico tierra- agua, es razonable que refleje singularidades.

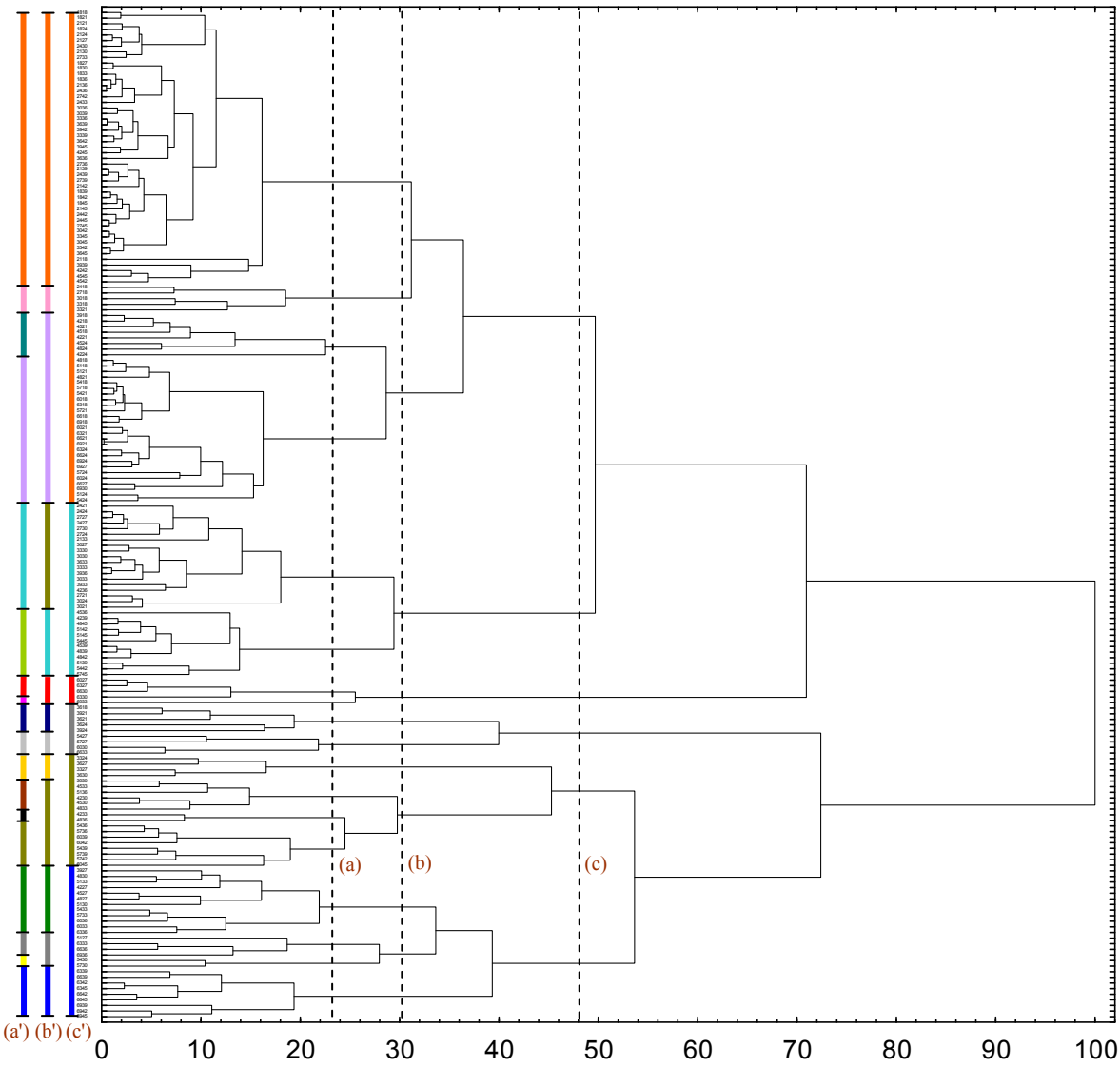
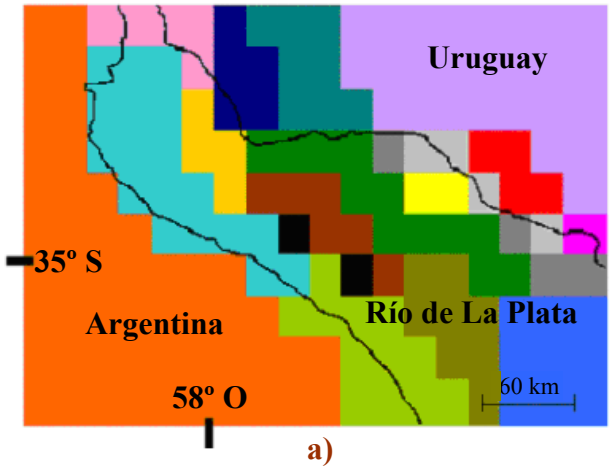
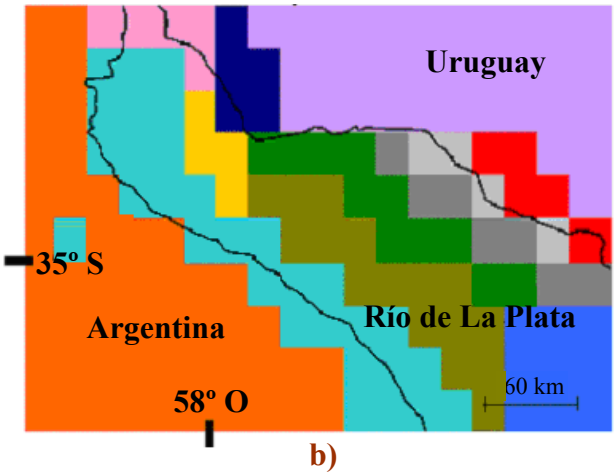


Figura V.27: Dendograma para el verano. La columna de números pequeños (solo legibles en formato digital) sobre el eje *Y* refiere a la identificación de cada uno de los 180 vectores en coordenadas arbitrarias, cada uno de ellos se corresponde con un pixel en la Figura V.28. El eje de las *X* representa a la distancia Euclídea al cuadrado que ha sido reescalada respecto de la máxima distancia por lo que aparece en %. Las tres distancias de corte seleccionadas (23, 30 y 48) se hallan identificadas con las líneas verticales a tramos. Para cada una de estas distancias (casos (a), (b) y (c) de la Figura V.28) cada grupo formado se halla identificado con un color según se muestra a la izquierda del eje *Y* ((a'), (b') y (c')).

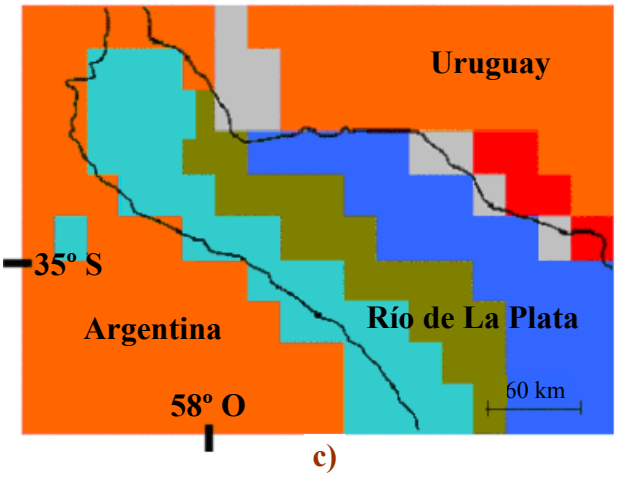
La Tabla V.8 muestra los valores de *SAD* entre dos estaciones continentales de Argentina (AER y EZE) y las estaciones ubicadas en el río y en la zona de Uruguay continental. El *SAD* promedio entre todas estas las estaciones es de 38.6%, más del doble que el de la Tabla V.7. Es de notarse que todas estas estaciones están ubicadas en diferentes subáreas de la Figura V.28b o c. El *SAD* promedio para PRE- COL- MGA- EZE es 32.4% mientras que el promedio PRE- COL- MGA- AER es de 43.9%. Cualquiera de estos valores es más del doble que el valor promedio de *SAD* para la zona de Argentina continental.



a)



b)



c)

Figura V.28: El rectángulo interior de la Figura V.25 se muestra dividido en:
a) 18 subáreas
b) 12 subáreas
c) 6 subáreas
Esta división se basa en las tres soluciones adoptadas en el proceso de análisis por conglomerados jerárquicos para la estación verano.

Cada uno de los 180 pixeles de la zona de estudio cubre un área aproximada de 22 (horizontal) x 28 (vertical) km x km. Estos píxeles algo rectangulares se aproximan a la forma que da el sistema de coordenadas gaussiano de la superficie terrestre en el rango de latitudes de trabajo.

Cada subárea (indicada con un color) reúne un número específico de pixeles según la distancia de corte; (a), (b) o (c) de la Figura V.27.

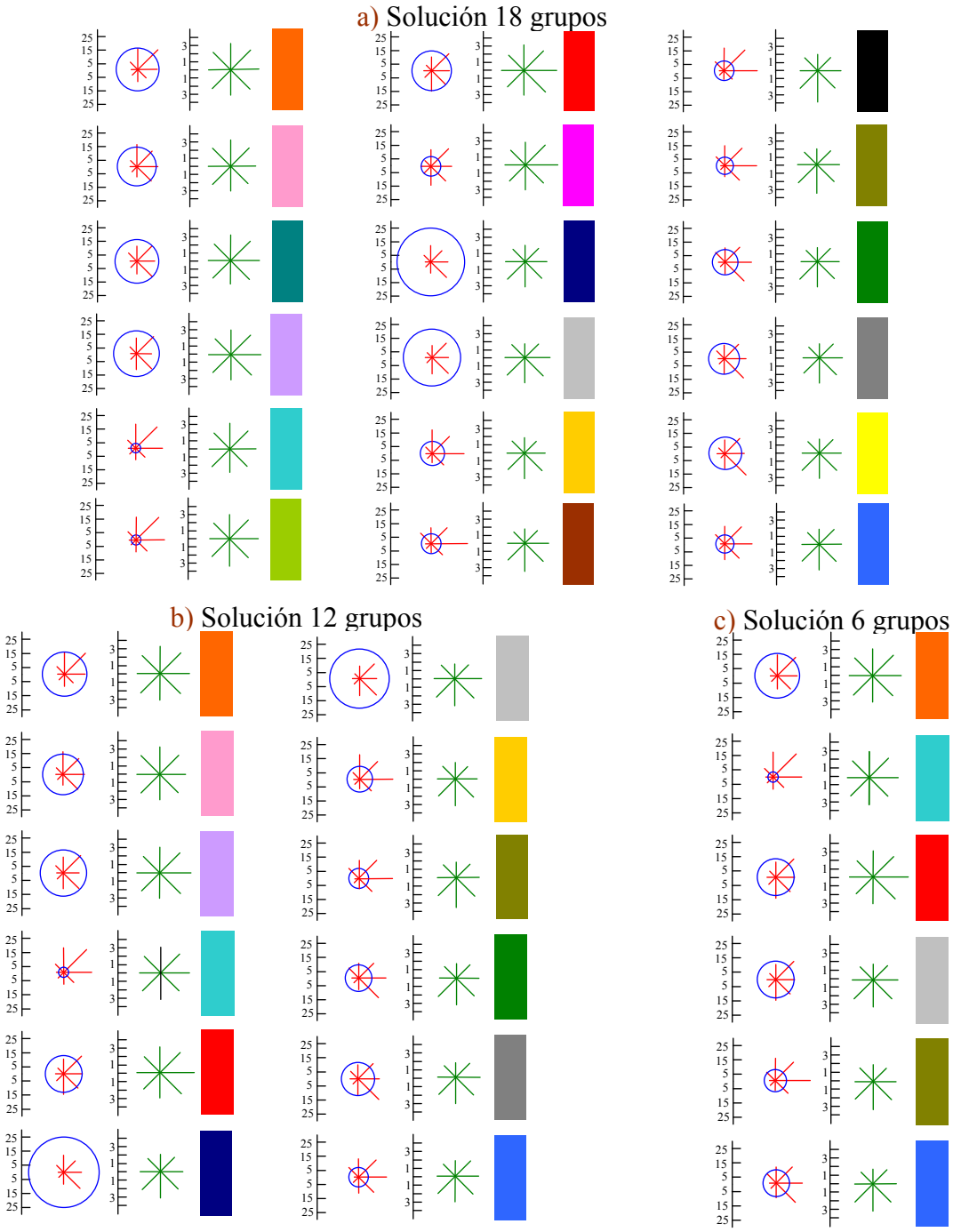


Figura V.29: Rosetas de viento resultantes (promedios) obtenidas para las tres soluciones adoptadas a partir del análisis por conglomerados. Las frecuencias de direcciones de viento incluyendo las calmas están dadas en porcentaje mientras que las velocidades medias de vientos están dadas en m s^{-1}. Lado izquierdo: rosetas de frecuencias de viento (líneas rojas) que incluyen las calmas (circunferencias azules), ambas expresadas en porcentaje de ocurrencias. El eje *Y* representa la frecuencia porcentual para las direcciones y las calmas. Lado derecho: rosetas de velocidades de viento (líneas verdes) expresadas en m s^{-1} (1 m s^{-1} equivale a 3.6 km h^{-1}) El eje *Y* representa la velocidad promedio para la dirección correspondiente. Cada roseta de vientos es el resultado de promediar los vectores correspondientes a las tres soluciones para el verano. Los rectángulos en color (este último asignado arbitrariamente) designan las subáreas que representan las rosetas involucradas en el mapa de la Figura V.28.
a) corresponde a 18 grupos (distancia de corte 23 en la Figura V.27) que se representan en la Figura V.28a.
b) corresponde a 12 grupos (distancia de corte 30 en la Figura V.27) que se representan en la Figura V.28b.
c) corresponde a 6 grupos (distancia de corte 48 en la Figura V.27) que se representan en la Figura V.28c.

Table V.8					
Sitio	**PRE**	**COL**	**MGA**	**AER**	**EZE**
PRE		24,8	45,0	32,0	15,9
COL			39,9	50,5	24,5
MGA				71,0	44,2

Tabla V.8: Valores de *SAD* para verano entre rosetas de direcciones de viento observadas en las distintas estaciones meteorológicas.

Por lo tanto, la comparación entre rosetas de viento observadas, pone en evidencia valores de *SAD* más grandes cuando las estaciones meteorológicas están ubicadas en distintas subáreas (definidas por el análisis por conglomerados) que cuando pertenecen solo a una subárea (Argentina continental). Por otra parte, los mayores valores de *SAD* se observan entre estaciones ubicadas en el río y en Uruguay continental, que es donde el análisis por conglomerados indica mayor concentración de subáreas (Figura V.28). Puede concluirse que las observaciones apoyan fuertemente las zonas definidas por el análisis por conglomerados a partir del modelo de mesoescala.

Aspectos climatológicos de los patrones espaciales

Más allá del grado de detalle que proveen las soluciones de 18, 12 y 6 grupos, la Figura V.28 permite apreciar que las tres comparten características similares. Las tres soluciones tienden a dar grupos a lo largo del río observándose más concentración de subáreas en las cercanías de la costa NE (Uruguay) que en la SE (Argentina). Esta disposición, dada por el agrupamiento de distintos puntos de la grilla del modelo, pone en evidencia los principales aspectos de los vientos de superficie en la zona del Río de La Plata que se caracteriza por la circulación de tipo brisa de mar- tierra. La ribera uruguaya sigue un patrón más accidentado cambiando el frente de la circulación a lo largo de la costa y dando lugar a más variación en los patrones de dirección de viento.

Como puede apreciarse en la Figura V.26a (Hora 6) las direcciones de viento en la zona uruguaya muestran un predominio del N y NE mientras que en Argentina los vientos se hallan distribuidos en varias direcciones siendo la N y la S algo más visibles. En la Figura V.26b (Hora 18) las estaciones de Uruguay muestran vientos predominantemente del E, SE y S mientras que en la Argentina los vientos son predominantemente del E.

En relación a la actual cantidad y distribución de estaciones meteorológicas el análisis por conglomerados da una cantidad y distribución similar a la existente aunque deja ver que se enriquecería la descripción de los vientos si hubiera más cantidad de estaciones sobre el río y sobre la costa uruguaya (Figura V.28b,c).

V.8.3 Homogeneidad de grupos de rosetas de viento utilizando Curvas de Andrews

En la Sección V.8.1 se adoptó, para los patrones horarios de dirección de viento, una distancia de corte tal que las rosetas estacionales quedaron representadas por 5 grupos, una posible solución dada por el análisis por conglomerados jerárquico.

En esta sección el objetivo es evaluar la homogeneidad de dichos grupos utilizando Curvas de Andrews (Andrews, 1972). Estas curvas permiten visualizar y explorar datos multidimensionales (en este caso rosetas de 16 direcciones) mediante gráficos en dos (o tres) dimensiones (Unwin, 2008; Moustafa, 2011) habilitando la posibilidad de detectar estructuras en los datos originales (García Osorio y Fyfe, 2005). La importancia de este método reside por un lado, en su simplicidad (resulta muy apropiado cuando la reducción de dimensionalidad aplicada a los datos originales proveen soluciones en más de tres dimensiones- casos en que los gráficos tradicionales se vuelven complejos de manejar- y por otro lado, dadas sus propiedades matemáticas, se hace posible relacionar los resultados

obtenidos con los de otros métodos (por ejemplo, análisis por conglomerados). Andrews mostró que la diferencia entre dos curvas dadas es proporcional a la distancia Euclídea, es decir, puntos cercanos en el espacio multidimensional serán evidenciados como Curvas de Andrews cercanas en el plano. Para más detalles ver la Sección 2 de Ratto et al. (2014b).

Cada punto del espacio *p*-dimensional $z = \{ z_1, z_2, ..., z_p \}$ define una función periódica dada por:

$$f(t) = \frac{z_1}{\sqrt{2}} + z_2 \ \sin(t) + z_3 \ \cos(t) + z_4 \ \sin(2t) + z_5 \ \cos(2t) + z_6 \ \sin(3t) + z_7 \ \cos(3t) ..... \text{(ec. V.4)}$$

que es llamada Curva de Andrews en la que *t* queda definido en el rango [*-180, 180*] dado en grados sexagesimales.

En el presente caso, la aplicación de Curvas de Andrews tiene un carácter cualitativo; las mismas son empleadas para asistir el análisis de la homogeneidad de los grupos, esto involucra (tal como lo hacen otros enfoques de minería visual de datos) el sistema de percepción visual del analista como parte del procesamiento de la información (Motta García et al., 2012).

Dado un conjunto de datos que forman grupos, las Curvas de Andrews darán a simple vista las características diferenciales de tales grupos permitiendo detectar patrones característicos. Si estas curvas fueran muy similares a lo largo de todos los grupos esto implicaría que los datos no tienen estructura fuerte de grupo.

Como puede apreciarse en la ecuación V.4, *f(t)* depende del orden asignado a las variables (z_i). La primera coordenada en la ecuación enfatiza frecuencias bajas que tenderán a dominar el gráfico (Carr, 1998). Sin embargo, esto no influye en la aplicación de estas curvas para detectar estructura de grupo o valores atípicos (Seber, 1984) porque cualquier orden escogido permitirá evaluar diferencias entre curvas (la información inherente es la misma). Gnanadesikan (1997) señala que, cuando no es posible asignarle a las variables distinto grado de importancia, puede recurrirse a comparar el resultado de varias permutaciones de ellas logrando así un mayor conocimiento de las características de los datos de trabajo.

En la presente aplicación se recurrió a seguir el "orden natural" dado por las componentes principales (CP) (Sección V.5), lo cual provee una solución al problema del orden de asignación de las variables (Spencer, 2003) y sienta bases para futuras comparaciones con otros conjuntos de datos. Además, el método de CP brinda una fundamentada reducción de dimensionalidad posible (Wilks, 2006), reteniendo una alta proporción de la varianza total lo cual, permite realizar el cálculo de las Curvas de Andrews con menos variables.

Existen dos posibles contratiempos al aplicar Curvas de Andrews a un conjunto dado de datos (Seber, 1984; Gnanadesikan, 1997; Chan, 2006): el tiempo de cómputo y el efecto de confusión dado por el ensimismamiento de las curvas. En el presente caso ninguno de los dos factores son importantes. El método de CP reduce la dimensionalidad de 16 a 5 (como se demostrará más adelante en esta misma sección) lo cual simplifica los cálculos de las Curvas de Andrews mientras que el tamaño de la muestra (24 vectores representados por 24 Curvas de Andrews) resulta fácil de manejar desde el punto de vista de la inspección visual.

Se toma como punto de partida el dendograma de la Figura V.21 y la solución de 5 grupos (cuyos promedios se hallan representados en la Figura V.22 (Sección V.8.1)).

Antes de representar las Curvas de Andrews a partir de las rosetas de viento de la Figura V.20 (Sección V.8.1) se recurrió al cálculo de las CP. Para ello se utilizó el software *Statistica 8.0* utilizándose la matriz de covarianzas en las variables. Con propósitos ilustrativos, se muestra la configuración de puntos correspondientes a las dos primeras CP obtenidas (Figura V.30), que explican el 86.7% de la varianza total (Figura V.31 y Tabla V.9).

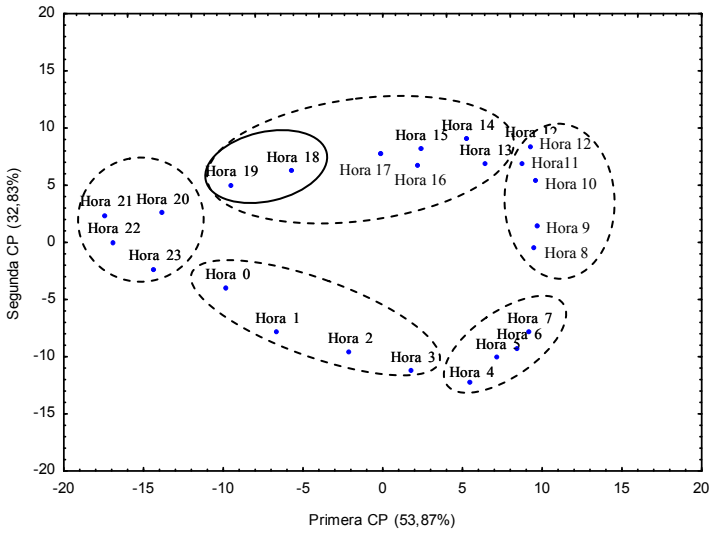


Figura V.30

Figura V.30: Los puntos en el plano representan a las rosetas de viento del dendograma de la Figura V.20 expresadas por las dos primeras componentes principales. Las líneas envolventes de trazos indican los grupos determinados por el dendograma para una distancia de corte de alrededor del 50%.
La línea continua que envuelve a las horas 18 y 19 indica un posible subgrupo. Ninguna de las líneas envolventes reflejan la forma de los grupos, han sido dibujadas solo con fines ilustrativos para mostrar la estructura de grupo.
Los valores sobre el eje de las X divididos por $\sqrt{2}$ constituyen el primer término en la ecuación V.4. y el valor constante para cada una de las curvas de la Figura V.32 desde la (a1) hasta la (e1).

En la Figura V.30 es posible apreciar la existencia de una estructura de grupos y la ausencia de valores atípicos (no conclusivo).

La Figura V.31 muestra el diagrama de sedimentación correspondiente a todas las componentes principales. Es posible apreciar que luego de los primeros 5 o 6 autovalores la curva se hace muy aplanada. Esto implica que con pocas componentes (nuevas variables) es posible representar a las 16 variables originales.

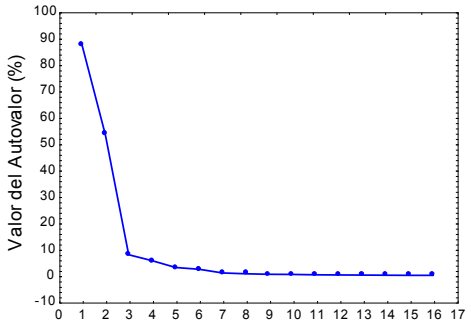


Figura V.31: Diagrama de sedimentación. Ayuda a determinar el número de autovalores a retener.

Tabla V.9

Número de Autovalor	Varianza (%)
1	53,87
2	86,70
3	91,52
4	94,95
5	96,80

Tabla V.9: Varianza acumulada según el número de autovalor.

Como puede apreciarse en la Tabla V.9, la varianza acumulada para las cinco primeras componentes es mayor al 95%, por lo que las Curvas de Andrews pueden ser graficadas solamente a partir de estas primeras cinco CP.

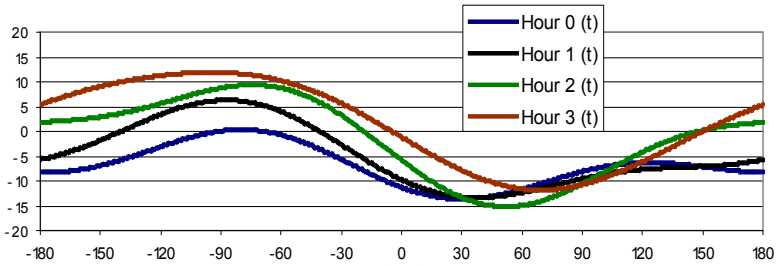
La Figura V.32 muestra las Curvas de Andrews correspondientes a cada roseta de la Figura V.20 para una distancia de corte de alrededor del 50% (5 grupos).
Una vista panorámica de la Figura V.32 de (a1) a (e1) permite determinar que existe buena homogeneidad en cada uno de los grupos, es decir, las curvas individuales tienden a estar cercanas unas a otras y sus formas son similares. En casi todos los grupos las curvas que pertenecen a los extremos del intervalo (por ejemplo, Hora 0 y Hora 3 en el Grupo 1, Hora 13 y Hora 19 en el Grupo 4, etc.) son las más diferentes entre sí (considerando distancia o forma). A través de los grupos la ocurrencia de picos y valles (que le dan identidad al grupo) son diferentes: la Figura V.32 de (a2) a (e2) – línea sólida- permite observar, en promedio, esta característica. En la misma figura las líneas de punto indican el promedio general del grupo que se corresponde con el primer término de la serie de Andrews (influenciado por la primera componente principal). Para el Grupo 1 este promedio es –2.8, para el Grupo 2 es 5, para el Grupo 3 es 6.6, para el Grupo 4 es 0.1 y para el Grupo 5 es –11. Esto significa que, en algunos casos, es posible distinguir importantes diferencias entre grupos (por ejemplo, entre el Grupo 3 y el 4 o entre el Grupo 4 y el 5) solamente a partir de la primera componente. Algunos grupos muestran pocas oscilaciones (por ejemplo, el Grupo 1) lo cual implica que hay más influencia de las funciones *sen(t) y cos(t)* (que se corresponden con la segunda y tercera CP) que de las funciones *sen(2t)* y *cos(2t)* (que se corresponden con la cuarta y quinta CP). Lo contrario ocurre con el Grupo 5 donde las funciones *sen(2t)* y *cos(2t)* asociadas a más oscilaciones son fáciles de notar.

En el Grupo 4 (Figura V.32d1) las curvas correspondientes a las Horas 18 y 19 aparecen como algo distintas al resto de las curvas del grupo, además no se ven similares a curvas de grupos vecinos (es decir, del Grupo 3 o del 5). Comparando curvas vecinas (Hora 17 con Hora 18 y Hora 19 con Hora 20) existen diferencias que no parecen ser muy fuertes. Calculando el promedio general para un potencial subgrupo Hora 13- Hora 17 (2.3) así como el promedio general para el otro potencial subgrupo Hora 18- Hora 19 (–5.3) es posible apreciar una diferencia relevante entre estos dos subgrupos.

En el Grupo 5 (Figura V.32e1) la curva para la Hora 23 muestra un patrón algo distinto del resto de los miembros del grupo. Comparando la Hora 23 con la Hora 0 (el vecino más cercano al Grupo 1) y con la Hora 22 (el vecino más cercano dentro del grupo) no es posible concluir que la Hora 23 constituya un individuo mal clasificado. A pesar de ser algo distinta al resto no es posible considerarla un atípico.

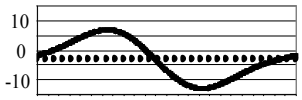
En síntesis, por un lado parece pertinente reagrupar a los miembros del Grupo 4 en dos nuevos grupos: Hora 13- Hora 17 y Hora 18- Hora 19. Esto se halla de acuerdo al dendograma (Figura V.21) para una distancia de corte de alrededor del 40%. Por lo tanto, a partir de Curvas de Andrews puede visualizase que las 24 rosetas de viento originales (verano en Punto J) parecen quedar mejor agrupadas en 6 grupos que en 5. Notar también que la estructura moderada de grupo que posee el verano en el Punto J (mostrada en la Sección V.5.5.5) se hace visible en la Figura V.30 si se comparan, por ejemplo, la distancia entre la Hora 18 y 19 y entre la Hora 19 y la 20.

Figura V.32a: Grupo 1



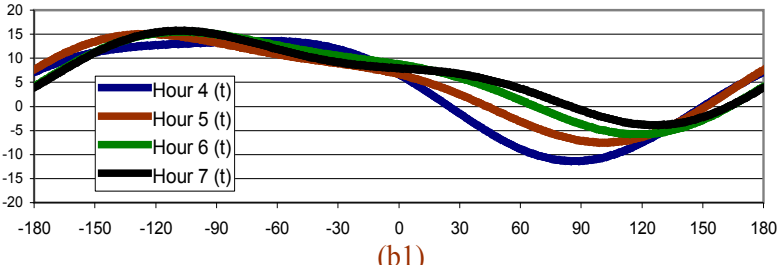
(a1)

(a1) Curva de Andrews individual
(a2) Curva promedio del grupo (línea sólida) y promedio general (línea de puntos)



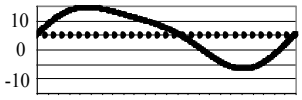
(a2)

Figura V.32b: Grupo 2



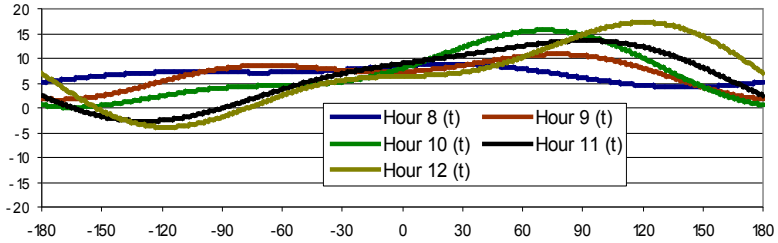
(b1)

(b1) Curva de Andrews individual
(b2) Curva promedio del grupo (línea sólida) y promedio general (línea de puntos)



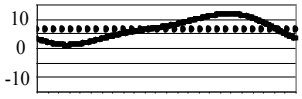
(b2)

Figura V.32c: Grupo 3



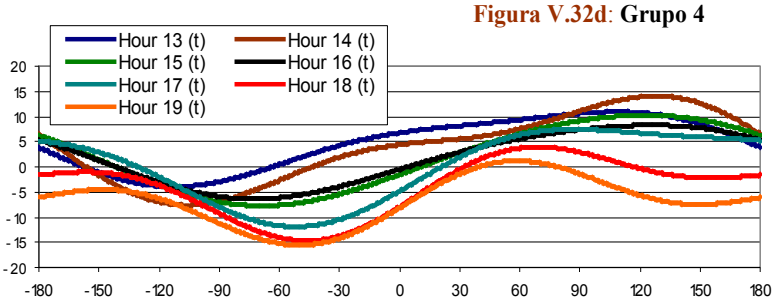
(c1)

(c1) Curva de Andrews individual
(c2) Curva promedio del grupo (línea sólida) y promedio general (línea de puntos)



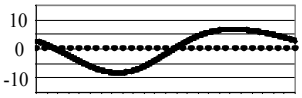
(c2)

Figura V.32d: Grupo 4



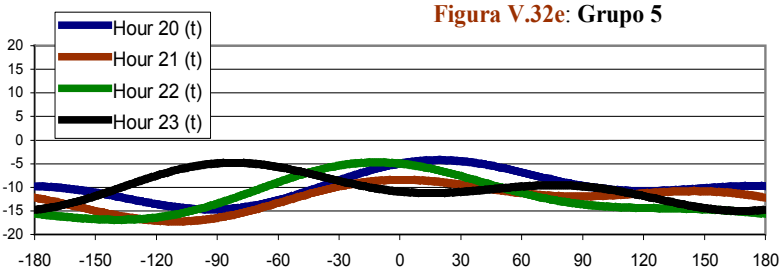
(d1)

(d1) Curva de Andrews individual
(d2) Curva promedio del grupo (línea sólida) y promedio general (línea de puntos)



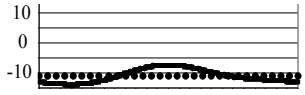
(d2)

Figura V.32e: Grupo 5



(e1)

(e1) Curva de Andrews individual
(e2) Curva promedio del grupo (línea sólida) y promedio general (línea de puntos)



(e2)

Figura V.32: Curvas de Andrews para las rosetas horarias de la Figura V.20. Cada curva fue construida a partir de las primeras cinco componentes principales empleadas como variables en la ecuación V.4. El eje *X* cubre el intervalo *t* [-180, 180]. El eje *Y* corresponde a *f(t)* (ver ecuación V.4).

Implicaciones meteorológicas

La Figura V.33 (columna de la derecha) muestra las rosetas de viento de los nuevos grupos formados (Grupo 4* y Grupo 5*) que modifica a la Figura V.22 (columna de la izquierda) como resultado de la aplicación de las Curvas de Andrews. El asterisco (*) indica que el grupo es nuevo respecto de la clasificación dada en la Figura V.22.

El Grupo 6 en la columna de la derecha es el Grupo 5 de la Figura V.22 que ha sido renombrado con fines prácticos.

Notar que la obtención de seis grupos es coincidente con lo hallado en la Sección V.5.5.5 mediante otros indicadores.

Una de las ventajas de la Figura V.33 (columna derecha) comparada con la Figura V.22 es que permite apreciar un cambio más gradual de los vientos dominantes desde el mediodía hasta el atardecer. El Grupo 5* muestra el decrecimiento del viento N y la importancia de los vientos del E hacia el atardecer (efecto que no era captado por el Grupo 4 surgido un agrupamiento de las 24 rosetas originales en 5 grupos).

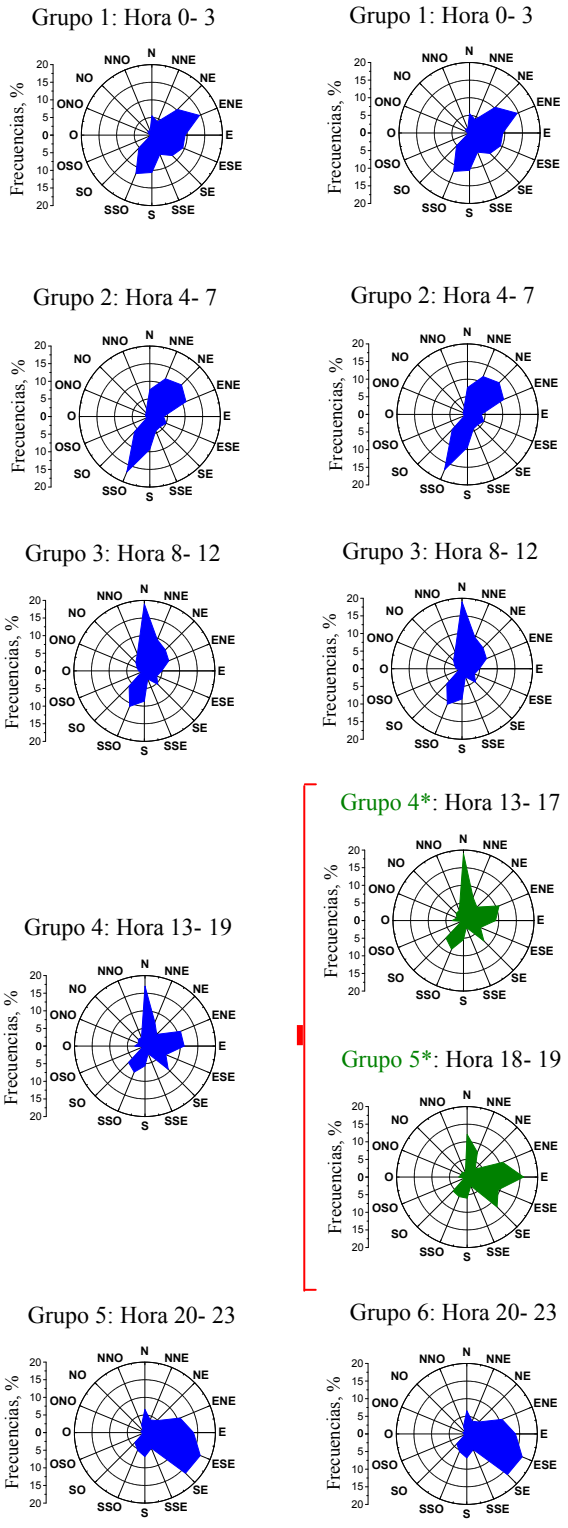


Figura V.33: La columna izquierda de esta figura repite la configuración de rosetas de la Figura V.22. La columna derecha introduce los nuevos grupos hallados en concordancia con el dendograma de la Figura V.21 para una distancia de corte de 40%. El Grupo 4 de la Figura V.22 ha dado lugar al Grupo 4* y Grupo 5*.

V.8.4 Encontrar grupos teniendo en cuenta restricciones

En algunos casos de aplicación, cuando el objetivo es encontrar grupos en un conjunto dado de datos, existe información externa que debe ser tenida en cuenta. La necesidad de que los individuos dentro de un grupo tengan contigüidad espacial o consecutividad temporal representan casos típicos de restricciones. Gordon (1999) y Everitt et al. (2011) presentan de manera sintética la problemática involucrada mientras que Basu et al. (2009) compilan una vasta cantidad de métodos y ejemplos de análisis por conglomerados que requieren el empleo de información adicional. Como se señaló en la Sección V.8.1 las rosetas horarias de viento que forman un grupo deben ser consecutivas (al menos en las cercanías de la distancia de corte adoptada) puesto que de lo contrario habrá pérdida de interpretabilidad debilitándose así las conclusiones.

Un ejemplo de formación de grupos con miembros discontinuos puede apreciarse en el dendograma correspondiente a la primavera en el Punto J tal como lo es para distancias de alrededor del 50 % (ver Figura V.34). Como puede observarse, para una distancia de corte correspondiente a 6 grupos existe un grupo que contiene a las Horas 9, 10 y 12 y otro grupo que contiene a las Horas 11, 13, 14, 15, 16, 17 y 18 (ver recuadros sobre el eje *Y* izquierdo de la figura). Esta situación comienza a una distancia de alrededor de 35% (9 grupos) y se soluciona para una distancia de alrededor de 54% (que corresponde a la solución de 5 grupos adoptada en Ratto et al. (2010b)).

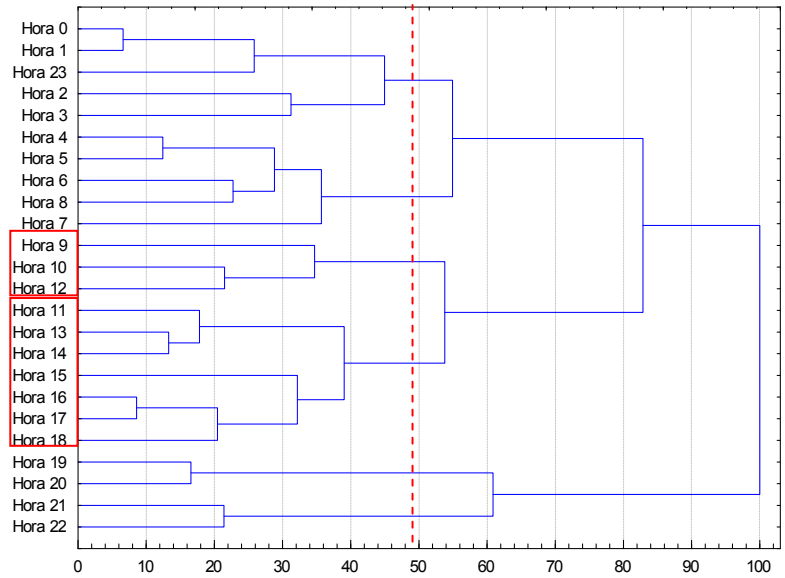


Figura V.34

Figura V.34: Dendograma correspondiente a rosetas de frecuencias horarias de vientos por dirección de la primavera en el Punto J durante el período 1998- 2003. El eje *X* son distancias Euclídeas al cuadrado reescaladas. El dendograma fue obtenido normalizando los datos con media y desvío estándar. La distancia Euclídea al cuadrado es la medida de disimilitud adoptada y el criterio de aglomeración es el UPGMA.

Supongamos que dado un dendograma único o un conjunto de ellos (tal como el caso de las estaciones del año) sea conveniente adoptar un número de grupos determinado y que esa elección implique la detección de discontinuidades. En estos casos, es deseable que exista algún tipo de transformación de los datos, de tal manera que los grupos tengan miembros consecutivos.

En el Anexo V.5 (pág. 185) se describe, a modo de ejemplo, un enfoque sencillo propuesto por Maronna (CP) que arroja luz sobre el significado de trabajar con restricciones.

V.8.5 Siluetas

Como se expresó en la Sección V.1, el empleo del análisis por conglomerados jerárquicos se basa en el desconocimiento *a priori* del número de grupos que pueda contener un conjunto dado. Una vez elegida la distancia de corte en el dendograma quedan definidos los grupos y el promedio o centroide se adopta típicamente como "representante" de cada grupo (Kaufman y Rousseeuw, 2005; Mirkin, 2005). Dicho representante no es ningún miembro "real" del grupo.

Existen otras formas de encontrar un representante de grupo en el que uno de los vectores originales es designado como tal por sus características, o sea, constituye un prototipo (Mirkin, 2005). Rousseeuw (1987) presenta un método que sirve para la interpretación y validación de los grupos que no depende del algoritmo utilizado para hallarlos. El diagrama de siluetas involucra un recurso gráfico que permite establecer la fortaleza de la membresía de cada individuo en su grupo (pertenencia), informa sobre la relación de cada miembro de un grupo con respecto a grupos vecinos, establece el representante de cada grupo y da una idea sobre el nivel de estructuración de los datos originales (criterio de realidad de los grupos encontrados- Sección V.5.6).

Sea *i* un objeto (individuo, miembro, vector) cualquiera del conjunto original de datos que está constituidos por *k* grupos (A, B, C, etc.).
Supongamos que *i* fue asignado al grupo A (donde hay otros objetos) entonces es posible definir a *a(i)* como la disimilitud (por ejemplo, la distancia Euclídea) promedio del individuo *i* en relación a todos los miembros de su grupo (grupo A).

Consideremos un grupo C (≠A) y definamos *d(i, C)* como la disimilitud promedio entre *i* (de A) y todos los objetos del grupo C. Esto puede hacerse para todos los grupos existentes distintos de A. Entonces se define *b(i)* como el valor mínimo de las disimilitudes encontradas entre *i* y todos los miembros de los grupos (≠A) o sea:

$$b(i) = mín\ d(i,C) \qquad \text{con } C \neq A$$

Al grupo que cumple con la condición de dicho mínimo es llamado el "vecino" del objeto *i* y será la segunda mejor elección de membresía.
Entonces mientras *a(i)* da idea de la cohesión que tiene el objeto *i* respecto del grupo al cual pertenece (a menor *a(i)* mayor es la cohesión interna), *b(i)* da una idea del grado de aislamiento que tiene el objeto *i* con respecto a los grupos a los cuales no pertenece.

Combinando *a(i)* y *b(i)* se define (Rousseeuw, 1987):

$$s(i) = \frac{b(i) - a(i)}{máx\ \left[a(i), b(i)\right]} \qquad -1 \leq s(i) \leq 1$$

Cuando un grupo contiene un solo individuo *s(i)* se define como cero (neutralidad).
Si *s(i)* tiene valores altos (cercanos a 1) implica que la disimilitud dentro del grupo dada por *a(i)* es baja con respecto al mínimo de disimilitud con otros grupos (dado por *b(i)*). En este caso *a(i)* muestra buena cohesión interna y *b(i)* buena separación y, por lo tanto, el objeto o vector *i* se halla bien asignado como miembro de un grupo. Cuando *s(i)* adopta valores cercanos a cero, *a(i)* y *b(i)* adoptan valores similares y no resulta clara la membresía del objeto *i* (si pertenece por ejemplo al grupo A o al B). Cuando *s(i)* adopta valores negativos da indicios de que el objeto *i* ha sido mal clasificado, cosa que se incrementa cuanto más cercano a –1 esté. En este sentido el diagrama de Siluetas puede

utilizarse para proponer "mejoras" en una clasificación hallada. En los casos en que gran parte de los *s(i)* son bajos, implicará que no hay una estructura de grupos en el conjunto de datos analizados o que la misma es muy débil.
En síntesis, *s(i)* mide cuan bien clasificado se halla el objeto *i* en un grupo y cuan distinto es de los demás grupos. El objeto de mayor *s(i)* es denominado representante de dicho grupo.

Cada grupo es representado por una "silueta" que es un tipo de gráfico que permite visualizar un perfil (Gordon, 1999), cada silueta queda definida por los valores de *s(i)* de los miembros del grupo. La representación de todos los grupos permite ver la calidad de la aglomeración lograda. El *s(i)* promedio de un grupo se llama ancho de la silueta. El promedio general de los *s(i)*, llamado $s_{prom}(k)$, da una idea de la bondad de la estructura de grupos encontrados.

En caso de trabajar con algún método de partición (tal como el de las *k*-medias o el PAM "partición alrededor de medioides", etc.) se puede efectuar un barrido desde k=2 hasta k=n-1 con el método de las siluetas y determinar el máximo de los $s_{prom}(k)$ (llamado coeficiente de silueta, *SK*); será posible así determinar el número óptimo de grupos.
Kaufman y Rousseeuw (2005) dan una tabla orientativa para interpretar el $s_{prom}(k)$ hallado para un caso particular.

Tabla V.10

SK	Interpretación sobre la estructura de grupos
0.71- 1.00	fuerte
0.51- 0.70	razonable
0.26- 0.50	débil, puede ser artificial (probar otros métodos)
≤ 0.25	no hay estructura sustancial

Tabla V.10: Coeficientes de Siluetas. Tomada del Capítulo 2 de Kaufman y Rousseeuw (2005).

A continuación y a modo de ejemplo, se muestra el dendograma para rosetas horarias anuales de direcciones de viento publicado en Ratto et al. (2010a) (Figura V.35) y el diagrama de Siluetas correspondiente para la solución adoptada de 8 grupos (Figura V.36).

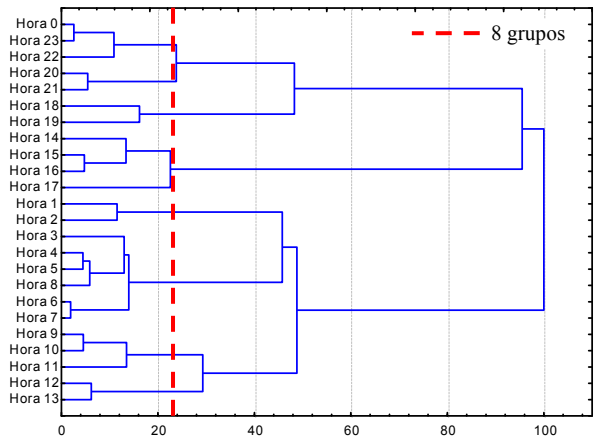


Figura V.35

Figura V.35: Dendograma correspondiente a rosetas de frecuencias horarias anuales de vientos por dirección observadas en el Punto A durante el período 1997- 2000.
El eje *X* son distancias Euclídeas al cuadrado reescaladas. El dendograma fue obtenido normalizando los datos con media y desvío estándar. La distancia Euclídea al cuadrado es la medida de disimilitud adoptada y el criterio de aglomeración es el UPGMA.
Se indica, en línea cortada, la solución adoptada en la publicación de referencia para una distancia de corte de aprox. 24% (solución para 8 grupos).

Esta última figura (gráfico de las siluetas) muestra que los grupos poseen, en general, una buena calidad de membresía (todos excepto uno son valores positivos). El $s_{prom}(k)$ da 0.528 que según la Tabla V.10 se corresponde con una estructura inherente razonable en los datos. Existen cuatro individuos que poseen valores de *s(i)* menores que 0.25. El caso

extremo es el de la Hora 14 que aparece como mal clasificada (*s(i)*= - 0.257) en el grupo 5. El resto de los individuos corresponden a las Horas 11, 19 y 22 y presentan poca fortaleza de membresía (grupos 3, 6 y 8 respectivamente).

Si, a la luz de los resultados provistos por la Figura V.36, se quisiera ver como se pueden agrupar "mejor" los vectores horarios, podría recurrirse (puesto que se ha definido el número de grupos en 8 en base al dendograma) a un criterio no jerárquico como el bien conocido método de las *k*- medias (Ver Anexo V.6, pág. 188).

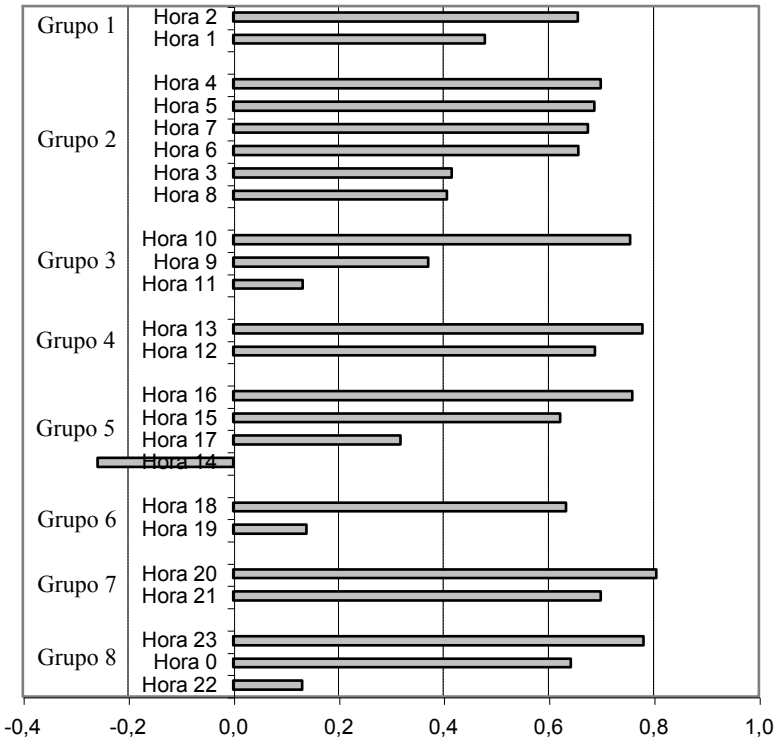


Figura V.36: El eje de las *X* son las *s(i)* para cada uno de los vectores originales pertenecientes a un grupo. El eje *Y* representa las rosetas horarias y los grupos formados según el dendograma de la Figura V.35 para una distancia de corte de 24%.

Los representantes de grupo son:
Grupo 1: Hora 2
Grupo 2: Hora 4
Grupo 3: Hora 10
Grupo 4: Hora 13
Grupo 5: Hora 16
Grupo 6: Hora 18
Grupo 7: Hora 20
Grupo 8: Hora 23

Figura V.36

La aplicación de este método da como resultado los siguientes grupos (Tabla V.11):

Tabla V.11

Grupo	**Miembros**
Grupo 1	Hora 1- Hora 2
Grupo 2	Hora 3- Hora 8
Grupo 3	Hora 9- Hora 11
Grupo 4	Hora 12- Hora 13
Grupo 5	Hora 14- Hora 16
Grupo 6	Hora 17- Hora 18
Grupo 7	Hora 19- Hora 21
Grupo 8	Hora 22- Hora 0

Tabla V.11: Grupos obtenidos mediante el método de las *k*-medias (utilizando el software *Statistica 8.0*).

La Tabla V.11 refleja que, respecto de los grupos dados por el dendograma (Figura V.35), se aprecian algunas modificaciones en los integrantes que forman los grupos 5, 6 y 7 mientras que el resto permanece igual. Con el fin de poner en evidencia si el arreglo de grupos dado por el método de las *k*-medias es más adecuado que aquel obtenido recurriendo exclusivamente al dendograma, se procedió a aplicar nuevamente el diagrama de las Siluetas (Figura V.37):

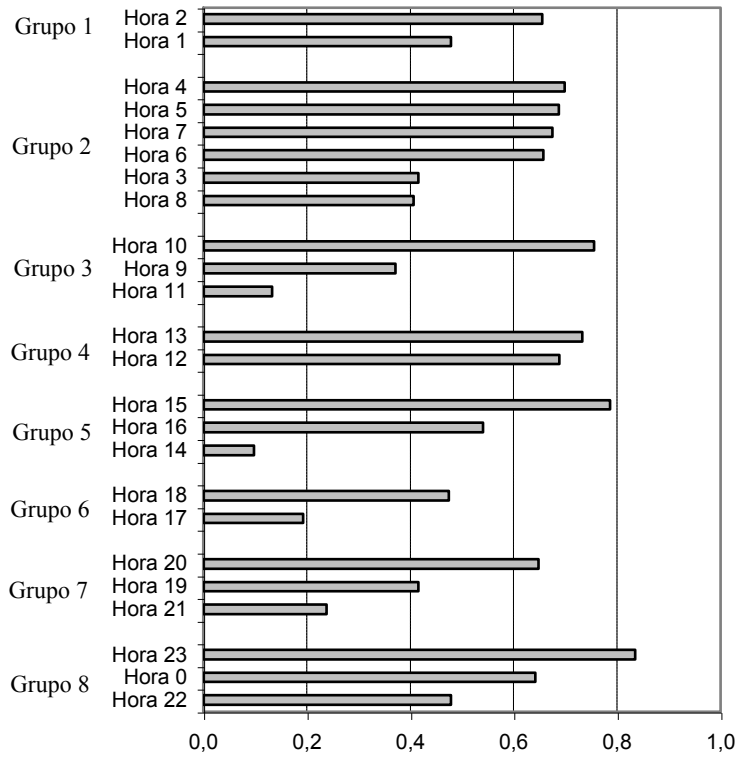


Figura V.37

Figura V.37: El eje de las *X* son las *s(i)* para cada uno de los vectores originales pertenecientes a un grupo. El eje *Y* representa las rosetas horarias y los grupos formados según el método de las *k*- medias aplicado a los datos de trabajo de la Figura V.35.

Los representantes de grupo son:
Grupo 1: Hora 2
Grupo 2: Hora 4
Grupo 3: Hora 10
Grupo 4: Hora 13
Grupo 5: Hora 15
Grupo 6: Hora 18
Grupo 7: Hora 20
Grupo 8: Hora 23

La Figura V.37 muestra que los grupos poseen, en general, buena calidad de membresía. Existen cuatro individuos que poseen valores de *s(i)* menores que 0.25 y son Hora 21, 17, 11 y 14, sin embargo, todos los integrantes poseen valores positivos evidenciando que no quedan individuos mal clasificados. Por lo tanto, ha sido posible mostrar de una manera sencilla, que la aplicación del método de las *k*- medias puede mejorar el arreglo inicial dado por el procedimiento de aglomeración jerárquica. Este recurso suele utilizarse con distinto grado de sofisticación (Kaufmann y Whiteman, 1999) para producir grupos más homogéneos; en el caso aquí presentado el diagrama de las Siluetas sirve para visualizar la mejora realizada.

Por otra parte, la obtención de representantes "reales" de grupo resulta potencialmente interesante para los trabajos de campo dada la existencia de horas específicas del día que definen el alcance de mediciones realizadas. Por ejemplo, llevar a cabo mediciones de alguna especie contaminante a la Hora 4 (ver el valor relativo de *s(i)* en la Figura V.37) tendrá vientos representativos de una franja de horas (Hora 3- Hora 8) en donde se esperan vientos muy similares (ver siluetas del Grupo 2 en la Figura V.37).

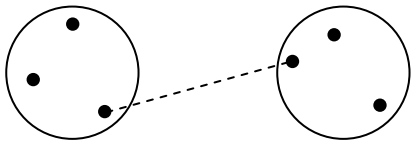
Anexo V.1

Criterios de agrupamiento (discusión)

Enlace Simple (también llamado Encadenamiento Simple, de la Distancia Mínima, del Vecino más próximo o criterio del Mínimo)

$$d_{g_1,g_2} = mín\ (d_{rs}) \text{ donde } r \in g_1 \text{ y } s \in g_2$$

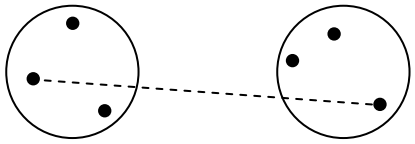
Aquí la distancia entre el grupo 1 y el grupo 2 se adopta como la menor de las distancias que existe entre cada objeto del grupo 1 y cada objeto del grupo 2



Enlace Completo (también llamado Encadenamiento Completo, de la Distancia Máxima, del Vecino más lejano o criterio del Máximo)

$$d_{g_1,g_2} = máx\ (d_{rs}) \text{ donde } r \in g_1 \text{ y } s \in g_2$$

Aquí la distancia entre el grupo 1 y el grupo 2 se adopta como la mayor de las distancias que existe entre cada objeto del grupo 1 y cada objeto del grupo 2.

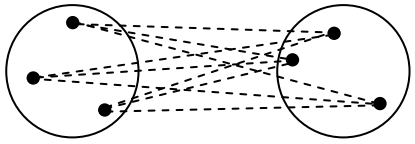


Enlace Promedio (también llamado de Distancia Promedio, Promedio no ponderado- **UPGMA** -unweighted pair-group method using arithmetic averages -, Promedio entre grupos)

$$d_{g_1,g_2} = \frac{1}{n_r\ n_s} \sum_{1}^{n_r} \sum_{1}^{n_s} d_{rs} \text{ donde } r \in g_1 \text{ y } s \in g_2,$$

n_r es el número de objetos en g_1, n_s es el número de objetos en g_2

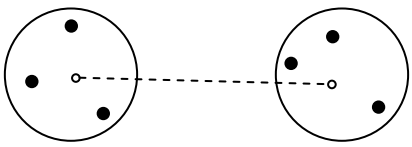
Aquí la distancia entre el grupo 1 y el grupo 2 se adopta como el promedio de todas las distancias entre pares de objetos del grupo 1 y del grupo 2.



Enlace Centroide (UPGMC- unweighted pair-group method using the centroid approach)

$d_{g_1,g_2} = \left(\sum_{1}^{p} (\bar{x}_{g_1} - \bar{x}_{g_2})^2 \right)^{1/2}$ donde $\bar{x}_{g_1}$ y $\bar{x}_{g_2}$ son el centro geométrico (centroide) de cada grupo

Aquí la distancia entre el grupo 1 y el grupo 2 se adopta como la distancia entre los vectores *p*-dimensionales promedio (centroides) de cada grupo.



Los enlaces “Promedio” y “Centroide” tienen sus contrapartes “pesadas” por el número de miembros del grupo. Suelen llamarse "Enlace Promedio Ponderado (WPGMA- weighted

pair-group method using arithmetic average)" y "Enlace Mediana (o del centroide pesado o WPGMC- weighted pair-group method using the centroid approach)" respectivamente.

Regla de Ward: este criterio es algo distinto a los anteriores puesto que no opera sobre la matriz de distancias (sino sobre la de datos). Se realiza una suma de cuadrados (*SC*) entre objetos o grupos y se elige luego la menor *SC* para una determinada instancia de aglomeración.

$SC = \sum_{i=1}^{g_1+g_2} \sum_{1}^{p} (x_{i,p} - \bar{x}_p)^2$ donde $x_{i,p}$ representa a la variable del objeto *i* de dimensión *p*

$\bar{x}_p$ es la media en la variable *p* de todos los objetos para los que se calcula la SC

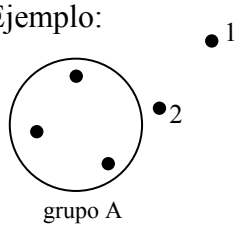
i designa el número de objeto de *p* variables

g_1 es el número de objetos en el grupo 1 (es un único objeto en el paso inicial)

g_2 es el número de objetos en el grupo 2 (puede ser un único objeto)

En el paso inicial de aglomeración (todos los objetos forman *n* grupos de un individuo) *SC*=0. En el paso final (todos los objetos iniciales forman un solo grupo), *SC* tiene un valor máximo. En cada paso se reúnen todos los posibles objetos de a pares (individuos o grupos según corresponda) y se calculan los *SC*. Se elige el individuo o grupo (que agregado al existente) produzca el menor *SC*.

Ejemplo:



Se unen los miembros de A con el objeto 1 y se calcula SC_{A1}. De forma análoga se calcula SC_{A2}. Puesto que $SC_{A2} < SC_{A1}$ entonces corresponde que el punto 2 pase a formar parte del grupo A, formándose así un grupo de cuatro miembros.

Antes de realizar algunos comentarios sobre los criterios arriba presentados conviene mostrar -a manera de ejemplo- como en un conjunto de datos se pueden distinguir subgrupos con criterios muy distintos, o dicho de otra manera, como la definición de un criterio puede afectar la búsqueda de grupos. El conjunto de puntos mostrados en la Figura 1a puede ser subdividido según se muestra en la Figura 1b o como en la Figura 1c.

Lorr (1983) nombra a estas dos principales posibilidades como grupos serpentinos (Figura 1b) o grupos compactos (Figura 1c).

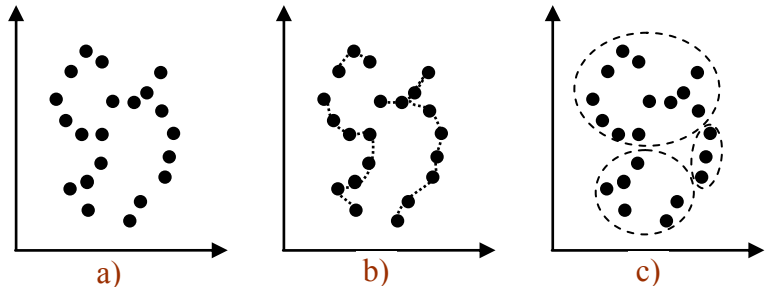


Figura 1: Conjunto de datos y dos posibles formas de agrupamiento.
a) puntos en el plano b) agrupamiento elongado c) agrupamiento esferoide

Con este ejemplo se quiere enfatizar el hecho de que cuando se trata de puntos multidimensionales (*p*>3) en donde la representación no es simple las suposiciones sobre la "forma" de los agrupamientos (que en general se desconoce *a priori*) queda implícita en el criterio que se seleccione. Este hecho también explica la no existencia de un "criterio

óptimo" universal (Timm, 2002) y que la aplicación de distintos criterios puede conducir a resultados muy distintos (Maronna, CP).

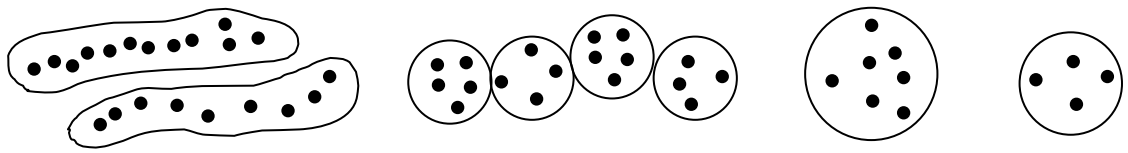
El criterio del enlace simple fue introducido por Sneath en 1957. Este criterio tiende a unir grupos aún cuando un solo punto de uno de los grupos se halle cerca del otro grupo. Además tenderá a producir uniones entre un grupo y un punto individual cercano y luego, añadirá otro punto individual al grupo formado tendiendo a producir un "efecto en cadena", o sea, puede evidenciarse una tendencia a unir puntos a grupos más que a realizar uniones entre grupos; al menos, en una parte considerable del proceso de aglomeración. Esta tendencia puede también hacer que puntos muy lejanos de un grupo queden formando el nuevo grupo produciendo así una "contracción" del espacio de referencia (Legendre y Legendre, 1998) (se brinda un ejemplo comparativo en el Anexo V.3, pág. 183). De proseguir este efecto en cadena el criterio dará lugar a la formación de grupos "elongados". Sin embargo, se debe tener presente que hay casos en que se busca detectar la presencia de configuraciones alargadas (porque se conoce de antemano o se suponen) y entonces este criterio resultará apropiado. Cabe agregar que Jardine y Sibson (1968) señalan que el encadenamiento no debe ser considerado un "defecto" del enlace simple (puesto que como cita Timm (2002) es una característica intrínseca de los métodos jerárquicos), sino un criterio que lo pone más en evidencia.

El criterio del enlace completo fue propuesto por Sorensen en 1948. Es el "opuesto" del anterior en cuanto a que tiende a "dilatar" el espacio de referencia produciendo grupos compactos pero que se enciman entre sí y los puntos alejados (potenciales valores atípicos) solo pasarán a formar parte de los grupos en las etapas finales del proceso de aglomeración.

Tanto el criterio del enlace simple como el del enlace completo son invariantes ante transformaciones monótonas de escala (tales como las estandarizaciones más utilizadas) y satisfacen la condición ultramétrica (Sección V.3, pág. 113).

El criterio del enlace promedio (UPGMA) fue presentado por Sokal y Michener en 1958. Es un promedio no ponderado en el sentido de que todos los objetos reciben igual peso (o ponderación). Este criterio tiende a dar grupos esferoides, es relativamente robusto a la presencia de valores atípicos y es "conservativo" del espacio (Rencher, 2002).

Estos tres criterios descriptos son útiles en diferentes tipos de aplicaciones (Kaufman y Rousseeuw, 2005) y sus tendencias para identificar grupos pueden ejemplificarse en el siguiente gráfico tomado de Kaufman y Rousseeuw (2005):



formas de agrupamiento "esperables" con el enlace simple

formas de agrupamiento "esperables" con el enlace completo

formas de agrupamiento "esperables" con el enlace promedio

El enlace centroide (UPGMC) fue presentado por Sokal y Michener en 1958 y, como se describió en el principio de este anexo, es de fácil interpretación. El enlace mediana (WPGMC) presentado por Gower en 1967 es igual al anterior solo que se ponderan los grupos por su tamaño con la finalidad de darle igual importancia que al grupo en

formación. Estos dos últimos tipos de enlaces están sujetos a reversiones debido a que no cumplen con la propiedad ultramétrica (Sección V.3, pág. 113). La regla de Ward fue introducida por J. H. Ward en 1963. Este criterio, también llamado de la varianza mínima (Wilks, 2006), es sensible a la presencia de valores atípicos (Milligan, 1980) y tiende a dar grupos hiperesféricos de igual tamaño.

La adopción de distintas medidas de similitud o disimilitud dará, en general, distintos resultados para un mismo conjunto de datos (Timm, 2002; Everitt et al., 2011). Sin embargo, y a pesar de que hay muchos estudios que discuten la performance de los distintos tipos de proximidades, no es posible llegar a una conclusión general y la elección queda en su mayor parte dependiendo del tipo de variables involucradas y del criterio del investigador (Baxter, 1994). Cunningham y Ogilvie (1972) en un estudio de patrones en el plano, dan al enlace promedio "UPGMA" como el que mejor responde al evaluar conglomerados jerárquicos comparando medidas de bondad de ajuste. Sneath y Sokal (1973) y Maronna (CP) recomiendan elegir la medida más sencilla de tal manera que sea la de más fácil interpretación.

Cabe agregar que cada uno de estos criterios (excepto la regla de Ward) pueden expresarse en términos de combinaciones lineales de distancias entre individuos (combinatoriedad) para cada paso de aglomeración (esto se trata en detalle en el Capítulo 7 de Gan et al. (2007)). Existe para la mayoría de los criterios una forma generalizada (por Lance y Williams) de los mismos a partir de una fórmula recurrente (Capítulo 4 de Everitt et al. (2011)).

Anexo V.2
Método de las Componentes Principales

El presente anexo describe de manera sintética y sin rigurosidad matemática el Método de Componentes Principales en su versión "clásica".

Parte a: Planteo del Problema

El investigador se halla frecuentemente frente a un conjunto de datos que pueden agruparse en una matriz de datos.

Sea una matriz X (*n-objetos* x *p-variables*) donde las filas representan a los objetos y las columnas a las variables continuas aleatorias de una muestra de un sistema multivariado:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & \dots \\ \dots & \dots & x_{ij} & \dots \\ x_{n1} & .. & \dots & x_{np} \end{pmatrix} \quad \text{donde} \quad i=1,n \quad \text{y} \quad j=1,p$$

Es posible representar a ese conjunto inicial de datos como un conjunto de vectores fila $X_i = (x_{i1}, x_{i2}, x_{i3} \dots . x_{ip})$ donde cada X_i (vector de objetos) representa a un objeto dado de la matriz X.

$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ es el vector columna (vector de variables) de la matriz X para cada j desde $j=1$ hasta p y contiene los valores de una variable determinada *(j)* en cada uno de los objetos ($i=1,n$). Si se toma la primera variable, o sea, $j=1$ y se promedian los elementos de X_j desde $i=1$ hasta n se obtendrá el valor $\bar{x}_{1j} = \dfrac{\sum_{i=1}^{n} x_{i1}}{n}$, continuando este proceso se obtiene el vector de medias en las variables dado por:

$\bar{X}_j = \begin{pmatrix} \bar{x}_{1j} \\ \bar{x}_{2j} \\ \dots \\ \bar{x}_{nj} \end{pmatrix}$ de *(1 x p)* con $j=1, p$

Por otro lado, la matriz de datos X tendrá asociada una matriz de covarianzas *(p x p)* en las variables dada por:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1p} \\ \Sigma_{21} & \dots & \dots & \dots \\ \dots & \dots & \Sigma_{ij} & \dots \\ \Sigma_{p1} & .. & \dots & \Sigma_{pp} \end{pmatrix}$$ donde Σ_{ij} = Varianza si $i=j$, o sea $\Sigma_{ii} = \dfrac{\sum_{i=1}^{p} (x_i - \bar{x}_i)^2}{n-1}$

y Σ_{ij} = Covarianza si $i \neq j$, o sea $\Sigma_{ij} = \dfrac{\sum_{i=1}^{p}(x_i - \bar{x}_i)((x_j - \bar{x}_j)}{n-1}$ con $i=1,p$ y $j=1,p$.

Σ es una matriz cuadrada con rango completo p y simétrica respecto de la diagonal principal (varianzas).

Parte b: Autovalores y Autovectores

Cualquier matriz cuadrada (p x p) puede expresarse en función de escalares llamados autovalores (valores propios o valores característicos de la matriz) y vectores llamados autovectores (vectores propios o vectores característicos de la matriz) que son de p x 1 no nulos. Tomando como referencia la matriz de covarianzas resulta:

$$\boldsymbol{\Sigma Y} = \lambda \boldsymbol{Y} \qquad \text{ec. 1a}$$

que puede ser expresado como

$$(\boldsymbol{\Sigma} - \lambda \boldsymbol{I})\boldsymbol{Y} = 0 \qquad \text{ec. 1b}$$

donde $\boldsymbol{Y}$ es el autovector $\boldsymbol{Y} = \begin{pmatrix} a_1 \\ \dots \\ a_p \end{pmatrix}$ y λ es el autovalor asociado, $\boldsymbol{I}$ es la matriz identidad.

Siendo $\boldsymbol{Y}$ no nulo queda:

$$|\boldsymbol{\Sigma} - \lambda \boldsymbol{I}| = 0 \qquad \text{ec. 2a}$$

$$\boldsymbol{Y} = 0 \qquad \text{ec. 2b}$$

En relación a la ec.2a se formará un sistema de p ecuaciones con p incógnitas que podrán expresarse como un polinomio de grado p donde las c_i son constantes dadas por la combinación del sistema de ecuaciones, o sea:

$$c_1 \lambda^p + c_2 \lambda^{p-1} + c_3 \lambda^{p-2} + \dots\dots + c_p \lambda + c_{p-1} = 0$$

cuyas raíces son los autovalores (λ) de $\boldsymbol{\Sigma}$. Cada autovalor tendrá asociado un autovector $\boldsymbol{Y}$ que satisface la ec. 1b.

Una propiedad es que:

$$\sum_{J=1}^{p} \lambda_j = tr(\Sigma)$$

o sea, que la sumatoria de los autovalores encontrados a partir del polinomio será igual a la traza de la matriz covarianza que por definición de traza: $tr(\Sigma) = \sum_{J=1}^{p} \Sigma_{ii}$. Esta propiedad de los autovalores es importante, ya que cuando se calculan a partir de la matriz de covarianza, la suma de los autovalores es igual a la suma de las varianzas de las variables incluidas en la matriz, o sea dicha suma da la variación total.

Si se vuelve al sistema de p ecuaciones con p incógnitas mencionado (basado en la ec.1) para cada valor de λ hallado se obtendrá un sistema compatible de p ecuaciones con p incógnitas pero indeterminado (infinitas soluciones). Esto se resuelve imponiendo la condición de módulo unitario a cada autovector tal que $\boldsymbol{Y'Y}=1$. Por lo tanto, la solución se hace determinada y para cada valor de λ habrá un autovector $\boldsymbol{Y}$.

Para hacer más tangible lo descripto en esta sección se recurrirá a un ejemplo tomando como punto de partida una matriz de covarianzas de dos dimensiones (2x2).

Sea $\Sigma = \begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix}$ entonces la ec. 1b podrá escribirse como

$$\left[\begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right] x \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 6-\lambda & 3 \\ 3 & 4-\lambda \end{pmatrix} x \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

efectuando el producto de estas dos matrices se obtiene un sistema de dos ecuaciones con dos incógnitas:

$(6- \lambda)\ a_1 + 3\ a_2 = 0$ ec. 3

$3\ a_1 + (4- \lambda)\ a_2 = 0$ ec. 4

Sustituyendo se llega a que

$(4- \lambda)\ (6- \lambda) = 9$ que puede expresarse como

$\lambda^2 - 10\lambda + 15=0$ (polinomio de grado *p* con raíces reales en λ.

La solución a este polinomio son dos autovalores:

$\lambda_1 = 8.16$

$\lambda_2 = 1.84$

Si con λ_1 vamos a la ec. 3 tendremos:

$-2.16\ a_1 + 3\ a_2 = 0$

$3\ a_1 - 4.16\ a_2 = 0$

sumando en ambos miembros

$0.84\ a_1 - 1.16\ a_2 = 0$ por lo que se tiene que

$a_1 = 1.38\ a_2$ ec. 5

O sea que las ecuaciones 3 y 4 forman un sistema compatible indeterminado. Para que el mismo tenga solución única se adopta la restricción de que el módulo del autovector que se quiere determinar sea unitario ($\boldsymbol{Y'Y}=1$) que se puede expresar como $a_1^2 + a_2^2 = 1$ entonces $a_2 = (1- a_1^2)^{1/2}$ que reemplazando en la ec. 5 se obtienen $a_1 = 0.81$ y $a_2 = 0.59$ (soluciones positivas) que son los elementos del autovector $Y = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.81 \\ 0.59 \end{pmatrix}$ cuyo autovalor es $\lambda_1 = 8.16$. Es posible apreciar que $(0.81)^2 + (0.59)^2 = 1$.

De la forma análoga se obtiene el autovector para λ_2.

<u>Parte c</u>: Obtención de las Componentes Principales

Operando sobre el conjunto original de *n* objetos de la matriz *X* de *p* variables es posible describir a ese sistema multivariado con nuevas variables Z_k con $k=1, p$ tal que estas nuevas variables sean combinaciones lineales de las variables originales (X_j). Si estas nuevas variables cumplen con el requisito de estar incorrelacionadas entre sí y dan cuenta de gran parte de la variabilidad (varianza) del sistema con pocas de ellas, entonces estas nuevas variables se denominan componentes principales (CPs).

No es necesario que el conjunto multivariado siga una distribución conocida pero, si la muestra se comporta como multinormal, las CPs obtenidas tendrán la característica de ser independientes.

La primera CP puede expresarse como:

$$Z_1 = a_{1j} X_j \qquad j=1,p$$

donde a_{11}, a_{12}, a_{13}, ...a_{1p} son los pesos o cargas ("loadings") que son los escalares que forman el autovector. Estas cargas permiten transformar a las variables originales X_j en la nueva variable Z_1. Habrá tantas Z_i como variables en el sistema original.
Se debe buscar el vector a_{1j} que maximice la varianza de Z_1. Es posible demostrar que dicho vector es el autovector que corresponde al autovalor de mayor valor (surge de la ec.1a) sujeto a la restricción de módulo unitario. Por lo tanto,

$$\text{Var}\,(Z_1) = \lambda_1$$

Geométricamente, el primer autovector indica la dirección en que los datos exhiben de manera conjunta la mayor variabilidad. El conjunto de los autovectores definen un nuevo sistema de coordenadas en los que pueden ser vistos los datos.

Para determinar la segunda CP se debe cumplir que

$$Z_2 = a_{2j} X_j \qquad j=1,p$$

que deberá cumplir con la condición de estar incorrelacionada con Z_1 (o sea, la Cov (Z_1,Z_2) =0) y tener máxima varianza después de la Var(Z_1). Siguiendo el mismo razonamiento que en el caso anterior se llega a que

$$\text{Var}\,(Z_2) = \lambda_2$$

finalmente se llega a que $Z_j = \sum_{j=1}^{p} a_{ij}\ X_j$

Desarrollando las ecuaciones se tiene que:
Primera CP : $Z_1 = a_{11} X_1 + a_{12} X_2 + ... + a_{1p} X_p$
Segunda CP: $Z_2 = a_{21} X_1 + a_{22} X_2 + ... + a_{2p} X_p$
Tercera CP : $Z_3 = a_{31} X_1 + a_{32} X_2 + ... + a_{3p} X_p$
Etc.

Parte d: Representación de los Objetos en función de las CP

Cada objeto de la matriz original puede quedar representado en función de las CP. El valor particular que adopta una CP para un objeto cualquiera de la matriz original se llama marcador ("score"). Si un objeto se quiere representar, por ejemplo, en sus dos primeras componentes principales tendrá las coordenadas z_1 y z_2 (un punto en el plano de las dos primeras componentes principales), o sea, el objeto queda definido por los marcadores correspondientes.

Parte e: Misceláneas

1) Una vez calculadas las componentes principales es deseable conocer (para el conjunto de datos de aplicación) que porcentaje de varianza total original se explica con las nuevas variables.
Por ejemplo, %Var (explicado por Zj) = $\lambda_j / \Sigma\, \lambda_j$
Existen en la literatura muchos criterios (Reinmann et al., 2008) para determinar cuantas componentes principales explican bien al sistema original y los distintos criterios llevan a resultados distintos. El más difundido es tomar un número de CP tales que expliquen un valor alto de varianza, por ejemplo un 80%.

2) En la mayor parte de las aplicaciones se resta el vector de medias *Xj* en la matriz original de datos para evitar que la primera componente tenga valores distorsionados.

3) Es posible trabajar con la matriz de correlaciones ***R*** en lugar de la matriz ***Σ***. Los resultados son en general distintos (Jolliffe, 2002) y la elección queda librada a criterio del investigador; Jollife (2002), Rencher (2002) y Varmuza y Filzmoser (2009) dan lineamientos al respecto. Sin embargo, cuando las variables originales son de la misma naturaleza no conviene trabajar con correlaciones puesto que este procedimiento tenderá a equiparar artificialmente a todas las variables distorsionando los valores de las primeras componentes principales.

Se han descripto hasta aquí lineamientos generales del proceso de las CP. Cabe agregar que se han propuesto mejoras al enfoque mostrado: por ejemplo, "robustizando" el método de tal manera que sea menos vulnerable a la presencia de valores atípicos (Maronna et al., 2006). El software *Scout 1.0* brinda la posibilidad de calcular CP robustas. También, existen propuestas que tienen en cuenta las no linealidades de la matriz original de datos.

Anexo V.3
Coeficiente cofenético y esquema de aglomeración

El coeficiente cofenético que se adopta típicamente es el de correlación de Pearson (ρ) que ha sido renombrado en el campo de las ciencias biológicas (Sokal y Rohlf, 1962) y su uso se ha generalizado con ese nombre (la palabra cofenético viene de "co": estar con y "fenético": taxonomía basada en similitudes y diferencias medibles). También suelen utilizarse estimadores no paramétricos de correlación mencionados en la Sección V.5.6.2 (pág. 134). Todos estos estimadores dan una idea de la consistencia interna del proceso de aglomeración (Chagoyen et al., 2006) pero también pueden ser utilizados para comparar dendogramas de los mismos datos (Brunet et al., 2004) utilizando distintos criterios de aglomeración o evaluar dendogramas en distintos niveles de aglomeración (Legendre y Legendre, 1998). Los coeficientes no paramétricos hacen más incapié en la estructura geométrica de las matrices que se comparan y no tanto en el ajuste paso a paso de los mismas. Seber (1984) recomienda la aplicación de un coeficiente que correlacione rangos cuando el criterio de agrupamiento es el enlace simple o el completo puesto que ambos son invariantes a transformaciones monótonas de cambios de escala (por ejemplo, cuando se aplica logaritmo para reescalar). Cunningham y Ogilvie (1972) emplearon un índice de estrés basado en distancias para evaluar jerarquías: $E_i = \sum (d_{ij} - d_{ij}^*)^2 / \sum d_{ij}^2$

donde d_{ij} es la distancia Euclídea entre los elementos i y j de la matríz de distancias original entre pares de datos (n (n-1)/2) y d_{ij}* es la distancia vía el dendograma (distancia cofenética). En su estudio, que abarca la mayoría de los criterios más utilizados de aglomeración, estos autores mostraron que el E_i dio resultados muy similares al índice τ de Kendall. Sin embargo, el principal hallazgo de su estudio es la influencia que tienen sobre estos índices (τ y E_i) la dependencia de la interacción entre la configuración de los datos (estructuras simuladas) y el criterio de aglomeración empleado.

Sea un conjunto n de individuos de p- dimensiones, se toman de a pares y se calcula alguna medida de disimilitud (por ejemplo, distancia Euclídea al cuadrado) entonces se podrá formar una matriz simétrica con ceros en la diagonal principal (matriz original o de entrada) que tendrá [n (n-1)]/2 elementos por debajo de la diagonal principal.
Por otra parte, y una vez obtenido el dendograma correspondiente o cualquier otro esquema de aglomeración (previamente se ha elegido una medida y un criterio de aglomeración) es posible obtener las disimilitudes (en nuestro ejemplo, distancia Euclídea al cuadrado) entre todos los pares de individuos pero "vía el dendograma" o sea, la matriz cofenética (con distancias al cuadrado). El coeficiente cofenético dará una idea de la distorsión que el proceso de aglomeración produce en los datos.
En la Figura 1 (idéntica a la Figura V.14 del cuerpo del Capítulo V) se muestra un dendograma sobre el que se trabajará para ilustrar el uso del coeficiente cofenético. El eje de las X no se halla reescalado con la finalidad de guardar correspondencia con las distancias mostradas en el esquema de aglomeración.
En la Tabla 1 se muestra el esquema de aglomeración de los 24 vectores conforme avanza el proceso. El "paso" (primera columna de la tabla) se refiere al nivel o instancia de aglomeración. La D_{Enlace} (segunda columna) es la distancia Euclídea al cuadrado calculada con el criterio del promedio entre grupos (UPGMA). La tercera columna muestra la composición del grupo que se arma por fusión de individuos o grupos en un determinado paso de aglomeración. La última columna indica la instancia en donde un grupo ya formado se fusiona para formar un grupo mayor.

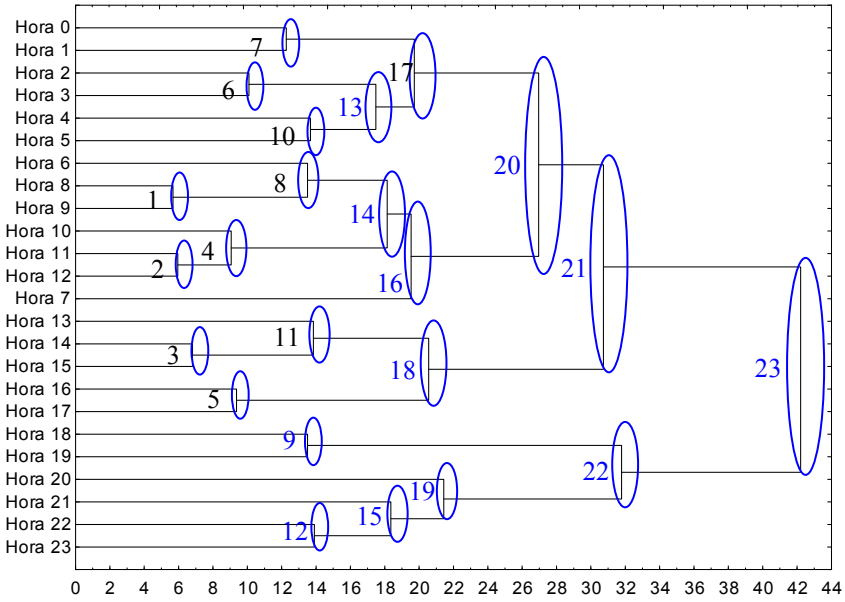


Figura 1

Figura 1: Dendograma de 24 rosetas horarias promedio de vientos correspondiente al invierno en el Punto J para el período 1998- 2003.
El eje de las *Y* cada "Hora" representa un vector de 16 direcciones de frecuencia de vientos.
En el eje de las *X* se halla representada la distancia Euclídea al cuadrado. Los óvalos y sus números indican el paso de aglomeración según el esquema de la Tabla 1.

Como se indicó en la Sección V.5.6.2 (pág. 134) el cálculo del coeficiente cofenético implica la determinación de la matriz original de distancias (Figura 2a) y la matriz cofenética (Figura 2b). Puesto que para el caso de estudio ambas son de 24x24, por razones de espacio, se mostrará una parte de ellas con el objetivo de hacer más tangible el procedimiento de cálculo del coeficiente.

Tabla 1

Paso	D_{Enlace}	Miembros del Grupo	Próximo Paso
1	5,673	Hora 8- Hora 9	8
2	5,957	Hora 11- Hora 12	4
3	6,812	Hora 14- Hora 15	11
4	9,061	Hora 10- Hora 11- Hora 12	14
5	9,376	Hora 16- Hora 17	18
6	10,094	Hora 2- Hora 3	13
7	12,274	Hora 0- Hora 1	17
8	13,507	Hora 6- Hora 8- Hora 9	14
9	13,508	Hora 18- Hora 19	22
10	13,677	Hora 4- Hora 5	13
11	13,845	Hora 13- Hora 14- Hora 15	18
12	13,904	Hora 22- Hora 23	15
13	17,482	Hora 2- Hora 3- Hora 4- Hora 5	17
14	18,141	Hora 6- Hora 8- Hora 9- Hora 10- Hora 11- Hora 12	16
15	18,351	Hora 21- Hora 22- Hora 23	19
16	19,536	Hora 7- Hora 6- Hora 8- Hora 9- Hora 10- Hora 11- Hora 12	20
17	19,720	Hora 0- Hora 1- Hora 2- Hora 3- Hora 4- Hora 5	20
18	20,555	Hora 13- Hora 14- Hora 15- Hora 16- Hora 17	21
19	21,437	Hora 20- Hora 21- Hora 22- Hora 23	22
20	26,961	Hora 0- Hora 1- Hora 2- Hora 3- Hora 4- Hora 5- Hora 7- Hora 6- Hora 8- Hora 9- Hora 10- Hora 11- Hora 12	21
21	30,726	Hora 0- Hora 1- Hora 2- Hora 3- Hora 4- Hora 5- Hora 7- Hora 6- Hora 8- Hora 9- Hora 10- Hora 11- Hora 12- Hora 13- Hora 14- Hora 15- Hora 16- Hora 17	23
22	31,786	Hora 18- Hora 19- Hora 20- Hora 21- Hora 22- Hora 23	23
23	42,209	Todos los objetos iniciales quedan fusionados en un solo grupo	

Tabla 1: Esquema de aglomeración obtenido con el software *SPSS Versión 13.0* correspondiente al dendograma de la Figura V.14. Ejemplo: para una distancia aproximada de 5.7 en el dendograma (óvalo con el número 1) se forma el primer grupo (Hora 8- Hora 9) tal como lo indica la presente tabla en el paso 1.

	Hora 0	Hora 1	Hora 2	Hora 3	Hora 4	Hora 5	Hora 6	Hora...
Hora 0	0							
Hora 1	12.3	0						
Hora 2	21.4	14.4	0					
Hora 3	27.3	15.8	10.1	0				
Hora 4	22.6	12.8	17.6	12.9	0			
Hora 5	24.2	19.3	22.9	16.5	13.7	0		
Hora 6	33.4	30.7	40.3	25.5	27.6	31.7	0	
Hora...	…	…	…	…	…	…	…	…

a)

	Hora 0	Hora 1	Hora 2	Hora 3	Hora 4	Hora 5	Hora 6	Hora...
Hora 0	0							
Hora 1	12.27	0						
Hora 2	19.72	19.72	0					
Hora 3	19.72	19.72	10.09	0				
Hora 4	19.72	19.72	17.48	17.48	0			
Hora 5	19.72	19.72	17.48	17.48	13.67	0		
Hora 6	26.29	26.29	26.29	26.29	26.29	26.29	0	
Hora...	…	…	…	…	…	…	…	…

b)

Figura 2

Figura 2: Matrices de distancias involucradas en el cálculo del coeficiente cofenético.

a)Fracción de la matriz original de distancias (matriz de un modo). Esta matriz muestra las distancias Euclídeas al cuadrado entre pares de objetos al inicio del procedimiento cuando no se han formado grupos.

b) Fracción de la matriz cofenética que resulta de todo el proceso de aglomeración. Esta matriz muestra las distancias Euclídeas al cuadrado (Enlace Promedio) entre pares de objetos (individuos o grupos) "vía el dendograma", o sea, cuando todos los objetos han sido agrupados.

El coeficiente cofenético se calcula como el ρ de Pearson entre ambas matrices.

La Figura 3 se obtuvo graficando el conjunto de pares de puntos que relacionan ambas matrices en un diagrama de dispersión utilizando el criterio del Enlace Promedio. Este gráfico se denomina, según Legendre y Legendre (1998), diagrama tipo- Shephard por analogía con los diagramas distancia- distancia empleados en el método de escalamiento multidimensional no métrico propuesto por Shepard en 1962.

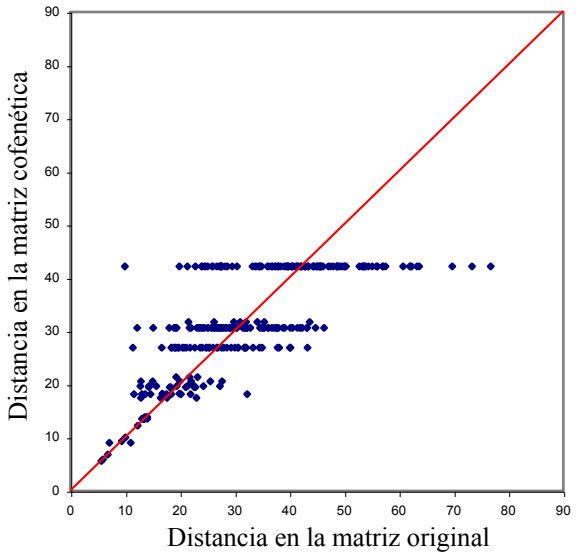


Figura.3

Figura 3: Diagrama tipo- Shephard con enlace promedio.
En el eje de las *X* han sido graficadas las distancias Euclídeas al cuadrado de la matriz original. En el eje de las *Y* las correspondientes distancias "vía el dendograma". La repetición de valores en el eje de las *Y* se debe a que la matriz cofenética limita su número de valores a *n-1* tal como lo muestra la Tabla 1 mientras que las distancias en la matriz original son de $n(n-1)/2$. Ocurre que para algunos valores distintos de la matriz original existe un solo valor correspondiente en la matriz cofenética. La línea a 45° ha sido trazada como referencia.

El valor del coeficiente cofenético encontrado para este caso es de 0.728, valor que resulta aceptable (Sección V.5.6.2, pág. 134).

Si se realizan los dendogramas para el criterio del Enlace Simple y el Enlace Completo (Sección V.4, pág. 115 y Anexo V.1, pág. 171) junto a los coeficientes cofenéticos calculados utilizando los estimadores de Pearson y de Spearman se obtiene la Tabla 2.

Tabla 2

	Enlace Simple	Enlace Completo	Enlace Promedio
Pearson (ρ)	0,580	0,701	0,728
Spearman (Sr)	0,588	0,725	0,734

Tabla 2: Coeficientes de correlación de Pearson y Spearman para tres criterios de enlace.

Esta tabla permite apreciar que el mejor ajuste lo da el Enlace Promedio siendo el Enlace Simple el que más "distorsiona". Por otra parte *Sr* da siempre algo superior debido a su mayor robustez.

La Figura 4 es el diagrama tipo- Shephard cuando el dendograma se lleva a cabo con el Enlace Simple. Es apreciable como las distancias vía el dendograma (eje *Y*) son más pequeñas que las distancias en el eje *X* (dadas por la matriz original de distancias). De esta manera queda ejemplificado el efecto de "contracción del espacio" que produce este criterio que incorpora al vecino más cercano (Anexo V.1, pág. 174). El efecto contrario es apreciable en la Figura 5. Siguiendo la misma perspectiva la Figura 3 representa un criterio de "conservación del espacio".

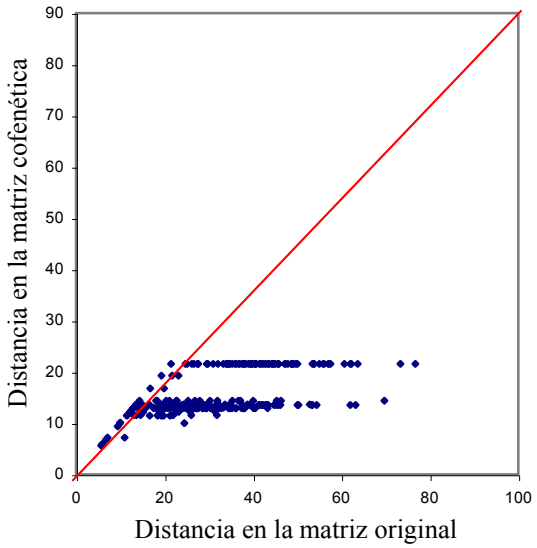


Figura 4: Diagrama tipo- Shepard utilizando el criterio del Enlace Simple ("single linkage"). La recta trazada a 45 grados ha sido trazada como referencia permite evidenciar la "contracción" del espacio inducida por este tipo de criterio.

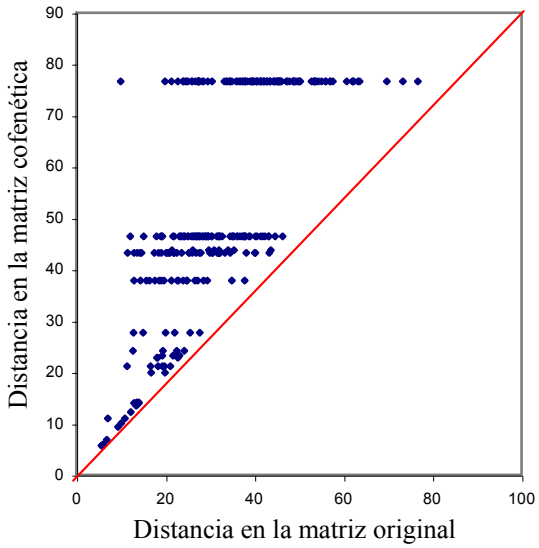


Figura 5: Diagrama tipo- Shepard utilizando el criterio del Enlace Completo ("complete linkage"). La recta trazada a 45 grados ha sido trazada como referencia permite evidencia la "expansión" del espacio inducida por este tipo de criterio.

Anexo V.4
Secuencia de pasos para el cálculo de una configuración de EMD

A continuación se presenta una secuencia simplificada de pasos posibles para obtener una configuración con estrés mínimo.

1) Se calcula la matriz de disimilitud $\mathbf{\Delta}_{n\ x\ n}$ en el espacio *p*-dimensional de las *n* observaciones.

2) Se define una configuración inicial de *n* puntos en la dimensión *k* (típicamente *k*=2)

3) Se normaliza (por ejemplo, con media y desvío estándar) la matriz de puntos definida en el paso 2)

4) Se calcula $\mathbf{D}_{n\ x\ n}$ (matriz de distancias Euclídeas entre puntos de la configuración)

5) Se ordenan los elementos de **Δ** (paso 1) en orden ascendente (o descendente)

6) Se ordenan los elementos de **D** siguiendo el mismo orden de la matriz **Δ**

7) Se calcular la matriz de disparidades $\hat{\mathbf{D}}$ que tendrá elementos $\hat{d}$ que serán el resultado de reemplazar algunos elementos de **D** que no satisfagan la condición de monotonicidad (tal como el punto 4 de la Figura V. 16 (Sección V.6.1, pág. 140)).

8) Se calcula *S* (ec. V.2 Capítulo V) y se lo lleva a porcentaje (Sección V.6.1, pág. 142).

9) Puesto que *S* será al principio alto (esto dependerá de lo cercana que esté la configuración inicial a la final) se deberá cambiar la configuración inicial. Para esto se recurre a un algoritmo específico (Sección V.6.1, pág. 140).

10) Se siguen los pasos 3 a 10 hasta que, por ejemplo, dos valores consecutivos de STRESS no difieran significativamente.

Anexo V.5
Encontrar grupos con restricciones
(enfoque)

Maronna (CP) propone, a modo de ejemplo, un enfoque sencillo que permite al investigador tomar contacto con la problemática de encontrar grupos con restricciones.

El objetivo general es afectar la matriz inicial de distancias de tal manera que las distancias más bajas correspondan a individuos más cercanos entre si (consecutivos) y las más altas a aquellos pares de individuos que se hallan opuestos entre sí. Esta "configuración" propuesta es debida al carácter cíclico de los vectores considerados (Sección V.8.1, pág. 145).

Sea la matriz de datos (de dos modos) M, en el caso de estudio $M^{Primavera\ J}_{n\ x\ p}$ (Sección V.8.4, pág. 166) donde n es el número de datos y p es el número de variables. Se normalizan los datos con algún criterio (en este caso con promedio y desvío estándar (Ratto et al., 2010b)). Luego se calcula la matriz (de un modo) D' de distancias Euclídeas al cuadrado entre objetos $D'^{Primavera J}_{n\ x\ n}$ (siguiendo la selección hecha previamente). A esta matriz se la divide por el máximo de sus elementos de tal manera de obtener una matriz de distancia escalada $D^{escal}{}^{Primavera\ J}_{n\ x\ n}$ (o simplemente D^{escal}) con valores entre 0 y 1.

Por otro lado, se representan los números del 0 al 23 en un círculo ("reloj" de la Figura 1). Se busca una distancia mínima en el recorrido del reloj. Por ejemplo, entre horas contiguas esa mínima distancia será 1 mientras que entre horas opuestas será 12. La Hora 1 distará 1 de la Hora 0 mientras que 12 de la Hora 13. O sea, a medida que se produce un alejamiento entre horas habrá mayor distancia, el incremento es 1.

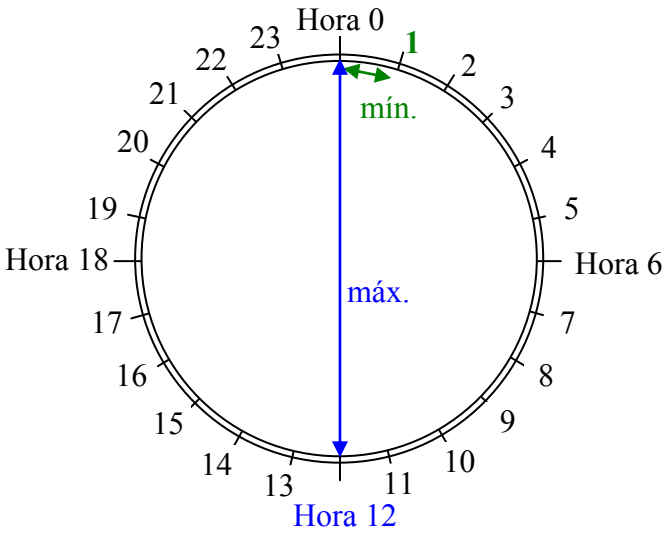


Figura 1

Figura 1:
Las doble flechas indican una de las distancias mínimas y una de las distancias máximas posibles en el reloj.

Buscamos una matriz $P'^{Primaveras\ J}_{n\ x\ n}$ (o simplemente P') tal que cumpla estrictamente que entre individuos consecutivos la distancia sea mínima y que entre individuos opuestos sea máxima cubriéndose por analogía los casos intermedios.
La matriz P' que se muestra a continuación fue realizada considerando las distancias entre horas del reloj de la Figura 1. Se han marcado algunos números con un círculo para

mostrar, por ejemplo, que entre la Hora 5 (columnas) y la Hora 3 (filas) hay una distancia de 2.

$P' =$

Hora 0					Hora 5				Hora 9			Hora 12	Hora 14									Hora 23	
0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1
1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2
2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3
3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4
4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5
5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6
6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7
7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8
8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10	9
9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11	10
10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	11
11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12
12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11
11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10
10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9
9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8
8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7
7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6
6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5
5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4
4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2	3
3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0	1
1	2	3	4	5	6	7	8	9	10	11	12	11	10	9	8	7	6	5	4	3	2	1	0

Hora 3

Hora 13

Esta matriz (que se utilizará para asignar pesos) cumplirá la función de "penalizar" a la matriz D^{escal} de tal manera de restringirla como para que las nuevas distancias entre objetos formen grupos con vectores sucesivos en el tiempo. Primeramente se debe dividir a P' por el máximo de sus elementos de tal manera de tener, también aquí, una matriz escalada de penalización que llamaremos P^{escal}.

Se propone obtener una matriz $D^{Primaveras\ J}_{n\ x\ n}$ (que llamaremos D) tal que:

$$D = D^{escal} + \alpha\ P^{escal}$$

donde α es un número real mayor que cero.

Se deberá buscar el menor α tal que se resuelva en el dendograma la no consecutividad de horas (para un número k de grupos deseado). Para ello se trabaja por prueba y error. Una vez determinado el menor α posible, quedará definida la matriz D y el dendograma obtenido será satisfactorio respecto de la restricción requerida.

Nota: Al "alterar" la matriz de distancias Euclídeas originales entre datos con la matriz de penalización la matriz resultante ya no cumplirá la desigualdad triangular (Sección V.3, pág. 113), sin embargo, esto no implica necesariamente que haya consecuencias indeseables para el análisis por conglomerados jerárquico, dado que este opera con cualquier matriz de distancias.

Para el caso de aplicación, se trabajó con un paso de 0.01 para α y se encontró que para $\alpha=0.05$ se resolvía el problema planteado (se lograba consecutividad en el dendograma). El resultado se corresponde con el dendograma de la Figura 2. Esta figura permite apreciar (ver rectángulos sobre el eje *Y*) que los miembros de grupo son ahora todos consecutivos.

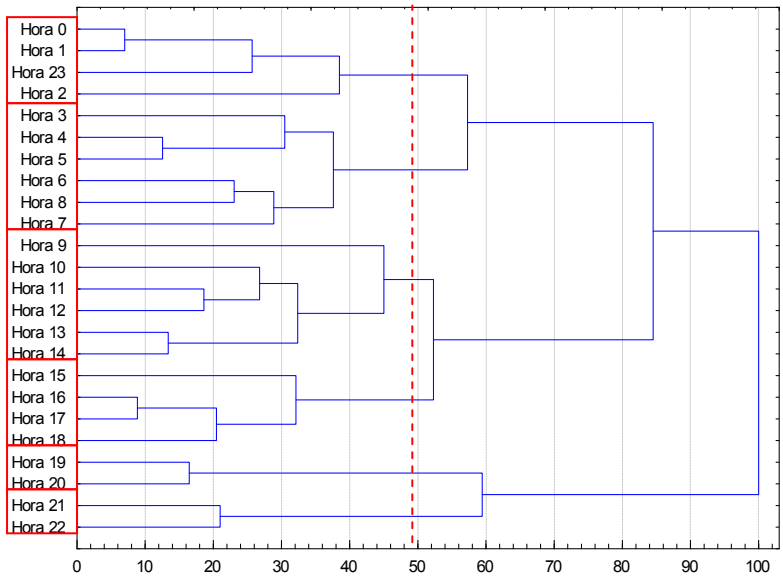


Figura 2

Figura 2:
Dendograma correspondiente a rosetas de frecuencias horarias de vientos por dirección de la Primavera Punto J durante el período 1998- 2003 obtenido considerando la restricción de consecutividad de los miembros de cada grupo para 6 grupos.
El eje *X* son distancias Euclídeas al cuadrado reescaladas. El dendograma fue obtenido normalizando los datos con media y desvío estándar. La distancia Euclídea al cuadrado es la medida de disimilitud adoptada y el criterio de aglomeración es el UPGMA.

En el caso planteado este enfoque sencillo ha mostrado ser satisfactorio pero, si la solución requiere de altos valores de α, el grado de alteración sería muy alto por lo que deberá buscarse otro procedimiento.

Anexo V.6
Método de las *k*-medias

Dada la popularidad de este método y por ser de aplicación secundaria en la tesis solo se describirán lineamientos generales (Seber, 1984; Sharma, 1996) con el objetivo de mostrar los pasos básicos que siguen muchas de las variantes establecidas para definir los integrantes de grupo. Como se señaló en la Sección V.1 (pág. 109). al aplicar un método de partición el número de grupos (*k*) debe estar previamente definido.

a) Seleccionar las "semillas" de los *k* grupos a determinar (*k* semillas).
   La semilla pueden ser algunos de las observaciones (datos) o valores distintos elegidos con algún criterio inicial (ver a_1 a a_4 en el cuadro de texto correspondiente).
b) Asignar cada observación (excepto las que fueron utilizadas como semilla si corresponde) a una de las *k* semillas (para formar *k* grupos) utilizando un criterio de asignación definido, por ejemplo, minimizando la distancia Euclídea entre la observación y las semillas de cada grupo (ver f_1 a f_3 en el cuadro de texto correspondiente).
c) Calcular el centroide (promedio) de cada grupo.
d) Reasignar cada observación a un grupo teniendo en cuenta el criterio definido en el paso b) para el centroide calculado en el paso anterior.
e) Calcular nuevamente el centroide de cada grupo.
f) Continuar el proceso entre d) y e) hasta verificar un criterio de convergencia determinado, por ejemplo, hasta que no haya diferencia significativa entre los centroides calculados en los dos últimos pasos realizados.

a_1) elegir las primeras *k* observaciones como semillas.

a_2) elegir una semilla para el primer grupo. La semilla del segundo se elige con un criterio de distancia máxima respecto de la primera. Y así se calculan todas las semillas respecto de la anterior.

a_3) se eligen las semillas con un criterio al azar.

a_4) utilizar semillas provistas por el usuario (por ejemplo, centroides de grupos obtenidos con aglomeración jerárquica).

f_1) Reasignar las observaciones cuya distancia al centroide sea mínima hasta que se cumpla con un criterio de convergencia.

f_2) Reasignar las observaciones según un criterio (que en general utilizan una función objetivo):
1) minimizar la Traza de la matriz covarianzas intragrupo (W).
2) minimizar el Determinante de la matriz covarianzas intragrupo.
3) minimizar la Traza del producto de la inversa de la matriz intragrupo por la matriz intergrupo (B), o sea, $W^{-1}B$.
4) el máximo autovalor de $W^{-1}B$.

El método de las *k*-medias es conocido por depender de las condiciones iniciales (Rencher, 2002). Esta dificultad puede, en la actualidad paliarse, debido a la velocidad de cálculo de las computadoras que permiten realizar los cálculos con distintos criterios de selección de semillas sin requerir de largas jornadas de operación para el investigador (aunque los cálculos pueden resultar según Steinley (2004) miles). Dado que este método es poco robusto, podrá aplicarse con más confianza, si previamente se realiza una exploración de los datos, con la finalidad de saber si hay atípicos importantes. De no haber atípicos se podrá trabajar directamente, caso contrario se puede recurrir a algún tipo de criterio de normalización robusta de las variables (Steinley, 2004) o a algún método similar pero más robusto (Kaufman y Rousseeuw, 2005).

"Quizás los dragones que amenazan nuestra vida solo aguardan un indicio de nuestra apostura y valentía".
Rainer Maria Rilke
(poeta)

"Everyone on earth will be an environmentalist in the not too distant future, driven there by necessity and experience".
Paul Hawken
(founder of the Natural Capital Institute)

"Ordinary risk analysis asks, how much environmental damage will be allowed? But the precautionary principle asks, "How little damage is possible?""
Thomas Prugh
(World Watch Institute)

"In order to slow, stop and ultimately reverse environmental degradation, we need to understand not only what is directly causing that degradation, but also how human society is contributing through its policies and decisions".
UNEP, Geo Cities Manual

Capítulo VI
Síntesis y conclusiones finales

VI.1 Introducción

La ciudad de La Plata y alrededores (ubicada en el Estuario del Río de La Plata) es una de las seis urbes más pobladas de la Argentina, posee una importante actividad económica, un gran parque industrial cercano al casco urbano, una central térmica de generación de energía, un astillero, un puerto naviero y gran actividad de tránsito vehicular. Como la mayoría de las grandes ciudades, debe poner en consideración la incidencia de enfermedades respiratorias debidas a la contaminación del aire y por su ubicación costera debe afrontar los desafíos del cambio climático global. En este contexto, dado que la ciudad se halla en una zona de la Argentina con escasa capacidad de depuración atmosférica y que no posee una red oficial de monitoreo continuo de los contaminantes del aire, sumado al hecho de que, en estudios prospectivos previos, se habían detectado niveles altos de algunos contaminantes, fue posible formular para este trabajo de tesis un conjunto de objetivos de estudio que permitan tanto enriquecer el conocimiento del ambiente en la zona, como sugerir estrategias para la mejora. Tales objetivos pueden definirse en términos de compilación de información ambiental (que era escasa, dispersa y de calidad disímil), entrenamiento en el manejo de equipamiento de monitoreo y el análisis estadístico de datos para el estudio de patrones de viento y sus dinámicas asociadas al transporte de los contaminantes industriales.

VI.2 En relación al empleo de técnicas espectroscópicas

La capacitación en el manejo de equipamiento para la medición de contaminantes constituye un potencial importante cuando se trata de poner a punto equipos de una red de monitoreo. Como parte del trabajo de tesis se realizaron pruebas de ajuste de una cámara de ensayos de diseño local para poder operar con gases de chimenea en valores de emisión. El doctorando adquirió experiencia tanto en el manejo de equipos ópticos y electroquímicos como en la operación de un laboratorio específicamente diseñado. Se realizaron ensayos de ajuste de cero y calibración de un equipo diseñado en el CIOp (Centro de Investigaciones Opticas) basado en un método no dispersivo para medir simultáneamente SO_2 y NO_2 en la región ultravioleta del espectro. El doctorando participó del ensamblado de un equipo dispersivo experimental (DOAS) destinado a realizar

medidas de prueba de NO_2 en valores de calidad de aire.

VI.3 En relación a los métodos estadísticos

Los métodos de análisis estadístico son de gran utilidad tanto para explorar conjuntos de observaciones como para realizar inferencias. La tesis discute un conjunto amplio de métodos, describiendo algunos aspectos teóricos que el investigador debe conocer desde el punto de vista de las aplicaciones; además, pone en contexto estos métodos en el campo de las ciencias ambientales. Los métodos analíticos y gráficos fueron empleados desde un punto de vista crítico y constituyeron un recurso para describir e interpretar fenómenos ambientales, permitiendo producir y sintetizar información que fundamente la toma de decisiones. El concepto de robustez aparece como imprescindible de considerar (constituye un tema transversal de la tesis), dado que asume trabajar tanto en la exploración de los potenciales valores atípicos como en la aplicación de métodos para el modelado. Para ello se recurrió a varias herramientas tales como los "QQ-Plots" (gráficos cuantil-cuantil) y el método de Componentes Principales (CP).

La detección de parecidos constituye un arte complejo pero, mediante el empleo de herramientas sencillas de correlación (para comparar "formas") y distancias (para comparar "tamaños"), el investigador se encuentra con perspectivas complementarias para comparar conjuntos de datos multivariados. Se presentaron y discutieron el uso de alternativas a cada una de estas herramientas.

Cuando se requiere describir el comportamiento de algunas variables (respuesta) en función de otras (explicativas), el investigador debe plantearse alternativas de regresión. Se presentaron y discutieron distintas aplicaciones de regresión paramétrica (para determinar coeficientes) y no paramétrica (para evaluar tendencias) utilizando procedimientos clásicos y robustos según el caso.

Dado un conjunto de datos multivariados la búsqueda de grupos constituyó uno de los temas de interés de la tesis. Se recurrió al análisis por conglomerados jerárquicos como herramienta principal y se discutió una posible secuencia de pasos para su implementación; se puso énfasis en la estandarización, la selección de un criterio de agrupamiento, la detección de atípicos, la determinación del número óptimo de grupos y la validación. El análisis por escalamiento multidimensional se aplicó como método complementario al análisis por conglomerados, pudiéndose visualizar una gran cantidad de datos de forma simultánea y haciéndose tangible el tipo de estructura de los mismos.

Las Curvas de Andrews se constituyeron como un enfoque eficaz para representar vectores multidimensionales, esto permitió profundizar en el estudio de las características de los grupos.

En relación a la búsqueda de grupos con restricciones, se propone un enfoque sencillo que permite abordar la problemática que se presenta cuando los miembros de un grupo deben satisfacer el requisito de ser consecutivos en el tiempo. Este enfoque, que puede ser útil para algunas aplicaciones, cumple la función principal (dada su sencillez) de introducir al investigador en la temática de encontrar grupos en conjuntos de datos considerando requisitos externos.

El método de las Siluetas sirvió para interpretar y validar grupos utilizando principalmente un medio gráfico. Este método, empleado recurrentemente (por ejemplo, junto al algoritmo de las *k*-medias), permitió realizar mejoras en la clasificación, independientemente del algoritmo utilizado para formar los grupos.

VI.4 En relación a la presencia de dióxido de azufre

a) El dióxido de azufre (SO_2) es un gas cuya presencia en la atmósfera ha sido considerada de gran relevancia a nivel planetario (Smith et al., 2010), tanto debido a su origen natural

(por ejemplo, emisiones de volcanes o de suelos sulfurosos (Macdonald et al., 2004)) como a su origen antropogénico (por ejemplo, industrias o producción de energía). Dada la importancia que este gas posee desde el punto de vista normativo tanto a nivel de lineamientos y leyes internacionales como nacionales, las características de la zona de estudio y los escasos antecedentes de medición del mismo en la zona, los resultados obtenidos en esta tesis (basados en monitoreo no sistemático) destacan la importancia que se le debe prestar al seguimiento de este agente contaminante.

Las observaciones realizadas en el Punto A durante el período 1996- 2000 pueden considerarse como una referencia "histórica" de la contaminación del aire en la zona. Los datos mostraron una tendencia creciente de los promedios anuales. La mayoría de estos promedios superan las 10 ppbv, lo cual en presencia de material particulado (cuyas fuentes principales son la industria y el parque automotor) tiene impacto en las enfermedades respiratorias. Por otro lado, la presencia de SO_2 resultó ser significativa como para afectar materiales y bienes culturales. Dada la existencia concomitante de otros contaminantes industriales, tales como compuestos orgánicos volátiles así como de contaminantes de origen vehicular (NO_2, $PM_{2.5}$, etc.), el monitoreo continuo de SO_2 se vuelve muy importante considerando la exposición general a la que se ve expuesto el habitante de La Plata y alrededores.

Considerando que el aeropuerto (Punto K) es de bajo tránsito y se halla ubicado lejos del Punto A y la ausencia de fenómenos naturales productores de SO_2, las rosetas de concentración permitieron verificar que no hubo para el año 2000 aportes significativos del tráfico de la avenida de alto tránsito más cercana (aporte de los vehículos diesel) al mismo tiempo que la zona industrial de Ensenada (Polo Petroquímico) es una fuente inequívoca de las emisiones de SO_2 quedando por discriminar los potenciales aportes del Puerto La Plata. El empleo de este recurso gráfico, que permite combinar concentraciones de contaminantes con direcciones de viento, dio lugar a la exploración de direcciones de viento que resultan significativas en relación al transporte de los contaminantes del aire, desde el área industrial hacia el casco urbano. Junto al empleo de otras herramientas pudo determinarse que las direcciones NNO- N- NNE y NE definidas como Sector 1 son de gran interés ambiental.

b) En la medida que un sitio de observación se halle más alejado de la fuente se espera una mayor dilución del contaminante al mismo tiempo que este tiene más posibilidades de reaccionar en la atmósfera (el SO_2 puede convertirse en ácido sulfúrico debido a la presencia de humedad). Los promedios diarios de SO_2 observados en el Punto D (alejado de las fuentes industriales de emisión) durante una campaña de 92 días (primavera de 2005) tuvieron un máximo de 8.5 ppbv y un mínimo de 1.6 ppbv. Los promedios diarios estuvieron por debajo del lineamiento WHO (2000a) que es de 48 ppbv pero, en dos ocasiones, se superó el lineamiento OMS (2006) que es de 7.6 ppbv. Esto implica que sobre 92 días este límite máximo recomendable fue superado el 2.2 % de las veces, que extrapolado anualmente equivale a aproximadamente 8 días al año en los que se supera el valor sugerido por el lineamiento. Los promedios horarios no sobrepasaron en ningún caso los estándares US EPA (75 ppbv).

En relación a las direcciones de viento y las concentraciones observadas en el Punto D, se encontró una alta correlación entre los vientos procedentes de las direcciones ENE- E- ESE y los promedios horarios de SO_2. De esta manera pudo definirse el "Sector 2" (ENE- E- ESE) de gran importancia para el transporte de los contaminantes de origen industrial hacia áreas residenciales (tales como Gonnet, City Bell, etc.). También se pudo determinar, para el período de estudio, el carácter lineal que guardan las concentraciones observadas de SO_2 y las frecuencias del Sector 2 observadas en distintos sitios y períodos de registro.

De **a)** y **b)**, y a pesar de los distintos períodos de muestreo y de la diferencia en la calidad de los datos, surge que el Punto A permite detectar valores de SO_2 altos debido a su cercanía a las fuentes industriales mientras que el Punto D permite cuantificar valores bajos de SO_2 debido principalmente al efecto de dilución por distancia a las fuentes. Además de sus respectivas posiciones estratégicas en relación a las fuentes industriales y dado que ambos sitios pertenecen a organismos que dependen de las autoridades estatales, los puntos A y D se muestran aptos como potenciales sitios de monitoreo de contaminantes de una red oficial de vigilancia.

VI.5 En relación a las frecuencias horarias de direcciones individuales de vientos observadas en los puntos A y J

Los puntos A y J demostraron tener durante el período 1998- 2003 y en términos generales patrones horarios similares en lo que hace a las ocurrencias de vientos por dirección. Las principales diferencias entre estos sitios son atribuibles al fenómeno de brisas de mar y tierra (principalmente en verano), a la rugosidad de los terrenos y a la calidad de los datos.
La comparación entre los dos sitios, utilizando el método de las distancias Euclídeas al cuadrado, mostró buena similitud en todos los casos mientras que, el método de correlación, ofreció un panorama irregular: no se encontró correlación lineal importante en algunas direcciones tales como las comprendidas entre OSO y NO (en sentido horario). Esto último implica que no es posible predecir la ocurrencia horaria de una dirección de viento de un sitio a partir del otro (al menos para algunas direcciones). Este hecho es importante de considerar cuando concentraciones observadas de un determinado contaminante del aire en un sitio alejado de los puntos A y J quieran ser relacionadas con direcciones individuales de vientos observados en A o en J.
Por otra parte, al correlacionar "sectores 1" (NNO-N-NNE-NE) observados desde distintos sitios (puntos A, D, J o K) la correlación lineal es en general alta. Esta homogeneidad posibilita establecer correlaciones entre este sector y las concentraciones observadas en cualquier lugar de la ciudad y alrededores. Esto parece indicar, en relación a las direcciones individuales, que hay algún fenómeno de compensación que se pone en juego al sumar direcciones individuales.
Algo análogo es observable para el Sector 2 (ENE-E-ESE) y el Sector 3 (ENE-E-ESE-SE-SSE-S-SSO-SO-OSO), vale decir que, estos sectores permanecen muy similares independientemente del sitio desde el cual son realizados los registros.

VI.6 En relación a algunos grupos de direcciones de viento (sectores 1 y 2)

Los sectores 1 y 2 no solamente son importantes porque transportan contaminantes desde el complejo industrial hacia el casco urbano (Sector 1) y hacia áreas residenciales (Sector 2) sino por su alto porcentaje de ocurrencias. En el período 1998- 2003 el promedio de ocurrencias del Sector 1 en los puntos A y J fue de 28.3 % mientras que las del Sector 2 fue de 24.2 %. Evaluando el perfil horario de estos sectores se determinó que las horas de máxima probabilidad de ocurrencia del Sector 1 tienen lugar entre la Hora 9 y la Hora 14 en verano (42.1 %) y entre la Hora 11 y la Hora 16 en invierno (33.2 %). Para el Sector 2 los máximos se dan entre la Hora 18 y la Hora 21 con un porcentaje de ocurrencias del 47.7 % para el verano mientras que de 25.8 % para el invierno. Estos porcentajes se mantienen similares si se amplían los sitios y/o los períodos de observación siendo la suma de ocurrencias de ambos sectores generalmente mayor al 50%.
Dada la relevancia de estos sectores se investigó la influencia que tienen el ciclo anual (importancia de las estaciones) y el ciclo diario (importancia de las horas del día).
El Punto A mostró mayor variación diurna que el Punto J siendo esta variación más pronunciada en el Sector 2 que en el Sector 1. Los vientos del Sector 1 están originados

principalmente en el anticiclón del Atlántico Sur teniendo aportes locales de la brisa marina de la mañana. La influencia del anticiclón es la misma para ambos sitios, que solo difieren en la rugosidad del terreno pero la brisa marina es más importante en el Punto A (próximo al río) que en el Punto J (que se halla tierras adentro). La circulación de la brisa marina influye algo más en las direcciones involucradas en el Sector 2 que en el Sector 1. Los vientos rotan desde el Sector 1 hacia el Sector 2, este último se va afianzando durante la tarde. Se cuantificó en qué medida el ciclo diario tiene más peso que el ciclo anual y se detectó que la brecha es mayor en el Punto A que en el Punto J.
Por otra parte, ambos sectores se mostraron estables (ausencia de tendencia creciente o decreciente) durante los períodos de estudio 1998- 2003 y 1998- 2009.

VI.7 En relación a las velocidades de viento

Las velocidades medias corregidas de los vientos mostraron ser muy similares en los puntos A y J en los períodos 1998- 2003 y 1998- 2009. El promedio general resultó de 7.4 km h^{-1} lo cual corresponde según la Escala Beaufort para tierra a "brisa suave". Las velocidades observadas en el Punto K (único sitio oficial de observaciones meteorológicas) durante 2001- 2010 mostraron ser alrededor de dos veces más altas ("brisa leve" en la Escala Beaufort), en promedio 14.0 km h^{-1}. Estas diferencias se justifican debido a las distintas rugosidades de los terrenos involucrados y la diferencia en la calidad de los datos. El Punto I mostró velocidades promedio de 13.0 km h^{-1} durante el período 1967- 1994 a partir de valores observados a 40 m de altura, el promedio corregido es de 9.2 km h^{-1}.
Distintos autores señalan que velocidades menores a 7.2 km h^{-1} contribuyen a la acumulación de contaminantes. Calculando los percentiles para las velocidades corregidas del Punto J (1998- 2007) se determinó que el 50% de las veces las velocidades son menores a 7.1 km h^{-1}.

VI.8 En relación a la presencia de calmas

VI.8.1 Caracterización de las calmas

La presencia de calmas puede constituir una condición meteorológica propicia para la acumulación de los contaminantes del aire en las cercanías de las fuentes de emisión. Las calmas fueron cuantificadas calculando el porcentaje de eventos por debajo del límite de detección del anemómetro (o sea, velocidades de viento < 1.6 km h^{-1}) respecto del total de eventos (vientos y calmas). Se estudiaron las curvas de distribución horaria según la estación del año en los puntos A (1997- 2003), J (1997- 2006) y K (1995- 2005) de observación. Los resultados muestran que existe un patrón espacial generalizado de las curvas, las cuales, guardan una relación directa con los procesos del ciclo diario de la capa límite planetaria. Los sitios aparecen bien correlacionados mostrando el grado de generalización de los patrones hallados. El promedio general de ocurrencia de calmas para la estación verano fue de 14.7%, para el otoño fue de 19.1%, para invierno de 12.8% y para la primavera de 11.6%, dando un promedio total de 14.6%.
Analizando la estructura de las calmas según su duración, se encontró que el 90.1% de estas tienen duraciones menores a 5 horas. Las calmas de 1 hora representan el 50.6% de los eventos y presentan picos a lo largo de las horas del día. Las calmas largas (de dos horas o más) muestran un patrón con picos en horas del amanecer y el anochecer.

VI.8.2 Patrones de viento inmediatamente después de las calmas

Las primeras direcciones de viento que aparecen luego de los períodos de calma son fundamentales para conocer el destino de los contaminantes que se han acumulado. Se diseñó una "roseta de vientos de salida de calmas" (RVSC), computando las direcciones de viento que aparecían durante la primera hora luego de transcurrido un episodio de calma a

partir de los datos del Punto J durante el período 1998- 2007. Las RVSC obtenidas (que implican mucho tiempo de cálculo) mostraron un patrón bastante similar al de las rosetas correspondientes de rangos completos de velocidades de viento (RRC). Es decir, con poco error (estimado aplicando *SAD*); pudo determinarse que el patrón de vientos de velocidades bajas es similar al patrón de vientos correspondiente al de RRC.
La presencia del Sector 1 en la RVSC fue para el período de estudio 24.1% mientras que la del Sector 2 de 23.5% lo cual da una idea de que ambos sectores se mantienen importantes a velocidades bajas (para el Sector 1 la velocidad promedio de la RVSC es de 2.8 km h^{-1} mientras que para el Sector 2 es de 2.6 km h^{-1}).

VI.9 En relación al efecto combinado de direcciones relevantes, calmas y velocidades de viento

En base a lo concluido en secciones anteriores, es posible resumir, considerando distintos sitios y períodos de muestreo, lo siguiente: los sectores 1 y 2 se hallan presentes en promedio más de la mitad del tiempo (>50%), las presencia de calmas puede estimarse en promedio en 14.6% de las veces y las velocidades de viento son la mayor parte del tiempo bajas como para permitir acumulación de contaminantes (el 50% de las veces menores a 7.1 km h^{-1}). Estos hallazgos guardan coherencia con el diagnóstico dado por Gassmann (1998) en relación a la baja capacidad de depuración atmosférica (Capítulo I- Sección I.1.1 y Capítulo III- Sección III.4) de la zona durante las estaciones de invierno y otoño.

VI.10 En relación a la ubicación de un sitio potencial para evaluar la contaminación de fondo

Dentro de los sitios de medición de agentes contaminantes o parámetros meteorológicos (todos ellos pertenecientes a instituciones del Estado), el Punto K aparece como el más adecuado para el seguimiento de la contaminación de fondo. Esto se puede verificar dado que existe un conjunto de direcciones de viento (ENE-E-ESE-SE-SSE-S-SSO-SO-OSO) a las que se agrupó con el nombre de Sector 3, que desde el punto de vista del Punto K no transportan contaminantes de origen industrial ni vehicular hacia la mayor parte de la población expuesta. El Sector 3 (que incluye las direcciones del Sector 2) tiene un promedio de ocurrencias del 62.4 % para las cuatro estaciones del año.

VI.11 En relación a los patrones horarios de vientos en La Plata y alrededores

Las 24 rosetas horarias de frecuencias de viento que describen el "día" para cada estación del año y cada sitio de monitoreo (Punto A y Punto J) en un período determinado de tiempo, constituyen un volumen importante de información. Dados los vientos dominantes y el ciclo diario, que tiene lugar en la capa límite planetaria junto al fenómeno de brisa de mar y tierra, se espera que tales patrones horarios posean una estructura de grupo. Aplicando análisis por conglomerados jerárquicos y escalamiento multidimensional fue posible determinar y hacer visible un panorama generalizado de los patrones de direcciones de viento en la zona de La Plata y alrededores.
La aplicación de análisis por conglomerados determinó que 5 grupos (es decir 5 etapas del día representadas por rosetas horarias de direcciones de viento) describían bien las ocurrencias diarias. Se pudieron visualizar de manera sencilla, a través de las resultantes de los promedios de grupo, diferencias y similitudes entre estaciones del año y sitios de observación. Pudo apreciarse la presencia de los vientos dominantes y la rotación de N a SE en sentido horario entre la mañana y la noche como fenómenos característicos. También se pudo constatar que las direcciones observadas en los puntos A y J para el período 1998- 2003 presentan patrones similares a los de una zona más amplia (Estuario del Río de La Plata).

El "mapa" de las rosetas horarias obtenido mediante EMD permitió visualizar la influencia de la brisa marina, la cual se mostró más presente en las estaciones de verano y primavera que en otoño e invierno. También pudo evidenciarse a partir de este método, la mayor sensibilidad del Punto A a la brisa marina con respecto al Punto J.

VI.12 En relación a los patrones espaciales de viento en el estuario del Río de La Plata a partir de un modelo de mesoescala

Las salidas de un modelo de mesoescala, diseñado por el Dr. G. Berri y sus colaboradores, que predice rosetas de direcciones y velocidades de viento en una zona amplia del Estuario del Río de La Plata, son tomadas como datos de entrada para la aplicación de un método de análisis por conglomerados con el objetivo de sintetizar información espacial e identificar áreas de alta homogeneidad de vientos.
A partir del análisis por conglomerados jerárquicos se propusieron tres soluciones posibles; las tres permiten distinguir la existencia de grupos a lo largo del río, observándose mayor discriminación de áreas homogéneas en la costa noreste (uruguaya) que en la sureste (argentina). Esto pone en evidencia los principales aspectos de los vientos de superficie en la zona, que se ven afectados fuertemente por la presencia de la brisa de mar y tierra en donde pueden apreciarse la influencia de los accidentes costeros.
En relación a la actual cantidad y distribución de estaciones meteorológicas, el análisis realizado da una cantidad y distribución similar a la existente, pero permite inferir la necesidad de instalar más estaciones sobre el río y sobre la costa uruguaya.

VI.13 En relación al empleo de Curvas de Andrews

La transformación de rosetas de viento de 16 direcciones en Curvas de Andrews permitió visualizar el grado de homogeneidad de los grupos obtenidos mediante análisis por conglomerados jerárquicos y detectar la existencia de singularidades.
Una de las ventajas del nuevo arreglo de grupos encontrado mediante el empleo de estas curvas es que permitió apreciar el cambio paulatino en los vientos dominantes, desde el mediodía donde hay predominio de vientos del N, hasta el atardecer donde hay predominio de vientos del E. El resultado obtenido estuvo apoyado mediante la aplicación de otros indicadores, tales como el índice de Calinski y Harabasz.

VI.14 En relación al Método de las Siluetas

El ejemplo mostrado pone en evidencia mejoras en el reordenamiento de los integrantes de grupo obtenidos por otros métodos (conglomerados jerárquicos y *k*- medias). Dado que el Método de las Siluetas asigna como representante de grupo a un miembro "real" (y no a un promedio de algunos miembros), es posible utilizar esta ventaja en el diseño de trabajos de campo. Por ejemplo, se pueden realizar mediciones de contaminantes en una hora del día que sea la más representativa de los vientos de una franja horaria determinada.

VI.15 En relación a un criterio alternativo de muestreo

Existen circunstancias en las que el seguimiento continuo de un determinado contaminante del aire ya no es necesario (valores históricos muy bajos) o justificable (debido a costos). Dado que no es aconsejable abandonar totalmente el seguimiento de dicho contaminante se debe recurrir a un método discreto (tal como el de la pararosanilina para el SO_2). Se mostró, a modo de ejemplo, un procedimiento estadístico que permite reemplazar el muestreo continuo por uno discreto de manera controlada. Es posible determinar, mediante un método de regresión lineal robusta, cual es la hora del día que mejor representa los promedios diarios observados durante un período determinado. Una vez determinada la hora es posible definir la frecuencia de muestreo a realizar con el método discreto.

La metodología propuesta alcanza a cualquier gas, aerosol o material particulado ambiental y permite también realizar el seguimiento de picos de concentración.

VI.16 Perspectivas

Un programa ambiental

La prosperidad de una ciudad se define en base a los niveles y características de productividad, desarrollo de infraestructura, calidad de vida, equidad e inclusión social y la sustentabilidad de su ambiente (UN-HABITAT, 2012). Las actividades económicas desarrolladas en una región deben guardar una relación coherente con las agendas ambientales; ambas dependen en gran proporción de las políticas de estado que deben ser integradoras de los medios naturales y de todos los actores sociales. En este equilibrio, el derecho ambiental juega un rol fundamental; un cambio de paradigma (del antropocentrismo al biocentrismo) en donde "se reconoce a la naturaleza como sujeto de derechos" según se estableció en la nueva Constitución Nacional de Ecuador en el 2008 (Prieto Méndez, 2013) puede ser muy conducente. Son muchas y graves las consecuencias del desequilibrio entre productividad económica y ambiente (NU, 2009); todos los recursos naturales se ven afectados impactando sobre el presente cercano y dejando sus huellas para el futuro pero "ya no basta decir que debemos preocuparnos por las futuras generaciones… lo que está en juego es nuestra propia dignidad… [la] del propio paso por esta tierra" (CEP, 2015). Desde un punto de vista práctico y en relación al recurso atmósfera "la única forma de saber con certeza si existen, si se están generando, o si están empeorando los problemas de la contaminación del aire es mediante la medición de los contaminantes" (Kork y Sáenz, 1999). Existen metodologías que permiten realizar una evaluación integrada del ambiente (UNEP, 2010) como punto de partida de diversos proyectos ambientales. Abundante bibliografía internacional provee de recomendaciones y protocolos para monitorear la calidad del aire (PNUMA-OMS, 2002; EPA, 2008, 2013; WMO, 2008; PNUMA, 2012). A nivel de América Latina y el Caribe existen instituciones como el Foro de Ministros de Medio Ambiente que impulsan planes de acción regionales (UNEP, 2014b).

Cualquier programa ambiental de largo plazo requiere de un marco legal e institucional así como de recursos económicos que aseguren su implementación y mejora sostenidas. Sería muy propicio la creación de una ley que imponga, como punto de partida, la instalación progresiva de redes de monitoreo continuo de los contaminantes del aire en ciudades de más de 300 000 habitantes, teniendo en cuenta que ciudades más pequeñas, tales como Tandil en la Provincia de Buenos Aires afrontan perspectivas problemáticas (Sosa, 2015).

Un programa de vigilancia ambiental no solo debe contemplar la instalación de una red de monitoreo para el seguimiento continuo de las especies contaminantes del aire, tales como SO_2, NO_x, CO, PM_{10}, $PM_{2.5}$, PM_1, material particulado total en suspensión, HAPs, O_3 y COVs, sino que también, debe llevar a cabo la instalación de estaciones meteorológicas con capacidad para medir parámetros tales como velocidad y dirección de vientos, temperatura, humedad, presión, radiación solar, perfiles de temperatura y velocidad de viento en altura y altura de capa de mezcla (estos tres últimos parámetros muy importantes para la formulación de modelos de dispersión en base a observaciones). Dentro de tal programa, y dado que existe un inventario de emisiones industriales (aunque no accesible públicamente), se constituye como un requisito (Friedirch y Reis, 2004) conocer de forma detallada la estructura de las fuentes urbanas de emisión con el objetivo de generar un inventario de las mismas.

Algunas referencias sobre equipos de monitoreo

Un equipo optoelectrónico similar al DOAS es el LIDAR ("Light Detection and Ranging"-

detección y escaleo de luz) (Weitkamp, 2005) que es utilizado con éxito entre otras aplicaciones, para medir la evolución diaria de la capa de mezcla (Seibert et al., 2000; Sicard et al., 2006; Emeis et al., 2008). A este respecto, cabe agregar que en Argentina (DEILAP -Departamento de Investigaciones en Láseres y Aplicaciones-, dependiente de CONICET) hay experiencia en el manejo de este tipo de instrumentos (Fochesatto et al., 1995; Lavorato et al., 2002; Otero et al., 2006, 2011). Para el seguimiento de los perfiles de viento en tiempo real existe equipamiento como el SODAR ("Sound Detection and Ranging"- detección y escaleo de ondas acústicas) recomendado por la EPA (2000).
La red de monitoreo podrá contar principalmente con analizadores puntuales (del tipo de la unidad mostrada en la Sección II.3.3- Capítulo II o equipos como el FH62-C14 de Termo Scientific® con capacidad para muestrear y consecutivamente medir por atenuación beta material particulado total en suspensión, PM_{10}, $PM_{2.5}$ y PM_1 proveyendo valores en tiempo real) pero, será importante poner en consideración la instalación de equipamiento DOAS que puede cubrir varios kilómetros de monitoreo en varias direcciones, con capacidad para medir simultáneamente varios gases y visibilidad. Una abarcativa obra sobre el ambiente, los métodos de medición de las especies contaminantes y DOAS es la de Platt y Stutz (2008). Los equipos DOAS tienen gran versatilidad en relación a los objetivos del monitoreo: Edner et al. (1993) monitoreaban varios gases ambientales utilizando un DOAS con tres fuentes de luz cubriendo distancias de 200, 1600 y 2000 m respecto del sistema de recepción, en la ciudad de Lund (Suecia) de aprox. 50 000 hab. Las alturas a las que estaban instaladas las fuentes variaban entre 10 y 20 metros. Kourtidis et al. (2000) realizaron el seguimiento de benceno y tolueno procedente de fugas y productos no quemados de la combustión de naftas en una zona urbana de Tesalonika (Grecia) con un equipo comercial DOAS ubicado sobre edificios a 50 m de altura; los autores ponen en evidencia la confiabilidad del equipamiento utilizado. Avino y Manigrasso (2008) destacan las ventajas de realizar el seguimiento de benceno, tolueno, NO_2, O_3 y SO_2 con un equipo DOAS (modelo AR 500 fabricado por Opsis de Suecia) ubicado 10 m sobre el nivel del suelo en áreas urbanas de Roma durante el período 1991- 2000. Zoras et al. (2008) emplearon un DOAS para seguir varios contaminantes en un cañón de ciudad (zona acanalada formada entre la calle y altos edificios de ambos lados) en Kozani (Grecia) entre 10 y 15 metros de altura sobre el nivel del suelo cubriendo una distancia aprox. de 300 m de longitud. Lee et al. (2005) utilizaron un DOAS para medir BTX (benceno- tolueno- xileno) a 12 metros de altura cubriendo una distancia de 740 m de longitud en un área urbana de Seul (Corea). Por otro lado, Chiu et al. (2005) realizaron mediciones con un DOAS en el centro de una refinería de petróleo, ubicada en el sur de Taiwan, para seguir las concentraciones de HCHO (formaldehído), NO_2, SO_2, O_3, benceno y tolueno. Los autores destacan la confiabilidad y el bajo costo de mantenimiento del equipo utilizado. Kim y Kim (2001), Lee et al. (2005) y Zoras et al. (2008) hallan muy buenas compatibilidades entre las unidades puntales y los equipos DOAS. En particular, Lee et al. (2005) indican las causas potenciales de las diferencias entre los valores obtenidos con una unidad puntual y un equipo DOAS.

Algunas pautas para el dimensionamiento de una red

Dadas las características de La Plata y alrededores, sería óptimo instalar un DOAS que cubra el área del Parque Industrial de Ensenada y otro que cubra el área central del Casco Urbano. Al menos un equipo LIDAR para el seguimiento de la altura de la capa de mezcla será de gran importancia, su ubicación debe surgir a partir de estudios preeliminares. Tanto el DEILAP como el CIOp cuentan con experiencia en el desarrollo y operación de este tipo de instrumentos.
Los puntos A, D, J y K parecen apropiados, en primera instancia, para la instalación de una

red de analizadores continuos puntuales (Sección IV.6.10- Capítulo IV) pero, otros sitios tales como el Punto B (centro del Casco Urbano) y otras ubicaciones dentro del parque industrial de Ensenada, en el puerto y en el astillero aparecen también como necesarias. También resultan de interés sitios donde exista predominio de los aportes vehiculares, algunos de estos lugares se hallan indicados en MLP-UNLP (2001) como zonas de congestión de tránsito y otros son sugeridos en AAPLP (2006).

Además de los equipos DOAS, el punto de partida podría estar constituido por una red inicial de 5 o 6 sitios (CPCB, 2003) que realice monitoreo continuo con analizadores puntuales durante un período de dos o tres años en fase de diagnóstico, a partir de los cuales habrá elementos fundados para realizar un redimensionamiento (ampliación y reubicación) de la red. La red inicial podrá estar apoyada, complementada y extendida con medidores pasivos (por ejemplo, para el seguimiento de NO_2 de origen vehicular) que, en la actualidad, son utilizados con éxito (tal como el caso de Rosario (PAR, 2012) en Argentina).

Posteriores ampliaciones deberán incluir puntos de monitoreo en áreas urbanas de Ensenada y en áreas urbanas e industriales de Berisso (tendientes a cubrir el Gran La Plata), en la zona del Parque Industrial de La Plata (hacia el sudoeste del Casco Urbano) y en los Centros Comunales de mayor cantidad de habitantes tales como Villa Elvira, Los Hornos, Tolosa y City Bell.

Siguiendo la línea costera y en distintas profundidades, tanto tierra como mar adentro, sería muy importante establecer puntos de observación meteorológicos con la finalidad de caracterizar las celdas de circulación de la brisa de mar y tierra (determinando su penetración, condiciones de formación, frecuencia, etc.) y poder luego, estudiar su importancia en relación al transporte y reciclado de los contaminantes del aire.

A modo de ejemplo de relación población/superficie/sitios de monitoreo, se citan algunas ciudades con distinto grado de experiencia en la operación de redes de monitoreo de los contaminantes del aire:

- La ciudad de Cracovia (Polonia) contaba con aprox. 700 000 hab. en 1999 en un área metropolitana de aprox. 1000 km^2; 17 estaciones formaban parte de la red (Jedrychowsky et al., 1999).
- La ciudad de Guatemala (República de Guatemala) cuenta con aprox. 3 millones de habitantes en un área de aprox. 850 km^2, posee una red de monitoreo incipiente de 6 sitios para los contaminantes criterio (USAC- MAG, 2012).
- La región metropolitana de Montevideo (Uruguay) cuenta con aprox. 1,2 millones de hab. en un área aprox. de 530 km^2, posee desde 2004 una red de monitoreo y al 2011 contaba con 8 estaciones permanentes de calidad de aire (IACA, 2011).
- El área metropolitana de Rosario (Argentina) poseía alrededor de 1,2 millones de habitantes al 2010 en un área de aprox. 178 km^2. A partir de 2004 comenzó a funcionar una red de monitoreo, en 2012 la ciudad contaba con 25 sitios de seguimiento de NO_2 (PAR, 2012).
- El área metropolitana de San José de Costa Rica (Costa Rica) es de aprox. 2000 km^2 y alberga alrededor de 2 millones de habitantes; se realizan monitoreos sistemáticos que dependen del contaminante; posee entre 9 y 25 sitios de monitoreo según la especie (CR, 2012; PNUMA, 2012).
- El área metropolitana de la Ciudad de México alberga alrededor de 20 millones de personas en una superficie de aprox. 7900 km^2. Cuenta con 50 estaciones de monitoreo (36 automáticas y 14 manuales), 15 estaciones meteorológicas independientes del servicio meteorológico y un sistema móvil (Perevochtchikova, 2009).

Cabe citar que ciudades avanzadas y con larga trayectoria en el monitoreo del aire, como Londres (Reino Unido), poseen una red para el seguimiento de la contaminación urbana,

otra para la vehicular, otra para la industrial y otra para la zona rural (LAQN, 2015).
Es oportuno agregar que lo presentado hasta aquí en esta sección, y en el contexto de toda la tesis, permite establecer algunas de las bases necesarias para la realización de un estudio de costos de inversión en equipamiento de monitoreo.

Beneficios potenciales de una red

Una red operativa del monitoreo de la calidad del aire constituye un sistema que se halla en constante mejora, para ello existen abordajes estadísticos que permiten ir optimizando la red (Borge et al., 2014). En ese contexto, y a partir de los datos observados provistos por las estaciones de monitoreo, será posible abordar modelos de predicción. Los trabajos de investigación, citados a lo largo de esta tesis sobre la calidad del aire en la zona, refieren a campañas cortas llevadas a cabo en puntos focales; la instalación de una red posibilitará a los investigadores tener perspectivas de largo plazo tal como el seguimiento del impacto de la calidad del aire en la salud, la influencia de los fenómenos de isla de calor e isla fría en la calidad de vida urbana (Rosenzweig et al., 2011) o tener registros locales de el índice de ventilación de la ciudad. La escala espacial se verá también enriquecida pudiéndose establecer relaciones con fenómenos de mesoescala y escala sinóptica. La red de monitoreo constituirá una herramienta fundamental para generar información de acceso público (PLN, 2004), evaluar el cumplimiento de los estándares ambientales, activar procedimientos para situaciones de alerta, alarma y emergencia que salvaguarden la salud de la población, planificar el desarrollo urbano e industrial y proveer elementos que den cuenta del estado y tendencias de la calidad del aire contribuyendo al enriquecimiento de "modelos de ciudad" que incluyen este parámetro (San Juan et al., 2006; UN- HABITAT, 2012). La **calidad del aire** podrá constituirse entonces en uno de los parámetros *conocidos* que hacen a la **calidad de vida** en La Plata y alrededores.

Epílogo

Al igual que los contaminantes fisicoquímicos del aire, el monitoreo continuo de los ruidos constituye un tema pendiente en la ciudad de La Plata y sus alrededores (MLP- UNLP, 2001; Rosenfeld et al., 2005; Dicroce et al., 2010). Una vez montada la infraestructura que exige una red como la sugerida en esta tesis y a pesar de las diferencias de criterio que se deben seguir, estará más facilitada la instalación del correspondiente equipamiento para el monitoreo continuo de los ruidos.

Indice de Figuras

Figura III.2	Escala idealizada de movimientos de la atmósfera. El eje de las *X* indica la duración del fenómeno (que se ha colocado a manera de ejemplo). El eje de las *Y* indica la extensión probable que alcance el fenómeno atmosférico (las magnitudes son solo indicativas). (*) Las trombas marinas (llamadas también mangas de agua) consisten en un intenso vórtice o torbellino que ocurre sobre un cuerpo de agua, usualmente conectado a una nube cumuliforme. (**) Este viento que se da en las Rocallosas durante los meses de invierno, es un fenómeno único que puede aumentar las temperaturas más de 20 grados centígrados en un día. (***) Las "westerlies" son circulaciones de viento en altura que ocurren en las latitudes medias de oeste a este en el hemisferio norte.
Figura III.3	Efecto de la velocidad horizontal del viento en la dilución de los contaminantes. Las partes superiores correspondientes fueron tomadas de Lutgen y Tarbuck (2013) mientras que las inferiores de Vallero (2008). Ambas representaciones permiten comparar el efecto de dilución cuando la velocidad se triplica. Por ejemplo, el viento en a) es de 36 km h^{-1} mientras que en b) es de 12 km h^{-1}. Las "esferas" mostradas en la parte de abajo de cada figura muestran las "unidades de masa" de aire contaminado en la unidad de longitud que se desplazan según la velocidad del viento. Es apreciable como una velocidad relativa más baja (del orden de tres veces tal como lo muestra la parte b) de la figura) induce mayor acumulación de contaminantes con la consecuente reducción de la visibilidad.
Figura III.4	a) Perfiles de velocidad horizontal de viento según la rugosidad del terreno. La velocidad máxima se corresponde para cada caso con el viento gradiente (un viento de velocidad constante que sopla paralelo a isobaras curvas) que tiene lugar en el límite de la CLP. Las escalas sobre los perfiles representan porcentajes de velocidad respecto del viento gradiente. En el eje de las *X* se ha puesto con fines comparativos un límite de 36 km h^{-1} como tope. b) Perfiles de viento con la altura según tres casos característicos de estabilidades atmosféricas (adaptada de Oke (1987)) b1) Neutra b2) Inestable y b3) Estable.
Figura III.5	a) Analogía que representa la atmósfera neutra en correspondencia con la Figura III.6a). b) Analogía que representa la atmósfera inestable, en correspondencia con la Figura III.6b). c) Analogía que representa la atmósfera estable en correspondencia con la Figura III.6c).
Figura III.6	Perfiles atmosféricos de temperatura. a) Atmósfera Neutra b) Atmósfera Inestable c) Atmósfera Estable débil y d) Atmósfera Estable fuerte. La curva a rayas en rojo representa la adiabática seca mientras que la curva en azul representa los distintos casos que puede tener el perfil de temperatura real del ambiente.
Figura III.7	a) Perfil de temperatura con dos tipos de inversiones b) Subsidencia c) Inversión nocturna. La zona celeste opaco en la parte b) indica la presencia de agentes contaminantes (zona gris) acumulados en las cercanías de la base de la capa de inversión. La parte c) muestra una atmósfera con acumulación de contaminantes (en proporción mayor que en la figura anterior) hasta llegar a la base de la capa de inversión. Las fotografías fueron tomadas de Lutgen y Tarbuck (2013).
Figura III.9	Algunas formas que adquieren las plumas de chimeneas según los distintos tipos de estabilidades atmosféricas a) Forma de remolino (predominio de turbulencia vertical- Figura III.4b2) b) Forma de cono (equilibrio entre turbulencia vertical y horizontal- Figura III.4b1) y c) Forma de tubo (predominio de turbulencia horizontal- Figura III.4b3).
Figura III.10	Gradientes de presión en dos áreas que se hallan a temperaturas distintas (un gradiente típico cercano a la superficie terrestre es de 1 hPa/8.6m). La superficie de presión homogénea **P** ha sido tomada como referencia y se halla a la misma altura en los dos casos. **P+1** indica una unidad arbitraria por encima de **P**, podría ser por ejemplo, 1 hPa (hecto Pascal). a) base a $\mathbf{T_1}$ tiene las superficies de igual presión separadas una cierta distancia x_1. b) base a $\mathbf{T_2}>\mathbf{T_1}$ muestra como la disminución de densidad del aire por la elevación de la temperatura de la base produce una mayor separación (x_2) entre las superficies de igual presión.
Figura III.11	Celda de circulación de la brisa marina. La denominación del fenómeno se debe al viento que sopla en la parte baja de la celda desde el mar hacia la tierra.

Figuras del Capítulo IV

Figura IV.18	El eje *Y* izquierdo refiere a las ocurrencias de vientos del Sector 2 observadas en los puntos A y J en primaveras de distintos períodos. El eje *Y* derecho indica la escala de las concentraciones horarias de SO_2 observadas en el Punto D durante una campaña corta en la primavera de 2005.
Figura IV.19	RR es la recta obtenida mediante un método robusto. CM es la recta obtenida mediante cuadrados mínimos.
Figura IV.20	Serie original del Sector 2. Influencia diaria y estacional sobre el Sector 2 en el Punto A (1998- 2003) y en el Punto J (1998-2003; 1998-2009). Residuos del Sector 2 en el Punto A y la curva de suavizado correspondiente. a) $Y_{S_2}^{A}(t)$ representa la frecuencia de ocurrencias de los vientos del Sector 2 observadas en el Punto A respecto del total de ocurrencias durante el período 1998-2003 (curva azul). $Y_{S_2}^{J}(t)$ ídem para el Punto J pero cubriendo el período 1998-2009. Cada punto del gráfico representa la frecuencia de vientos soplando desde el Sector 2 para una determinada hora (*t*) del día para una particular estación del año y para cada año del período especificado. Los valores de t están identificados cada 24 datos y están expresados de forma abreviada, por ejemplo, Ver 00 H0 indica la Hora 0 del Verano del año 2000. La cantidad total de datos es de 576 puntos para el Punto A (que cubre 6 años de observaciones) mientras que de 1152 datos para el Punto J (que cubre 12 años). b) El eje de las *Y* representa el porcentaje de ocurrencias del día promedio para el Sector 2 desde el punto de vista de los puntos A (líneas azules) y J para los dos períodos de estudio (líneas negras). El eje de las *Y* fue construido promediando cada hora acumulada según los años y las estaciones del año. c) El eje *Y* representa el porcentaje de ocurrencias del promedio de las estaciones. d) Residuos de la serie de la Figura IV.20a en el Punto A. La curva suavizada fue obtenida mediante la aplicación de un método de regresión local (LOESS) (Sección IV.3.3). Las líneas verticales señalan el inicio de año.
Figura IV.21	Distribución horaria de las calmas en distintos sitios de monitoreo para la estación verano elegida como referente y por cuestiones de espacio. El eje de las *Y* representa los promedios de frecuencias de ocurrencia de calmas en relación al total de ocurrencias expresadas en %. La curva llena suavizada (verde) representa el promedio de los tres sitios.
Figura IV.22	Calmas acumuladas (%) en intervalos de 1 hora para cada estación del año en los puntos A, J y K. Los porcentajes están expresados respecto del total de duraciones y horas del día.
Figura IV.23	Ubicación de las calmas (%) a lo largo del día según diferentes duraciones: a) **1 hora** de duración b) **2 horas** de duración c) **3 horas** de duración d) **4 horas** de duración e) **5 horas** de duración. Los porcentajes se hallan expresados respecto del total de duraciones (hasta 20 horas) a lo largo de una determinada hora. La línea recta horizontal central de cada gráfica representa el promedio de ocurrencia de la duración correspondiente. Las dos líneas con guiones por encima y debajo del promedio indican 1 y 2 desvíos estándar. La línea vertical a rayas indica el porcentaje de calmas para la Hora 9 a lo largo de las cinco duraciones.
Figura IV.24	Frecuencias de ocurrencia de vientos por dirección según una roseta de vientos de rango completo y la correspondiente roseta de salida de calmas para el verano.
Figura IV.25	Frecuencias de ocurrencia del Sector 1 en distintos sitios y períodos de tiempo y la curva promedio. a) Verano (promedio ponderado total 29.2%) b) Invierno (promedio ponderado total 28.4%).
Figura IV.26	Frecuencias de ocurrencia del Sector 2 en distintos sitios y períodos de tiempo y la curva promedio. a) Verano (promedio ponderado total 29.3 %) b) Invierno (promedio ponderado total 18.6 %).
Figura IV.27	Frecuencias de ocurrencia del Sector 3 en distintos sitios y períodos de tiempo y la curva promedio. a) Verano (promedio ponderado total 63.2 %) b) Invierno (promedio ponderado total 55.7 %).
Figuras de los Anexos del Capítulo IV	
	Anexo IV.1: no contiene figuras. **Anexo IV.2**: no contiene figuras.

	Anexo IV.3: no contiene figuras.
Figuras del Capítulo V	
Figura V.1	Ejemplo de Dendograma.
Figura V.2	Casos particulares de distancias de Minkowsky.
Figura V.3	a) Diagrama cuantil-cuantil correspondiente a las frecuencias de ocurrencia de la dirección N (norte). Eje *X* inferior: valores de los percentiles de la Distribución Normal Estándar (teórica). Eje *X* superior: percentiles expresados como probabilidad. Eje de las *Y*: Valores observados (datos). b) Densidad de distribución para las observaciones (barras) y para la curva teórica ajustada (forma de campana) de la Figura V.3a.
Figura V.4	a) Diagrama cuantil-cuantil correspondiente a las frecuencias de ocurrencia de la dirección ESE (este-sudeste). Eje *X* inferior: valores de los percentiles de la Distribución Normal Estándar (teórica). Eje *X* superior: percentiles expresados como probabilidad. Eje de las *Y*: Valores observados (datos). b) Densidad de distribución para las observaciones (barras azules) y para la curva teórica ajustada (rojo) de la Figura V.4a.
Figura V.5	a) Diagrama cuantil-cuantil correspondiente a las frecuencias de ocurrencia de la dirección O (oeste). Eje *X* inferior: valores de los percentiles de la Distribución Normal Estándar (teórica). Eje *X* superior: percentiles expresados como probabilidad. Eje de las *Y*: Valores observados (datos). b) Densidad de distribución para las observaciones (barras azules) y para la curva teórica ajustada (rojo) de la Figura V.5a.
Figura V.6	Aporte a la varianza total de cada una de las primeras cuatro componentes principales.
Figura V.7	Rosetas horarias expresadas en función de las dos primeras componentes principales.
Figura V.8	Rosetas horarias expresadas en función de las dos últimas componentes principales.
Figura V.9	La línea de rayas (roja) es la distancia del centroide de A-B hasta C trasladado para mostrar que no llega a D. Observar que no se mantiene la estructura anidada (jerarquía indexada).
Figura V.10	Ejemplo tomado de Suggar et al. (1999). a) Datos al azar en el plano b) Curva del W_k en función del número de grupos (gráfico de sedimentación).
Figura V.11	Ejemplo tomado de Tibshirani et al. (2001). a) Datos con estructura de grupo en el plano b) Curva del W_k en función del número de grupos.
Figura V.12	Dendograma de 24 rosetas horarias promedio de vientos correspondiente al Verano en el Punto J para el período 1998- 2003. En el eje de las *X* se halla representada la distancia Euclídea al cuadrado reescalada en % (para facilitar comparaciones con otros dendogramas). En el eje de las *Y* cada "Hora" representa un vector de 16 direcciones de frecuencia de vientos. La línea de trazos vertical cercana a una distancia de corte del 40% indica la solución dada por la mayoría de los criterios aplicados para la determinación del número óptimo de grupos.
Figura V.13	Diagrama de sedimentación para el dendograma de la Figura V.12
Figura V.14	Dendograma de 24 rosetas horarias promedio de vientos correspondiente al Invierno en el Punto J para el período 1998- 2003. En el eje de las *Y* cada "Hora" representa un vector de 16 direcciones de frecuencia de vientos. En el eje de las *X* se halla representada la distancia Euclídea al cuadrado. Los óvalos y sus números indican los sucesivos pasos de aglomeración.
Figura V.15	a) el eje de las *Y* es el *RMSSTD* y el eje de las X son los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14. b) el eje de las *Y* es el *SPR* y el eje de las *X* son los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14. c) el eje de las *Y* es el *RS* y el eje de las *X* son los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14. d) el eje de las *Y* están representadas las *CD* y en el eje de las *X* los pasos (o niveles) de aglomeración correspondientes al dendograma de la Figura V.14.
Figura V.16	Disimilitudes vs. distancias en la configuración.
Figura V.17	Configuración en dos dimensiones. La misma fue obtenida a partir de los coeficientes de correlación de la Tabla V.4 excepto los valores promedios de sitios y estaciones del año. Los ejes (dimensión 1 y dimensión 2) no tienen un significado absoluto sino que reflejan distancias relativas entre los puntos del plano (configuración hallada).

Figura V.18	STRESS (eje *Y*) versus número de dimensión (eje *X*).
Figura V.19	Configuración en tres dimensiones.
Figura V.20	Rosetas horarias promedio de frecuencias de viento por dirección observadas para la estación verano durante el período 1998- 2003 en el Punto J (Figura II.6). Los bloques horarios refieren a la Hora Local, por ejemplo, Hora 0 equivale a 00:00-00:59 Hora Local. Las calmas están expresadas como la cantidad de observaciones menores a 1.6 km h^{-1} respecto del total de observaciones. La velocidad media observada para esta estación durante el período es de 6.6 km h^{-1}.
Figura V.21	Dendograma de 24 rosetas horarias de viento correspondiente al verano en el Punto J para el período 1998- 2003. En el eje de las *X* se halla representada la distancia Euclídea al cuadrado reescalada en % (para facilitar comparaciones con otros dendogramas). En el eje de las *Y* cada "Hora" representa un vector de 16 direcciones de frecuencias de ocurrencias de vientos (Rosetas de la Figura V.20). Las líneas de trazos verticales indican posibles distancias de corte.
Figura V.22	Rosetas de viento promedio de cada grupo formado en el proceso de aglomeración jerárquico dado por la Figura V.21 para una distancia de corte de aprox. 50%. En la designación de cada roseta promedio el número de grupo asignado a cada grupo de horas es arbitrario y solo con fines prácticos.
Figura V.23	Vectores resultantes (de las rosetas de frecuencias de viento promedio por dirección) de grupo para cada estación y sitio de monitoreo. La flecha indica la dirección desde donde sopla el viento. El verano en Punto J se corresponde con las rosetas de la Figura V.22. Los números naturales del 1 al 5 (por ejemplo en "Grupo 1") señalan las cinco etapas en que ha quedado dividido el día a partir de los cinco conglomerados establecidos para cada estación y sitio. Los ejes en línea punteada indican la separación en cuadrantes con un predominio de los vientos en el primero y el cuarto (derecha arriba y abajo respectivamente).
Figura V.24	Salida de EMD. Cada punto del gráfico representa una roseta horaria de vientos de 16 direcciones (correspondiente a una estación del año y un sitio de monitoreo) que ha sido reducida a un punto en el plano aplicando EMD. Los ejes *X* e *Y* están dados en unidades arbitrarias. El número cercano a cada cuadrado o triángulo refiere a la hora del día de la roseta original, algunas etiquetas han sido omitidas por cuestiones de claridad. Las líneas que unen puntos (azules) para el Punto A y (rojas) para el Punto J han sido dibujadas como ayuda para la visualización. a) Veranos 1998- 2003 en los puntos A y J. b) Inviernos 1998- 2003 en los puntos A y J.
Figura V.25	Mapa parcial de la Argentina y países limítrofes. El rectángulo (rojo), que cubre aproximadamente 390 km en longitud y 285 km en latitud, es la zona de la cuenca del Río de La Plata en donde tiene alcance el modelo de predicción de vientos.
Figura V.26	Frecuencias promedio de direcciones de viento observadas entre 1994 y 2008 expresadas en porcentaje: a) Hora 6 y b) Hora 18. El rectángulo interior (rojo) indica la región en que se llevó a cabo el estudio de análisis por conglomerados. Las estaciones meteorológicas, en orden alfabético son: Aeroparque (AER), Carrasco (CAR), Colonia (COL), Ezeiza (EZE), Florida (FLO), La Plata Aero (LPA o Punto K en la Figura II.6- Capítulo II), Martín García (MGA), El Palomar (PAL), Punta Indio (PIN), Prado (PRA), Pontón Recalada (PRE), San Fernando (SFO) y Don Torcuato (TOR). La dirección Norte en los mapas se halla hacia arriba. La velocidad promedio total observada en el rectángulo en estudio para la estación verano fue de 16.2 km h^{-1} (4.5 m s^{-1}) que en la escala Beaufort (Sección III.4- Capítulo III) corresponde a Brisa leve.
Figura V.27	Dendograma para el verano. La columna de números pequeños (solo legibles en formato digital) sobre el eje *Y* refiere a la identificación de cada uno de los 180 vectores en coordenadas arbitrarias, cada uno de ellos se corresponde con un pixel en la Figura V.28. El eje de las *X* representa a la distancia Euclídea al cuadrado que ha sido reescalada respecto de la máxima distancia por lo que aparece en %. Las tres distancias de corte seleccionadas (23, 30 y 48) se hallan identificadas con las líneas verticales a tramos. Para cada una de estas distancias (casos (a), (b) y (c) de la Figura V.28) cada grupo formado se halla identificado con un color según se muestra a la izquierda del eje *Y* ((a'), (b') y (c')).

Figura V.28	El rectángulo interior de la Figura V.25 se muestra dividido en: a) 18 subáreas b) 12 subáreas c) 6 subáreas. Esta división se basa en las tres soluciones adoptadas en el proceso de análisis por conglomerados jerárquicos para la estación verano. Cada uno de los 180 pixeles de la zona de estudio cubre un área aproximada de 22 (horizontal) x 28 (vertical) km x km. Estos píxeles algo rectangulares se aproximan a la forma que da el sistema de coordenadas gaussiano de la superficie terrestre en el rango de latitudes de trabajo. Cada subárea (indicada con un color) reúne un número específico de pixeles según la distancia de corte; (a), (b) o (c) de la Figura V.27.
Figura V.29	Rosetas de viento resultantes (promedios) obtenidas para las tres soluciones adoptadas a partir del análisis por conglomerados. Las frecuencias de direcciones de viento incluyendo las calmas están dadas en porcentaje mientras que las velocidades medias de vientos están dadas en m s^{-1}. Lado izquierdo: rosetas de frecuencias de viento (líneas rojas) que incluyen las calmas (circunferencias azules), ambas expresadas en porcentaje de ocurrencias. El eje *Y* representa la frecuencia porcentual para las direcciones y las calmas. Lado derecho: rosetas de velocidades de viento (líneas verdes) expresadas en m s^{-1} (1 m s^{-1} equivale a 3.6 km h^{-1}) El eje *Y* representa la velocidad promedio para la dirección correspondiente. Cada roseta de vientos es el resultado de promediar los vectores correspondientes a las tres soluciones para el verano. Los rectángulos en color (este último asignado arbitrariamente) designan las subáreas que representan las rosetas involucradas en el mapa de la Figura V.28. a) corresponde a 18 grupos (distancia de corte 23 en la Figura V.27) que se representan en la Figura V.28a. b) corresponde a 12 grupos (distancia de corte 30 en la Figura V.27) que se representan en la Figura V.28b. c) corresponde a 6 grupos (distancia de corte 48 en la Figura V.27) que se representan en la Figura V.28c.
Figura V.30	Los puntos en el plano representan a las rosetas de viento del dendograma de la Figura V.20 expresadas por las dos primeras componentes principales. Las líneas envolventes de trazos indican los grupos determinados por el dendograma para una distancia de corte de alrededor del 50%. La línea continua que envuelve a las horas 18 y 19 indica un posible subgrupo. Ninguna de las líneas envolventes reflejan la forma de los grupos, han sido dibujadas solo con fines ilustrativos para mostrar la estructura de grupo. Los valores sobre el eje de las *X* divididos por $\sqrt{2}$ constituyen el primer término en la ecuación V.4. y el valor constante para cada una de las curvas de la Figura V.32 desde la (a1) hasta la (e1).
Figura V.31	Diagrama de sedimentación. Ayuda a determinar el número de autovalores a retener.
Figura V.32	Curvas de Andrews para las rosetas horarias de la Figura V.20. Cada curva fue construida a partir de las primeras cinco componentes principales empleadas como variables en la ecuación V.4. El eje *X* cubre el intervalo *t* [-180, 180]. El eje *Y* corresponde a *f(t)* (ver ecuación V.4). De (a1) a (e1) son curvas de Andrews individuales. De (b2) a (e2) son curvas promedio de grupo (linea sólida) y curva promedio general (línea de puntos).
Figura V.33	La columna izquierda de esta figura repite la configuración de rosetas de la Figura V.22. La columna derecha introduce los nuevos grupos hallados en concordancia con el dendograma de la Figura V.21 para una distancia de corte de 40%. El Grupo 4 de la Figura V.22 ha dado lugar al Grupo 4* y Grupo 5*.
Figura V.34	Dendograma correspondiente a rosetas de frecuencias horarias de vientos por dirección de la primavera en el Punto J durante el período 1998- 2003. El eje *X* son distancias Euclídeas al cuadrado reescaladas. El dendograma fue obtenido normalizando los datos con media y desvío estándar. La distancia Euclídea al cuadrado es la medida de disimilitud adoptada y el criterio de aglomeración es el UPGMA.
Figura V.35	Dendograma correspondiente a rosetas de frecuencias horarias anuales de vientos por dirección observadas en el Punto A durante el período 1997- 2000. El eje *X* son distancias Euclídeas al cuadrado reescaladas. El dendograma fue obtenido normalizando los datos con media y desvío estándar. La distancia Euclídea al cuadrado es la medida de disimilitud adoptada y el criterio de aglomeración es el UPGMA. Se indica, en línea cortada, la solución adoptada en la publicación de referencia para una distancia de corte de aprox. 24% (solución para 8 grupos).

Figura V.36	El eje de las *X* son las *s(i)* para cada uno de los vectores originales pertenecientes a un grupo. El eje *Y* representa las rosetas horarias y los grupos formados según el dendograma de la Figura V.35 para una distancia de corte de 24%. Los representantes de grupo son: Grupo 1: Hora 2, Grupo 2: Hora 4, Grupo 3: Hora 10, Grupo 4: Hora 13, Grupo 5: Hora 16, Grupo 6: Hora 18, Grupo 7: Hora 20, Grupo 8: Hora 23.
Figura V.37	El eje de las *X* son las *s(i)* para cada uno de los vectores originales pertenecientes a un grupo. El eje *Y* representa las rosetas horarias y los grupos formados según el método de las *k*- medias aplicado a los datos de trabajo de la Figura V.35. Los representantes de grupo son: Grupo 1: Hora 2, Grupo 2: Hora 4, Grupo 3: Hora 10, Grupo 4: Hora 13, Grupo 5: Hora 15, Grupo 6: Hora 18, Grupo 7: Hora 20, Grupo 8: Hora 23.
Figuras de los Anexos del Capítulo V	
Anexo V.1 **Figura 1**	**Anexo V.1** Conjunto de datos y dos posibles formas de agrupamiento. a) puntos en el plano b) agrupamiento elongado c) agrupamiento esferoide.
Anexo V.2	**Anexo V.2** No contiene figuras.
Anexo V.3 **Figura 1**	**Anexo V.3** Dendograma de 24 rosetas horarias promedio de vientos correspondiente al invierno en el Punto J para el período 1998- 2003. El eje de las *Y* cada "Hora" representa un vector de 16 direcciones de frecuencia de vientos. En el eje de las *X* se halla representada la distancia Euclídea al cuadrado. Los óvalos y sus números indican el paso de aglomeración según el esquema de la Tabla 1.
Figura 2	Matrices de distancias involucradas en el cálculo del coeficiente cofenético. a)Fracción de la matriz original de distancias (matriz de un modo). Esta matriz muestra las distancias Euclídeas al cuadrado entre pares de objetos al inicio del procedimiento cuando no se han formado grupos. b) Fracción de la matriz cofenética que resulta de todo el proceso de aglomeración. Esta matriz muestra las distancias Euclídeas al cuadrado (Enlace Promedio) entre pares de objetos (individuos o grupos) "vía el dendograma", o sea, cuando todos los objetos han sido agrupados.
Figura 3	Diagrama tipo- Shepard con enlace promedio. En el eje de las *X* han sido graficadas las distancias Euclídeas al cuadrado de la matriz original. En el eje de las *Y* las correspondientes distancias "vía el dendograma". La repetición de valores en el eje de las *Y* se debe a que la matriz cofenética limita su número de valores a *n-1* tal como lo muestra la Tabla 1 mientras que las distancias en la matriz original son de *n* (*n*-1)/2. Ocurre que para algunos valores distintos de la matriz original existe un solo valor correspondiente en la matriz cofenética. La línea a 45° ha sido trazada como referencia.
Figura 4	Diagrama tipo- Shepard utilizando el criterio del Enlace Simple ("single linkage"). La recta trazada a 45 grados ha sido trazada como referencia permite evidenciar la "contracción" del espacio inducida por este tipo de criterio
Figura 5	Diagrama tipo- Shepard utilizando el criterio del Enlace Completo ("complete linkage"). La recta trazada a 45 grados ha sido trazada como referencia permite evidencia la "expansión" del espacio inducida por este tipo de criterio.
Anexo V.4	**Anexo V.4** No contiene figuras.
Anexo V.5 **Figura 1**	**Anexo V.5** Las doble flechas indican una de las distancias mínimas y una de las distancias máximas posibles en el reloj.

Figura 2	Dendograma correspondiente a rosetas de frecuencias horarias de vientos por dirección de la Primavera Punto J durante el período 1998- 2003 obtenido considerando la restricción de consecutividad de los miembros de cada grupo para 6 grupos. El eje *X* son distancias Euclídeas al cuadrado reescaladas. El dendograma fue obtenido normalizando los datos con media y desvío estándar. La distancia Euclídea al cuadrado es la medida de disimilitud adoptada y el criterio de aglomeración es el UPGMA.
Anexo V.6	**Anexo V.6** No contiene figuras.

Figuras del Capítulo VI
(no contiene)

Indice de Tablas

Indice de Nomenclatura

Nomenclatura del Capítulo I

MAD: desvío absoluto de la mediana

s : desvío estándar

s_C : desvío estándar contaminado

$\bar{x}$: media (promedio aritmético)

$\bar{x}_C$: media contaminada

$y = \beta\, x + \varepsilon$: y (variable respuesta); x (variable explicativa); β (coeficiente de regresión); ε (error).

Nomenclatura del Capítulo II

adim.: adimensional

c es la concentración de una especie química

°C: grados Celsius

CLP: capa límite planetaria

Coordenadas geodésicas (ejemplo): 35°S 58°O se lee 35 grados de latitud sur y 58 grados de longitud oeste.

Δ_i : ancho de banda de un filtro

$\eta_{(\lambda)}$: eficiencia del filtro de luz [Amperes/Watt]

Hab.: habitantes

$I_0(\lambda)$ irradiancia emitida por la fuente de luz

$I(\lambda)$ irradiancia incidente en el detector [Watt cm^{-2}]

Km: kilómetros

Km^2= kilómetros cuadrados

L es la distancia que recorre la luz (camino óptico) [cm]

m: metros

m^3: metros cúbicos

MW: megawatts

mV: milivoltios

nm: nanómetros

®: marca registrada

ppbv: partes por billón (anglosajón) en volumen

ppmv: partes por millón en volumen

$\sigma(\lambda)$ es la sección eficaz del gas que se quiere medir [cm^2/moléculas]

V_i: señal en voltios para un dado conjunto de longitudes de onda

V_{300}: señal en mV centrada en 300 nm

Nomenclatura del Capítulo III

Γ (gamma mayúscula): gradiente adiabático de temperatura

h_r : altura a la que se midió la velocidad observada.

km h^{-1}: kilómetros por hora (velocidad)

P: presión atmosférica

p (exponente): está dado según la rugosidad del terreno y la estabilidad atmosférica dominante (Sección III.7).

$\left(-\frac{dT}{dz}\right)_{ambiental}$: gradiente ambiental de temperatura

$u_{(h_r)}$: velocidad del viento observada a una altura h_r.

$u_{(z)}$: velocidad del viento "corregida" a la altura ***z***.

z : altura a la que se desea obtener la velocidad corregida.

Nomenclatura del Capítulo IV

α es el nivel de significación

$Cov(x, y)$: covarianza en un sistema bivariado

DS($\hat{\mu}$ *)*: desvío estándar de la media

DS(ŷ): desvío estándar del modelo

D^2: distancia generalizada al cuadrado (puede ser Euclídea, de Mahalanobis, etc. según se especifique).

h: submuestra de *n* datos

IC: intervalo de confianza

MCD: mínimo determinante de la matriz de covarianzas (coeficiente de correlación)

μg: microgramos

$\hat{\mu}$ = estima de la media en el eje "y"

n: número de datos

r_i: residuos de regresión (diferencia entre el valor observado y_i y el predicho por el modelo $\hat{y}_i$)

ρ : coeficiente de correlación ("rho" de Pearson)

RVSC: roseta de vientos de salida de calmas

s ó $S_{D:}$ desvío estándar

SAD: suma de los valores absolutos de las diferencias (distancia)

t : "*t"* de Student o tiempo según corresponda

Nomenclatura del Capítulo V

$B_{(k)}$: suma de cuadrados entre grupos

CP: componentes principales

CH(k): índice de Calinski y Harabasz

d_{rs}: distancias en la configuración

δ_{rs}: disimilitudes en los datos originales

EMD: Escalamiento Multidimensional

F(z): función densidad de distribución

H(k): índice de Hartigan

k: número de grupos o número de dimensiones según se especifique

$KL_{(k)}$: Indice de Krzanowski y Lai

S: factor de STRESS estandarizado

S^*: factor de STRESS bruto

$s(i)$: coeficiente de silueta

STRESS: suma estandarizada de los residuos al cuadrado

σ: desvío estándar de la población

W_k: expresión general para designar una suma de cuadrados

Bibliografía

AAPLP (2006) *Análisis Ambiental del Partido de La Plata. Aportes al Ordenamiento Territorial,* Instituto de Geomorfología y Suelos –UNLP y Centro de Investigaciones de Suelos y Aguas de Uso Agropecuario (CISAUA), Provincia de Buenos Aires, Consejo Federal de Inversiones, Municipalidad de La Plata. Obtenido en 2012 de: http://sedici.unlp.edu.ar/handle/10915/27046.

Achad, M. (2015) Aerosoles: efectos sobre la Radiación UV-B y sobre la Calidad de Aire en la Región Central de Argentina, *Tesis Doctoral,* Universidad Nacional de Córdoba, Córdoba.

Afifi, A.A. y Clark, V. (1998) *Computer Aided Multivariate Analysis,* Second Edition, Chapman & Hall, Boca Raton.

Aggarwal, C.C. (2013) *Outlier Analysis,* Springer, New York.

Aggarwal, C.C. y Yu, P.S. (2001) Outlier Detection for High Dimensional Data, In: *Proceedings of the ACM SIGMOD '01 Conference on Management of Data,* New York.

Albritton, D.L. (1994) Atmospheric Chemistry and Global Change: the Scientist's Viewpoint. In: *The Chemistry of the Atmosphere: Its Impact on Global Change.* Ed. Calvert, J. G., IUPAC, Chemistry for the 21st Century, Blackwell Scientific Publications, Oxford.

Allende, D., Romero, G., Cremades, P., Mulena, G., Puliafito, S. (2013) Caracterización horaria y diaria de la concentración del número total de partículas en ambientes urbanos y suburbanos en Mendoza, *Libro de Actas de PROIMCA*, Universidad Tecnológica Nacional. Obtenido de: http://www.utn.edu.ar/secretarias /pp(Memorias)

Allende, D., Pascual Flores, R., Ruggeri, M., Roca, G. y Puliafito, S. (2015) Medición y caracterización de las fuentes de PM_{10}, $PM_{2.5}$ y PM_1 en las áreas urbanas y suburbanas del Gran Mendoza y Gran San Juan, *Libro de Actas de PROIMCA*, Universidad Tecnológica Nacional. Obtenido de: http://www.utn.edu.ar /secretarias /pp(Memorias)

Allison, P.D. (2001) *Missing Data,* Sage Publications, Thousand Oaks, California.

Alvarez Escudero, L. y Alvarez Morales, R. (2001) Climatología del Viento en Casablanca y sus Aplicaciones I. Climatología, *Boletín de la Sociedad Cubana de Meteorología*. Vol. 7 #2, Ciudad de La Habana, Cuba. Obtenido en Noviembre de 2006 de: http://www.met.inf.cu.

Alvarez Escudero, L., Alvarez Morales, R. y Roque Rodriguez, A. (2007) *Climatología del Viento y sus Aplicaciones II,* En: Contribución a la Educación y la Protección Ambiental. Cátedra de Medioambiente. Instituto Superior de Ciencias y Tecnologías Nucleares. Editorial Academia, La Habana, Cuba. ISBN 959-7136-09-0.

Alvarez Morales, R. y Alvarez Escudero, L. (2000) El efecto de acumulación y su influencia en el patrón de dispersión de contaminantes, *Revista Brasileira de Meteorologia,* 15 (A1): 103- 111.

Anderberg, M.R. (1973) *Cluster Analysis for Applications,* Academic Press, New York.

Anderson, H.R., Limb, E.S., Bland, J.M., Ponce de León, A., Strachan, D.P., Bower, J.H. (1995) Health effects of an air pollution episode in London, December 1991, *Thorax,* 50: 1188- 1193.

Andrade, M.I., Scarpati, O.E. (2008) Recent changes in flood risk in the Gran La Plata, Buenos Aires province, Argentina: causes and management strategy, *GeoJournal,* 70 (4): 245- 250.

Andrews, D.F. (1972) Plots of High-Dimensional Data, *Biometrics,* 28: 125- 136.

Arhens, C.D. (2009) *Meteorology Today.* An Introduction to Weather, Climate and Environment, Ninth Edition, Brookes/Cole Cengage Learning, USA.

Arkouli, M., Ulke, A.G., Endlicher, W., Baumbach, G., Schultz, E., Vogt, U. Muller, M.; Dawidowski, L., Faggi, A., Wolf-Benning, U., Scheffknecht, G. (2010) Distribution and temporal behavior of particulate matter over the urban area of Buenos Aires, *Atmospheric Pollution Research,* 1: 1- 8.

Arranz, G., Pereyra, M., Cifuentes, O. (2015) Herramienta de gestión: monitoreo perimetral en tiempo real de emisiones industriales de VCM (Caso Polo Petroquímico de Bahía Blanca), *Libro de Actas de PROIMCA*, Universidad Tecnológica Nacional. Obtenido de: http://www.utn.edu.ar/secretarias/pp(Memorias)

ARS (2015) Astillero Río Santiago, Información obtenida de: http://www.astillero.gba.gov.ar

Arya, P.S. (2001) *Introduction to Micrometeorology*, Second Edition, Academic Press, San Diego.

Avino, P. y Manigrasso, M. (2008) Ten-year measurements of gaseous pollutants in urban air by an open-path analyzer, *Atmospheric Environment,* 42: 4138– 4148.

Ayres, J., Harrison, R. M., Nichols, G. L., Mynard, R. L. (2010) *Environmental Medicine,* Hodder Education an HachetteUK Company, CRC Press, Taylor & Francis Group, LLC, Boca Raton.

Bard, D., Laurent, O., Havard, S., Deguen, S., Pedrono, G.,. Filleul, L., Segala, C., Lefranc, A., Schillinger C., Rivière, E. (2010) *Ambient air pollution, social inequalities and asthma exacerbation in Greater Strasbourg (France) metropolitan area: The PAISA study,* In: Air Pollution by Villanyi, V. (Ed.) Ed. Sciyo, Rijeka, Croatia.

Barnett, V. (2004) *Environmental Statistics - Methods and Applications,* John Wiley and Sons, Chichester.

Barnett, V. y Lewis, T. (1978) *Outliers in Statistical Data,* John Wiley and Sons, Chichester.

Barnett, V. y Lewis, T. (1994) *Outliers in Statistical Data,* Third Edition, John Wiley and Sons., Chichester.

Barros, V., Menéndez, A., Nagy G. (2005a) El Cambio Climático en el Río de La Plata, CIMA Textos del reporte técnico de los proyectos: *Impactos del Cambio Global en las áreas costeras del Río de la Plata y Variabilidad hidroclimática del estuario del Río de la Plata: Influencia humana, ENSO y estado trófico. Proyecto "Assessments of Impacts and Adaptations to Climate Change (AIACC)",* START-TWAS-UNEP.

Barros, V., Menéndez, A., Natenzón, C., Codignotto, J., Kokot, R., Bischoff, S. (2005a) *El cambio climático y la costa argentina del Río de La Plata,* Fundación Ciudad, Buenos Aires.

Bartkowiak, A. y Szustalewicz, A. (1997) The Grand Tour as a Method for Detecting Multivariate Outliers, *Machine Graphics and Vision,* 6: 487- 505.

Basu, S., Davidson, I., Wagstaff, K.L. (2009) *Constrained Clustering -Advances in Algorithms, Theory, and Applications,* Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC, Boca Raton.

Bates, D.V. (1995) The Effects of Air Pollution on Children, *Environmental Health Perspectives,* 103: 49-63.

Baxter, M.J. (1994) *Exploratory Multivariate Analysis in Archaeology,* Edinburgh University Press, Edinburgh.

Beaver, S., Palazoglu, A. (2006) Cluster analysis of hourly wind measurements to reveal synoptic regimes affecting air quality, *Journal of Applied Meteorology and Climatology,* 45:1710–1726.

Behrens, J.T. (1997) Principles and Procedures of Exploratory Data Analysis, *Psychological Methods,* 2 (2): 131- 162.

Bell M.L, Cifuentes, L.A., Davis D.L, Cushing, E., Gusman Telles, A., Gouveia, N. (2011) Environmental health indicators and a case study of air pollution in Latin American cities, *Environmental Research,* 111: 57–66.

Bell M.L, Davis, D.L, Gouveia, N., Borja-Aburto, V.H., Cifuentes, L.A. (2006). The avoidable health effects of air pollution in three Latin American cities: Santiago, São Paulo, and Mexico City, *Environmental Research,* 100 (3):431-40.

Belsley, D.A., Kuh, E., Welsch, R.E. (2004) *Regression Diagnostics. Identifying Influential Data and Sources of Collinearit,* John Wiley and Sons, New Jersey.

Bely, P.I., Christian, C. and Roy, J.R. (2010) *A Question and Answer Guide to Astronomy,* Cambridge University Press, Cambridge.

Bencalá K.E. y Seinfield, J.H. (1976) On Frequency Distributions of Air Pollutant Concentrations. *Atmospheric Environment,* 10: 941- 950.

Ben-Gal, I. (2005) Outlier detection, In: Maimon O. and Rockach L. (Eds.) *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers,* Kluwer Academic Publishers, Dordrecht.

Bennett, C.T., Macdonald, O., Denmead,T., White,I., Melville, M.D. (2004) Natural sulfur dioxide emissions from sulfuric soils, *Atmospheric Environment*, 38: 1473–1480.

Berri, G.J., Sraibman L, Tanco, R., Bertossa, G. (2010) Low-level wind field climatology over the La Plata River region obtained with a mesoscale atmospheric boundary layer model forced with local weather observations, *Journal of Applied Meteorology and Climatology,* 49 (6):1293–1305.

Berthouex P.M. y Brown, L.C. (2002) *Statistics for Environmental Engineers,* Second Edition, CRC Press LLC, Washington, DC.

Bilos, C, Colombo, J.C., Skorupka, C.N., Rodriguez Presa, M.J. (2001) Sources, distribution and variability of airborne trace metals in La Plata City area, Argentina, *Environmental Pollution,* 111: 149- 158.

Blanco, E.E. y Porta, A.A. (2013) *La contaminación atmosférica y la salud de la población en la micro región La Plata, Berisso y Ensenada. Definición de variables e indicadores de gestión en el marco de políticas públicas.* Reporte de la Editorial Universitaria de la Universidad Tecnológica Nacional (UTN – Argentina)**,** Obtenido de: http://www.edutecne.utn.edu.ar/coini_2013/trabajos/COA20_TC.pdf.

Blanco, J.E. y Berri, G.J. (2013) New indices for the spatial validation of plume forecasts with observations of smoke plumes from grassfires, *Atmospheric Environment,* 67: 313- 322.

Boeker, E. y Grondelle, R. van (1995) *Environmental Physics,* John Wiley and Sons, Chichester.

Bogo, H., Negri, R.M., San Román, E. (1999) Continuous measurement of gaseous pollutants in Buenos Aires City, *Atmospheric Environment,* 33: 2587- 2598.

Bonner, R.E. (1964) On some clustering techniques, *International Business Machines Journal of Research and Development,* 8: 22–32.

Borg, I. y Groenen, P.J.F. (2005) *Modern Multidimensional Scaling- Theory and Applications,* Second Edition, Springer, New York.

Borg, I., Groenen, P.J.F., Mair, P. (2013) *Applied Multidimensional Scaling,* Springer, New York.

Borge, R., De la Paz, D., Lumbreras, J., Pérez, J., Vedrenne, M. (2014) Analysis of Contributions to NO2 Ambient Air Quality Levels in Madrid City (Spain) through Modeling. Implications for the Development of Policies and Air Quality Monitoring, *Journal of Geoscience and Environment Protection,* 2 (1): 6-11.

Borque, P., Ruiz, J., Skabar, Y.G., Aldeco, L., Godoy, A., Nicolini, M. (2008) Numeric Simulation of a Real Sea Breeze Event in La Plata River, *XV Congreso Brasileño de Meteorología,* CBMET XV, Agosto de 2008, San Pablo.

Bower, J. (1997) *Ambient Air Quality Monitoring A review paper for the Royal Society of Chemistry,* AEA Technology, National Environmental Technology Centre, Oxfordshire, England.

Box, G., Jenkins, M., Reinsel, G. (2008) *Time Series Analysis: Forecasting & Control,* 3rd Edition, Wiley, New York.

Brereton, R. (1992) *Multivariate pattern recognition in chemometrics,* Elsevier, The Netherlands.

Brewer, G.D. (1999) The challenges of interdisciplinary, *Policy Sciences,* 32: 327- 337.

Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization, *Proceedings of the National Academy of Sciences,* 101(12): 4164–4169.

Butler, R.W., Davies, P.L., Jhun, M. (1993) Asymptotics for the minimum covariance determinant estimator, *The Annals of Statistics,* 21:1385–1400.

CAI (2012) *La Calidad del Aire en América Latina: Una Visión Panorámica.* Clean Air Institute, Autores: Green, J. y Sánchez, S., EUA, Washington D.C. Obtenido en Diciembre de 2013 en: http://www.cleanairinstitute.org/calidaddelaireamericalatina/TransporteyAireLimpio-cai-april2013.pdf

Calinski, T. y Harabasz, J. (1974) A Dendrite Method for Cluster Analysis, *Communications in Statistics,* 3: 1-27.

Caminos, J.A., Enrique, C., Ghirardi, R., Graizaro, A., Rusillo, S.L. y Pacheco, C.G. (2011) *Calidad de Aire en la Ciudad de Santa Fe*. Facultad Regional Santa Fe, Universidad Tecnológica Nacional, Editorial UTN.

Carr, D.B. (1998) Multivariate Graphics, In: Armitage, P. and Colton, T., Eds., *Encyclopedia of Biostatistics,* Wiley, Chichester, 2864-2886.

Carrizo, C., Berger, M. (2010) *Justicia Ambiental: Saberes prácticos para la efectiva vigencia de los derechos ambientales,* Narvaja Editor, ISBN: 978-987-530-104-7, Córdoba.

Carroll, J.D. y Arabie, P. (1980) Multidimensional Scaling, *Annual Review of Psychology,* 31:607-49.

Cator, E.A. y Lopuhaa, H.P. (2010). Asymptotic expansion of the minimum covariance determinant estimators, *Journal of Multivariate Analysis,* 101: 2372-2388.

Cattogio, J.A. (1990) *Fuentes de contaminación atmosférica. Tecnologías de control y sus impactos,* Latinoamérica. Medio Ambiente y Desarrollo, IEIMA (Instituto de Estudios e Investigaciones sobre Medio Ambiente, Bs. As.

Cattogio, J.A., Succar, S.D., Roca, A.F. (1989). Polynuclear aromatic hydrocarbon content of particulate matter suspended in the atmosphere of La Plata, Argentina, *Science of the Total Environment*, 79: 43- 58.

Celemín, H.A. (1984) *Meteorología Práctica,* Instituto Geográfico Militar, Ediciones de Autor, Mar del Plata.

CEP (2015) *Carta Encíclica Laudato Si del Santo Padre Franciscus sobre el cuidado de la Casa Común*, El Vaticano, Ciudad del Vaticano.

CEPAL (2006) Seminario regional: Las oficinas nacionales de estadística frente a los objetivos de desarrollo del milenio: una nueva evaluación. Tema*: "Propuesta de indicadores complementarios para el monitoreo de los objetivos de desarrollo del milenio en América Latina y El Caribe: ODM 7, Garantizar la sostenibilidad del Medio Ambiente",* Comisión Económica para América Latina y el Caribe (CEPAL), Santiago de Chile.

Chae, S.S. y Warde, W.D (2006) Effect of using principal coordinates and principal components on retrieval of clusters, *Computational Statistics & Data Analysis,* 50: 1407 – 1417.

Chagoyen, M., Carmona-Saez, P., Shatkay, P., Hagit, P. Carozo, J.M. (2006) Discovering semantic features in the literature: a foundation for building functional associations, *BMC Bioinformatics*, 7(41):1- 19.

Chambers, J.M., Cleveland, W.S., Kleiner, B. y Tukey, P.A. (1983) *Graphical Methods for Data Analysis,* Wadsworth and Brooks/Cole Publishers Company, California.

Chan, W.W.Y. (2006) *A Survey on Multivariate Data Visualization,* Report of the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong.

Chang, W.C. (1983) On using Principal Components before Separating a Mixture of Two Multivariate Normal Distributions, *Applied Statistics,* 32 (3):267-275.

Chatterjee, S. y Hadi, A.S. (2006) *Regression Analysis by Example,* Fourth Edition, John Wiley and Sons, New Jersey.

Cheng, S. y Lamb, K. (1998) An analysis of winds affecting air pollution concentrations in Hong Kong, *Atmospheric Environment,* 32: 2559- 2567.

Chiu, K.H., Sree, U., Tseng, S.H., Wu, C.H, Lo, J.G. (2005) Differential optical absorption spectrometer measurement of NO_2, SO_2, O_3, HCHO and aromatic volatile organics in ambient air of Kaohsiung Petroleum Refinery in Taiwan, *Atmospheric Environment,* 39: 941–955.

Cifuentes, L.A, Krupnick, A.J, O'Ryan, R., Toman, M.A. (2005). *Urban Air Quality and Human Health in Latin America and the Caribbea,* Organización Panamericana de la Salud, Washington DC.

Clarke, K.R. (1993) Non-parametric multivariate analyses of changes in community structure, Australian *Journal of Ecology,* 18: 117-143.

Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association,* 74: 829-836.

Cleveland, W.S. y Loader, C.R. (1996a) Smoothing by local regression: Principles and methods. In *W. Härdle and M. G. Schimek (Eds.), Statistical Theory and Computational Aspects of Smoothing,* pp. 10-49, Physica-Verlag, Heidelberg.

Cleveland, W.S. y Loader, C.R. (1996b) Rejoinder to Discussion of Smoothing by Local Regression: Principles and Methods, *Statistical Theory and Computational Aspects of Smoothing,* pp. 113-120, Physica-Verlag, Heidelberg.

Cleveland, W.S. y Devlin, S.J. (1988) Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association,* 83: 596-610.

CN (2001) *Censo Nacional 2001 República Argentina,* INDEC (Instituto Nacional de Estadísticas y Censos), Buenos Aires. Obtenido de: http://www.indec.gov.ar

CN (2010) *Censo Nacional 2010 República Argentina,* INDEC (Instituto Nacional de Estadísticas y Censos), Buenos Aires. Obtenido de: http://www.censo2010.indec.gov.ar.

Cochrane A. (2008) *Cities: Urban Worlds. In: An Introduction To Human Geography- Issues For The 21st Century,* Edited by Daniels P., Bradshaw, M., Shaw, D., Sidaway, J. Third Edition, Pearson Education Limited, Prentice- Hall, London.

Cohen, J., Cohen, P., West, S.G. and Aiken, L.S. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* Third Edition, Lawrence Erlbaum Associates Publishers, Mahwah, New Jersey.

Colman Lerner, J.E., Sanchez, E.Y., Sambeth J.E. y Porta A.A. (2012) Characterization and health risk assessment of VOCs in occupational environments in Buenos Aires, Argentina. *Atmospheric Environment,* 55: 440- 447.

Colman Lerner, J.E., Kohajda, T., Aguilar, M.E., Massolo, L.A., Sánchez, E.Y., Porta, A.A., Opitz, P., Wichmann, G., Herbarth, O., Mueller, A. (2014) Improvement of health risk factors after reduction of VOC concentrations in industrial and urban areas, *Environmental Science and Pollution Research,* DOI 10.1007/s11356-014-2904-x.

Colombo, J.C., Landoni, P., Bilos, C. (1999) Sources, distribution and variability of airborne particles and hydrocarbons in La Plata area, Argentina, *Environmental Pollution,* 104: 305- 314.

Cook, R.D. y Weisberg, S. (1999) *Applied Regression Including Computing and Graphics,* John Wiley and Sons, New York.

Corder, G.W. y Foreman, D.I. (2014) *Nonparam*etric Statistic*s, A Step-By-Step Approach*, John Wiley and Sons, New Jersey.

Cosemans, G., Kretzschmar, J., Mensink, C. (2008) Pollutant roses for daily averaged ambient air pollutant concentrations, *Atmospheric Environment,* 42: 6982–6991.

Cowen, M.P. (2010) Viejos problemas en ciudades nuevas. La Plata : agua potable y problemas sanitarios en la época fundacional, *Res Gesta,* 48. Disponible en: http://bibliotecadigital.uca.edu.ar

Cox, T.F. y Cox, M.A. (2001) *Multidimensional Scaling,* Second Edition, Chapman & Hall/CRC, New Jersey.

CPCB (2003) Guidelines for Ambient Air Quality Monitoring, Central Pollution Control Board Ministry of Environment & Forests, India. Disponible en: http://www.cpcb.nic.in

CR (2012) *Estado de la Calidad del Aire del Área Metropolitana de Costa Rica,* Informe Técnico Quinto, Ministerio de Salud de Costa Rica, Ministerio de Ambiente y Energía, Ministerio de Salud, Universidad de Costa Rica y Municipalidad de San José, San José. Obtenido en Noviembre de 2014 en: http://www.inecc.gob.mx

Croux, C. y Haesbroeck, G. (1999) Influence Function and Efficienty of the Minimum Covariance Determinant Scatter Matrix Estimator, *Journal of Multivariate Analysis,* 71: 161-190.

Cuadras, C.M. (1996) *Métodos de Análisis Multivariante,* EUB S.L., Barcelona.

Cuadras, C.M. (2012) *Nuevos Métodos de Análisis Multivariante,* CMC Editiones, Barcelona.

Cunningham, K.M. y Olgivie, J.C. (1972) Evaluation of hierarchical grouping techniques a preliminary study, *Computer Journal,* 15(3):209- 213.

Darby, L.S. (2005) Cluster analysis of surface winds in Houston, TX, and the impact of wind patterns on ozone, *Journal of Applied Meteorology,* 44: 1788–1806.

Dawidowski, L.E. (2016) *Comunicación Privada* con la Lic. Laura Dawidowski (jurado de esta tesis).

Deardorff, J.W. (1984) Upstream diffusion in the convective boundary layer with weak or zero mean wind. In: Fourth joint conference on application of air pollution meteorology, *American Meteorological Society,* Boston, Massachusetts.

Delahaye, J.P. (1997) Matematización del parecido, *Investigación y Ciencia* (Edición Española de Scientific American), 252: 78- 83.

Dicroce, L., Esparza, J., Díscoli, C. y Martini, I. (2010) Evaluación de contrastes urbanos a partir del grado de percepción en patologías urbano-ambientales presentes en el área del gran la plata, Avances en Energías Renovables y Medio Ambiente, Vol. 14 (*Reunión Nacional de ASADES- Asociación Argentina de Energías Renovables y Ambiente*). Obtenido de: http://www.cricyt.edu.ar/asades/

Diez, S., Fonseca, J., Piccioni, M., Britch, J. (2013) Dispersión de PM_{10} generado por el tráfico vehicular en la ciudad universitaria, Córdoba capital, *Libro de Actas de PROIMCA*, Universidad Tecnológica Nacional. Obtenido de: http://www.utn.edu.ar/secretarias /pp(Memorias)

Dimitriadou, E., Dolnicar, S. and Weingessel, A. (2002) An examination of indexes for determining the number of clusters in binary data sets, *Psychometrika,* 67, 137–159.

Díscoli, C.A. y Barbero, D.A. (2001) Insustentabilidad urbano-energética-ambiental. determinación y cuantificación de contaminantes aéreos y sumideros. Avances en Energías Renovables y Medio Ambiente Vol. 5 (*Reunión Nacional de ASADES- Asociación Argentina de Energías Renovables y Ambiente*). Obtenido de: http://www.cricyt.edu.ar/asades/

DLE (2003) *Diccionario de la Lengua Española,* Vigésimo segunda edición, Real Academia Española-Espasa Calpe, S.A., España.

Dragani, W., Martin, P., Simionato C., Campos, M. (2010) Are wind wave heights increasing in south-eastern south American continental shelf between 32 °S and 40°S ?, *Continental Shelf Research,* 30 (5):481-490.

Dudoit, S. y Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biology,* (3)7:1- 27.

Edelstein, H.A. (1999) *Introduction to Data Mining and Knowledge Discovery,* Third Edition, Two Crows Corporation, Potomac, MD.

Edner, H., Ragnarson, P., Spännare, S. and Svanberg, S. (1993) Differential Optical Absorption Spectroscopy (DOAS) system for urban atmospheric pollution monitoring, *Applied Optics,* 32 (3): 327- 332.

ELP (2011) *Estadísticas de La Plata*. Municipalidad de La Plata, Obtenido de: http://www.estadistica.laplata.gov.ar

Emeis, S. (2012) *Wind Energy Meteorology. Atmospheric Physics for Wind Power Generation,* Springer Heidelberg.

Emeis, S., Schäfer, K., Münkel, C. (2008) Surface-based remote sensing of the mixing-layer height - a review, *Meteorologische Zeitschrift,* 17 (5): 621-630.

EPA (1980) *Options for Reducing the Cost of Criteria Pollutant Monitoring,* EPA-450/4-86-014, Environmental Protection Agency, Washington.

EPA (2000) *Meteorological Monitoring Guidance for Regulatory Modeling Applications,* EPA-454/R-99-005, Environmental Protection Agency, Research Triangle Park, NC.

EPA (2006) *Guidance for Data Quality Assessment. Practical- Methods for Data Analysis*, EPA QA/G9, US EPA- EPA/240/B-06/003, Environmental Protection Agency, Washington.

EPA (2008) *Quality Assurance Handbook for Air Pollution Measurement Systems, Volume IV, Meteorological Measurements,* EPA-454/B-08-002, United States Environmental Protection Agency, Washington.

EPA (2009) *Scout 2008 Version 1.0 User Guide,* Second Edition, EPA/600/R-08/038, United States Environmental Protection Agency, Washington.

EPA (2010) *Reference Method for the determination of Sulfur dioxide in the atmosphere (pararosaniline method)*, 40 CFR, Part. 50, Appendix A-2 to Part 50, Environmental Protection Agency, Washington.

EPA (2013) *Quality Assurance Handbook for Air Pollution Measurement Systems, Volume II, Ambient Air Quality Monitoring Program,* EPA-454/B-13-003, United States Environmental Protection Agency, Washington.

EPA (2014) *Basic Air Pollution Meteorology*, SI- 409. http://yosemite.epa.gov/oaqps.

Escobar, G., Camilloni I., Barros, V. (2003) Desplazamiento del anticiclón subtropical del Atlántico Sur y su relación con el cambio de vientos sobre el estuario del Río de la Plata, *X Congreso Latinoamericano e Ibérico de Meteorología (CLIMET) y II Congreso Cubano de Meteorología,* SOMETCUBA y FLISMET, March 2003, La Habana, Cuba.

Escudero, L. F. (1977) *Reconocimiento de Patrones,* Paraninfo, Madrid.

Everitt, B.S., Landau, S., Leese, M. y Stahl, D. (2011*) Cluster Analysis,* Fifth Edition, John Wiley and Sons, Chichester.

FARN (2013) *Informe Ambiental 2013,* Eds. Di Paola, M. E., Sangalli, F., Ragaglia, J., Fundación Ambiente y Recursos Naturales, Buenos Aires.

Farris, J.S. (1969) On the cophenetic correlation coefficient, *Systematic Zoology,* 18: 279- 285.

Fauconnier, C. y Haesbroeck, G. (2009) Outliers Detection with the Minimum Covariance Determinant Estimator in Practice, *Statistical Methodology,* 6 (4) 363-379.

Fenger, J. (1999) Urban Air Quality, *Atmospheric Environment,* 33: 4877- 4900.

Fenger, J. (2009) Air pollution in the last 50 years – From local to global, *Atmospheric Environment,* 43:13–22.

Fensterstock, J.C. y Fraunkhouser, R.K. (1968) *Thanksgiving 1966 Air Pollution Episode in the Eastern United States,* National Air Pollution Control Administration Publication Nº AP-45, Durham, North Carolina.

Ferreira, H.G., Messina, J., Rigolini, J., López Calva, L.F., Lugo, A.M., Vakis, R. (2013*) La movilidad económica y el crecimiento de la clase media en América Latina,* Banco Internacional de Reconstrucción y Fomento- Banco Mundial, Washington.

Figueras, S. y Gargallo, P. (2003) *Análisis Exploratorio de Datos,* Obtenido en Mayo de 2009 de: http://www.5campus.com/leccion/aed

Filzmoser, P. (2004) A multivariate outlier detection method, In: Aivazian, S., Filzmoser, P., Kharin, Y. (eds.) *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, pp. 18–22. Belarusian State University, Minsk.

Filzmoser, P., Serneels, S., Maronna, R. and Van Espen, P.J. (2009) *Multivariate robust techniques,* In: Comprehensive Chemometrics, Eds. Walczak, B., Tauler, R. y Brown, S., 3: 681-722. Elsevier.

Finlayson- Pitts, B.J. y Finlayson- Pitts, J.N. (2000) *Chemistry of the Upper and Lower Atmosphere. Theory, Experiments, and Applications,* First Edition, Academic Press, California, USA.

Fleiss, J.L., y Zubin, J. (1969) On the methods and theory of clustering, *Multivariate Behavior Research,* 4, 235-250.

Fochesatto, G., Lavorato, M., Rosito, C., Quel, E., Guiraldez, A. (1995) Medición de Capa Límite Atmosférica mediante un Lidar, *Actas de la 79na Reunión AFA*, Vol. 7: 254- 256.

Fovell, R.G. y Fovell, M.C. (1993) Climate zones of the conterminous United States defined using cluster analysis, *Journal of Climate,* 6: 2103–2135.

Fox, J. (2000) *Non parametric Simple Regression – Smoothing Scatter Plots,* Sage Publications, Inc, Iowa.

Friedman, H.P. y Rubin, J. (1967) On some invariant criteria for grouping data, *Journal of the American Statistical Association,* 62, 1159–1178.

Friedirch, R. y Reis, S. (2004) *Emissions of Air Pollutants* - Measurements, Calculations and Uncertainties, Springer-Verlag, Heidelberg.

Fujiwara, F., Gómez, D., Faggi, A. (2013) Perfiles químicos y patrones espaciales del polvo de la calle colectado en la megaciudad de Buenos Aires, *Libro de Actas de PROIMCA*, Universidad Tecnológica Nacional. Obtenido de: http://www.utn.edu.ar/secretarias /pp(Memorias)

Gan, G., Ma, C., and Wu, J. (2007) *Data Clustering: Theory, Algorithms, and Application,* ASA-SIAM, Philadelphia.

Garcia-Huidobro, T., Marshall, F.M. Bell, J.N.B. (2001) A risk assessment of potential agricultural losses due to ambient SO2 in the central regions of Chile, *Atmospheric Environment,* 35: 4903–4915.

García-Osorio, C. y Fyfe, C. (2005) The Combined Use of Self-Organizing Maps and Andrews' Curves, *International Journal of Neural Systems,* 15: 197-206.

Garratt, J. R. (1992) *The Atmospheric Boundary Layer,* Cambridge University Press, New York.

Gasper, R., Blohm, A., Ruth, M. (2011) Social and economic impacts of climate change on the urban environment, *Current Opinion in Environmental Sustainability*, 3:150–157. Elsevier.

Gassmann, M.I. (1998) *Potencial de contaminación atmosférica en la República Argentina*, Tesis Doctoral, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires.

Gassmann, M.I. y Mazzeo, N.A. (2000) Air pollution Potential: Regional Study in Argentina, *Environmental Management,* 25 (4) :375-382.

Gassmann, M.I., Pérez, C.F. y Gardiol, J.M. (2002) Sea-land breeze in a coastal city and its effect on pollen transport, *International Journal of Biometeorology,* Vol. 46, 118-125.

Gigerenzer, G., Todd, P. y the ABC Research Group (1999) *Simple Heuristics That Make Us Smart,* Oxford University Press, Inc.

Gilbert, R.O. (1987) *Statistical Methods for Environmental Pollution Monitoring,* John Wiley and Sons. New York, New York.

Gnanadesikan, R. (1997) *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley and Sons, New York.

Gnanadesikan, R. y Kettenring, J.R. (1972) Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data, *Biometrics,* Special Multivariate Issue, 28 (1): 81-124.

Gnanadesikan, R., Kettenring, J.R. and Landwehr, J.M. (1977) Interpreting and assessing the results of cluster analyses, in Bulletin of the International Statistical Institute: *Proceedings of the 41st Session (New Delhi)* Book 2, 451–463. ISI, Voorburg, Netherlands.

Gnanadesikan, R., Kettenring, J.R. and Tsao, S.L. (1995) Weighting and selection of variables, *Journal of Classification,* 12, 113–136.

Godish, T. 1997. *Air Quality,* 3rd Edition Lewis Publishers, Boca Raton.

Godish, T. 2004. *Air Quality,* 4th Edition Lewis Publishers, Boca Raton.

Gong, X. y Richman, M.B. (1995) On the application of cluster analysis to growing season precipitation data in North America East of the Rockies, *Journal of Climate,* 8: 897- 931.

Gordon , A. (1999) *Classification,* Second Edition . London, UK : Chapman and Hall/CRC Press.

Gorunescu, F. (2011) *Data Mining. Concepts, Models and Techniques,* Springer-Verlag Berlin.

Gower, J.C (1966) Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika,* 53, 325-338.

Goyal, P. (2002) Effect of winds on SO_2 and SPM concentrations in Delhi, *Atmospheric Environment,* 36, 2925–2930.

Goyal, P. y Rama Krishna, T.V.B.P.S. (2002) Dispersion of pollutants in convective low wind: a case study of Delhi, *Atmospheric Environment,* 36: 2071–2079.

Graedel, T.E. (1994) Effects of Emissions to the Atmosphere on Materials and Cultural Artefacts. In: *The Chemistry of the Atmosphere: Its Impact on Global Change.* Ed. Calvert, J. G., IUPAC, Chemistry for the 21st Century, Blackwell Scientific Publications, Oxford.

Grubbs, F.E. (1969) Procedures for detecting outlying observations in samples, *Technometrics,* 11:1- 21.

Gurjar, B.R., Butler, T. M., Lawrence, M.G., Lelieveld, J. (2008) Evaluation of emissions and air quality in megacities, *Atmospheric Environment,* 42: 1593–1606.

Guthe, M., Borodin, P., Klein, R. (2005) Fast and Accurate Hausdorff Distance Calculation between Meshes, *The Journal of WSCG (recently* Winter School of Computer Graphics- Presently International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision*),* 13: 41- 48.

Hair, J.F., Black, W. C., Babin, B.J. and Anderson, R.E. (2010) *Multivariate Data Analysis*, Seventh Edition, Prentice Hall, Upper Saddle River, New York.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) On Clustering Validation Techniques, *Journal of Intelligent Information Systems,* 17 (2-3), 107-145.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2002a) Cluster Validity Methods: Part I. *Proceedings of the ACM SIGMOD Conference,* 31 (2): 40- 45.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2002b) Clustering Validity Checking Methods: Part II. *Proceedings of the ACM SIGMOD Conference,* 31 (3): 19- 27.

Hall, J.V., Brajer, V., Lurmann, F.W. (2010) Air pollution, health and economic benefits- Lessons from 20 years of analysis, *Ecological Economics,* 69: 2590–2597.

Hand, D., Mannila, H., Smyth, P. (2001) *Principles of Data Mining,* Massachusetts Institute of Technology, London.

Härdle, W. (1994) *Applied Nonparametric Regression,* Oxford University Press, Oxford.

Härdle, W. y Mammen, E. (1993) Comparing nonparametric vs. parametric regression fits, *The Annals of Statistics,* 21 (4): 1926- 1947.

Harlan, S.L. y Ruddel, D.M. (2011) Climate change and health in cities: impacts of heat and air pollution and potential co-benefits from mitigation and adaptation, *Current Opinion in Environmental Sustainability,* 3: 126–134.

Hartigan, J.A. (1975) *Clustering Algorithms,* John Wiley and Sons, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2011) *The Elements of Statistical Learning - Data Mining, Inference and Prediction,* Second Edition, Springer, New York.

Hawkins, D. (1980) *Identification of Outliers,* Chapman and Hall, New York.

Hay, W.W., Soeding, E., DeConto, R., Wold, C.N. (2002) The Late Cenozoic uplift – climate change paradox, *International Journal of Earth Sciencies,* 91:746–774.

Henry, R.C., Chang, Y.S., Spiegelman, C.H. (2002) Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction, *Atmospheric Environment,* 36: 2237–2244.

Hoaglin, D., Mosteller, F., Tukey, J. (1983) *Understanding Robust and Exploratory Data Analysis,* John Wiley and Sons, New York.

Hodge, V.J. y Austin, J. (2004) A survey of outlier detection methodologies, *Artificial Intelligence Review,* 22 (2):85-126.

Holland, D.M, Caragea, P., Smith, R.L. (2004) Regional trends in rural sulfur concentrations, *Atmospheric Environment,* 38 (2004) 1673–1684.

Holmes, D.E. y Jain, L.C. (2012) *Data Mining: Foundations and Intelligent Paradigms,* Springer-Verlag, Berlin.

Holzworth, G.C. (1967) Mixing depths, wind speeds and air pollution potential for selected locations in the United States, *Journal of Applied Meteorology,* 6: 1039-1044.

Hubert, M., Rousseeuw, P.J., Verdonck, T. (2012) A Deterministic Algorithm for Robust Location and Scatter, *Journal of Computational and Graphical Statistics,* 21(3): 618-637.

Husson, F., Lê, S., Pagès, J. (2011) *Exploratory Multivariate Analysis by Example Using R,* CRC Press Taylor & Francis Group, Boca Raton.

Huth, R., Memesova, I. and Klimperova, N. (1993) Weather categorization based on the average linkage clustering technique: an application to European mid- latitudes, *International Journal of Climatology,* 13: 817- 835

IAA (2006) *Informe Anual Ambiental 2006. Ciudad Autónoma de Buenos Aires.* Ley N° 303 de Información Ambiental Decreto N° 1325/06. Obtenido en Abril de 2016 de http://www.buenosaires.gob.ar/areas/med_ambiente

IACA (2011) *Informe de Calidad de Aire- Informe Anual, Montevideo.* Servicio Evaluación de la Calidad y Control Ambiental, Departamento de Desarrollo Ambiental, Intendencia de Montevideo, Uruguay.

IPA (1999). *La República Argentina y su Industria Petroquímica*. Special edition of the Argentinean Petrochemical Institute. Obtenido en Mayo de 2009 de: http://ipa.org.ar/publicaciones-a.htm

IPA (2011). *Perfiles de empresas productoras del sector petroquímico*. Obtenido en Marzo de 2013 de: http://ipa.org.ar/publicaciones-a.htm

Jaakkola J.J., Partti-Pellinen K., Marttila O., Miettinen P., Vilkka V., Haahtela T. (1999) The South Karelia Air Pollution Study: changes in respiratory health in relation to emission reduction of malodorous sulfur compounds from pulp mills, *Archives of Environmental Health,* 54: 254–263.

Jackson, I.J. y Weinand, H. (1995) Classification of tropical rainfall stations: a comparison of clustering techniques, *International Journal of Climatology,* 15: 985-994.

Jacob, D.J. y Winner, D.A. (2009) Effect of climate change on air quality, *Atmospheric Environment,* 43:51–63.

Jacobson, M.Z. (2002) *Atmospheric Pollution - History, Science and Regulation,* Cambridge University Press, New York.

Jacobson, M.Z. (2005) *Fundamentals of Atmospheric Modeling,* Second Edition, Cambridge University Press, Cambridge.

Jacoby, W.G. (1998) *Statistical graphics for visualizing multivariate data,* Sage University Papers Series on Quantitative Applications in the Social Sciences, Series Nº 07-120, Sage Publications, Inc.. Thousand Oaks, California.

Jain, A. y Dubes, R. (1988) *Algorithms for clustering data,* Englewood Cliffs, Prentice Hall, New York.

Jain, A.K., Murty, M.N. and Flynn, P.J. (2000) *Data Clustering: A Review,* ACM, Inc.

Jajuga, K. y Walesiak, M. (2000) Standardisation of data set under different measurement scales. In: *Classification and Information Processing at the Turn of the Millennium (R. Decker and W. Gaul, eds.)* 105–112 Springer-Verlag, Heidelberg.

Jardine, N. y Sibson, R. (1968) The construction of hierarchic and non-hierarchic classifications, *The Computer Journal,* 11 (2): 177-184. Obtenido en Noviembre de 2011 de: http://comjnl.oxfordjournals.org/

Jedrychowski, W., Flak, E. y Mróz, E. (1999) The Adverse Effect of Low Levels of Ambient Air Pollutants on Lung Function Growth in Preadolescent Children, *Environmental Health Perspectives,* 107 (8):669- 674.

Jimenez, P.A., Gonzalez-Rouco, J.F., Montalvez, J.P., Navarro, J., García- Bustamante, E. y Valero, F. (2008) Surface Wind Regionalization in Complex Terrain, *Journal of Applied Meteorology and Climatology,* 47:308- 325.

Jollife , I. (2002). *Principal component analysis,* Springer-Verlag, New York.

Jolliffe, I.T., Jones, B. and Morgan, B.J.T. (1986) Comparison of Cluster Analyses of the English Personal Social Services Authorities, *Journal of the Royal Statistical Society Series A,* 149, 253-270.

Kalkstein L.S., Tan, G., Skindlov J.A. (1987) An evaluation of three clustering procedures for use in synoptic climatological classification, *Journal of Climate and Applied Meteorolology,* 26: 717–730.

Kaufman, L. y Rousseeuw, P. J. (2005) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc., Hoboken, NJ.

Kaufmann, P. y Whiteman, C. D. (1999) Cluster-analysis classification of wintertime wind patterns in the Grand Canyon region*, Journal of Applied Meteorology,* 38, 1131–1147.

Kenkel, N.C. y Orlóci, L. (1986) Applying metric and nonmetric multidimensional scaling to ecological studies: some new results, *Ecology,* 67 (4): 919- 928.

Kettenring, J.R. (2006) The Practice of Cluster Analysis, *Journal of Classification,* 23(1):3- 30.

Khattree, R. y Naik, D.N. (2000) *Multivariate Data Reduction and Discrimination with SAS software,* John Wiley & Sons and SAS Institute, North Carolina, USA.

Kim, K.H. y Kim, M.Y. (2001) Comparison of an open path differential optical absorption spectroscopy system and a conventional in situ monitoring system on the basis of long-term measurements of SO2, NO2, and O3, *Atmospheric Environment,* 35: 4059–407.

Kim, S.T., Maedab, Y., Tsujino, Y. (2004) Assessment of the effect of air pollution on material damages, *Atmospheric Environment,* 38: 37- 48.

Kondrashov, D. y Ghil, M. (2006) Spatio-temporal filling of missing points in geophysical data sets, *Nonlinear Processes in Geophysics,* 13, 151–159.

Kork, M. y Sáenz, (1999) *Monitoreo de la calidad del aire en América Latina,* Programa de Control de Contaminación del Aire, CEPIS- OPS, Lima.

Kourtidis, K., Ziomas, I., Zerefos, C. Gousopoulos, A., Balis, D., Tzoumaka, P. (2000) Benzene and toluene levels measured with a commercial DOAS system in Thessaloniki, Greece, *Atmospheric Environment,* 34: 1471- 1480.

Kraas, F., Aggarwal, S., Coy, M., Mertins, G. (2014) *Megacities Our Global Urban Future,* Springer, Heidelberg.

Krämer U., Behrendt, H., Dolgner, R., Ranft, U., Ring, J., Willer, H. and Schlipköter, H.W. (1999) Airway diseases and allergies in East and West German children during the first 5 years after reunification. Time trends and the impact of sulfur dioxide and total suspended particles*, International Journal of Epidemiology,* 28(5):865–873.

Krämer, A., Khan, M.H. and Kraas, F. (2011) *Health in Megacities and Urban Areas,* Springer, Heidelberg.

Kruijt, D. y Koonings, K. (2009) The rise of megacities and the urbanization of informality, exclusion and violence. In: *Megacities: The politics of urban exclusion and violence in the global South,* (Koonings and Kruijt Eds.), Zed Books, London.

Kruskal, J.B. (1964a) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika,* 29:1- 28.

Kruskal, J.B. (1964b) Nonmetric multidimensional scaling: A numerical method, *Psychometrika,* 29:115-129.

Kruskal, J.B. y Wish, M. (1978) *Multidimensional Scaling,* Sage Publications, Inc., California.

Krzanowski, W.J. (2007) *Statistical Principles and Techniques in Scientific and Social Investigations*, Oxford University Press Inc., New York.

Krzanowski, W.J. y Lai, Y.T. (1988) A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering, *Biometrics,* 44 (1): 23-34.

LAQN (2015) *London Air Quality Network. Summary Report 2013,* Environmental Research Group, Kings College of London, London. Obtenido en Febrero de 2015 de: http://www.londonair.org.uk

Landsberg, H.E. (1981) *The urban climate,* Academic Press, New York.

Lavine, B.K. (2000) Clustering and classification of analytical data. In: *Encyclopaedia of Analytical Chemistry,* pp. 1- 21, Ed. R.E. Meyers, John Wiley and Sons, Chichester.

Lavorato, M., Cesarano, P., Pagura, M., Quel, E., Dworniczac, J.C., Flamant, P.H. (2002) Observación simultánea de parámetros atmosféricos con el Lidar Dual que opera en Buenos Aires con un nuevo sistema de detección, *Actas de la 86va Reunión AFA,* Vol. 14: 281- 283.

Lazaridis, M. (2011) *First Principles of Meteorology and Air Pollution,* Springer, Dordrecht.

Lazarsfeld, P.F. y Reitz, J.G. (1970) Toward a Theory of Applied Sociology, Report AD 715639, *Bureau of Applied Social Research,* Columbia University, New York.

Lebel, J. (2005) *Salud. Un enfoque ecosistémico,* Centro Internacional de Investigaciones para el Desarrollo. Ed. Alfaomega, Ottawa, Canadá.

Lee, C., Choi, I.J., Jung, J.S., Lee, J.S., Kim, K.H., Kim, Y.J. (2005) Measurement of atmospheric monoaromatic hydrocarbons using differentialopticalabsorption spectroscopy: Comparison with on-line gas chromatography measurements in urban air, *Atmospheric Environment,* 39: 2225–2234.

Legendre, P. y Legendre, L. (1998) *Numerical Ecology,* Second English Edition, Elsevier, Amsterdam.

Lesniok, M. (2011) Changeability of Air Pollution in Katowice Region (Central Europe, Southern Poland), In: *Advanced Air Pollution,* Chapter 10, Ed. Nejadkoorki, F.- InTech, Croatia.

Linares, G. (2001) Escalamiento Multidimensional: conceptos y enfoques, *Investigación Operativa,* 22 (2): 173- 183.

Ling, H., Schäfer, K., Xin, J.,Qin, M., Suppan, P., Wang, Y. (2014) Small-scale spatial variations of gaseous air pollutants e A comparison of path-integrated and in situ measurement methods, *Atmospheric Environment,* 92: 566- 575.

Lioy, P.J. (1990) Assessing total human exposure to contaminants, *Environmental Science and Technology,* 24, (7): 948- 945.

Lioy, P.J. (2006) Employing dynamical and chemical rocesses for contaminant mixtures outdoors to the indoor environment:The implications for total human exposure analysis and revention, *Journal of Exposure Science and Environmental Epidemiology,* 16: 207 –224.

Little, R.J.A y Rubin, D. B. (1987) *Statistical Analysis with Missing Data,* John Wiley and Sons, Chichester.

Loader, C. (1999) *Local Regression and Likelihood,* Springer, New York.

Lorr, M. (1983) *Cluster Analysis for Social Scientists,* The Jossey-Bass Social and Behavioral. San Francisco.

Lutgens, F. K. y Tarbuck, E. J. (2013) *The atmosphere: An Introduction to Meteorology,* 12th Edition, Pearson Inc., New York.

Lyall, C., Bruce, A., Tait, J., Meagher, L. (2011) *Interdisciplinary Research Journeys. Practical Strategies for Capturing Creativity,* 1st Edition, Bloomsbury Academic, London.

Macdonald, B.C.T., Denmead, O.T., White, I., Melville, M.D. (2004) Natural sulfur dioxide emissions from sulfuric soils, *Atmospheric Environment,* 38: 1473–1480.

Macedo I.M., Pereira Masi, B., Rosental Zalmon, L.l. (2006) Comparison of rocky intertidal community sampling methods at the northern coast of Rio de Janeiro state, Brazil, *Brazilian Journal of Oceanography,* 54(2/3):147–154.

Maddala, G.S. y Rao, C.R. (1997) *Handbook of Statistics Vol. 15*, Elsevier, Amsterdam.

Mahalanobis, P. C. (1936) On the generalized distance in statistics, *Proceedings of the National Institute of Science India,* Vol. II, N° 1, (12) 49–55.

Marañon Di Leo, J., Del Nero, S., Ragaini, J.C., Sacchetto, V., Colosqui, J., Colman, J., Boldes, U., Scarabino, A., Rosato, M., Reyna Almandos, J. (2004). Air Concentrations of SO_2 and Wind Turbulence near La Plata Petrochemical Pole (Argentina), *Latin American Applied Research,* 34: 55- 58.

Maronna (CP) *Comunicaciones Privadas* con el Dr. Ricardo Maronna.

Maronna, R. (1976) Robust M-estimators of multivariate location and scatter, *Annals of Statistics,* 4: 51-67.

Maronna, R., Martin, R., Yohai, V. (2006). *Robust Statistics. Theory and Methods,* John Wiley and Sons Ltd. London.

Maronna, R. y Yohai, V. (2014) High finite-sample efficiency and robustness based on distance-constrained maximum likelihood, *Computational Statistics and Data Analysis,* en prensa.

Marques de Sá, J.P. (2007) *Applied Statistics Using SPSS, STATISTICA, MATLAB and R.* Springer, Heidelberg.

Martinez, A.P. y Romieu, I. (1997) *Introducción al Monitoreo Atmosférico,* OPS/OMS, ECO- GTZ, Departamento del Distrito Federal de México, Ciudad de México.

Massolo, L., Müller, A, Tueros, M., Rehwagen, M., Frank, U., Ronco, A., Herbarth, O. (2002) Assessment of Mutagenicity and Toxicity of Different-Size Fractions of Air articulates from La Plata, Argentina, and Leipzig, Germany, *Environmental Toxicology,* 17: 219- 231.

Massolo, L., Rehwagen, M., Porta, A., Ronco A., Herbarth, O., Mueller, A. (2010) Indoor-outdoor distribution and risk assessment of volatile organic compounds in the atmosphere of industrial and urban areas, *Environmental Toxicology,* 25 (4):339-49.

Markatou, M. y Ronchetti, E. (1997) Robust Inference: the approach based on influence functions. In: *Handbook of Statistics,* Vol. 15, Maddala, G.S. and Rao, C.R. Eds., Elsevier, Amsterdam.

Mattio, C.A. (2009) Combinación de herramientas para el monitoreo y seguimiento de humo generado por incendios forestales y de pastizales en la República Argentina. *Reprints X Congreso Argentino de Meteorología,* Octubre 2009, Buenos Aires.

Mazzeo, N.A.; Nicolini, M.; Moledo, L.; Micheloni, R. (1971) Condiciones de Estabilidad Atmosférica y Capacidad de Dilución Vertical de Contaminantes en la Ciudad de La Plata. AIDIS, Buenos Aires pp. 101-114.

Mazzeo, N.A., Nicolini, M., Müler, C., Micheloni, R. (1974) Algunos aspectos climatológicos de la contaminación atmosférica en el área de La Plata (Prov. de Buenos Aires). *Meteorológica*, 3: 99- 134. Obtenido en 2006 de: http://www.cenamet.org.ar

Mazzeo, N.A. y Nicolini, M. (1974) Eficiencia de las dispersión atmosférica en la zona de La Plata (Prov. de Buenos Aires), *Meteorológica,* 5: 33- 43. Obtenido en 2006 de: http://www.cenamet.org.ar

Mazzeo, N.A. y Venegas, L.E. (1999) Atmospheric stagnation, recirculation and ventilation potential of several sites in Argentina, *Atmospheric Research,* 52: 43–57.

Mazzeo, N.A., Venegas L.E., Choren, H. (2005) Analysis of NO, NO_2, O_3 and NO_x concentrations measured at a green area of Buenos Aires City during wintertime, *Atmospheric Environment,* 39: 3055- 3068.

McCormik, R.A. (1968) Air Pollution Climatology. In: *Air Pollution* (Stern, A.) Vol. 1, Chapter 9 Second Edition, New York Academic Press, New York.

McCune, B. y Grace, J.B. (2002) *Analysis of Ecological Communities,* MjM Software Design Ed., Oregon.

McGreggor, G.R. (1999) Basic Meteorology, In: *Air Pollution and Health,* Eds. Holgate, S. T, Samet, J. M., Koren, H. S. y Maynard, R. L., Academic Press, San Diego.

McKnight, P.E., McKnight, K.M., Sidani, S. and Figueredo, A.J. (2007) *Missing Data: a gentle approach*, The Guilford Press, New York.

Milligan, G.W. (1980) An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, *Psychometrika,* 45: 325- 342

Milligan, G.W. y Cooper, M.C. (1985) An examination of procedures for determining the number of clusters in a data set, *Psychometrika,* 50: 159–179.

Milligan, G.W. y Cooper, M.C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavioral Research,* 21: 41–58.

Milligan, G.W. y Cooper, M.C. (1988) A study of standardization of variables in cluster analysis, *Journal of Classification,* 5, 181–204.

Ministerio de Salud (2012*) Natalidad, mortalidad general, infantil y materna por lugar de residencia.* Boletín Nro. 134, Sistema de Estadísticas e Información de la Salud, Ministerio de Salud de la Nación, Buenos Aires.

Miranda, J.J. (2006). *Impacto Económico en la Salud por Contaminación del Aire en Lima Metropolitana,* Programa de Investigaciones ACDI, IDRC (International Development Research Centre), Consorcio de Investigación Económica y Social (CIES), Instituto de Estudios Peruanos. Obtenido en Diciembre de 2014 de: http://redpeia.minam.gob.pe/

Mirkin, B.G. (2005) *Clustering for data mining: a data recovery approach*, Taylor & Francis Group, LLC, London.

Mirkin, B.G. (2011) *Core Concepts in Data Analysis: Summarization, Correlation and Visualization,* Springer-Verlag London Limited, London.

MLP- UNLP (2001) *Observatorio de Calidad de Vida La Plata. Diagnóstico de Calidad de Vida en el Partido de La Plata,* Municipalidad de La Plata y Universidad Nacional de La Plata, La Plata. Disponible solo en formato impreso en: Biblioteca Pública- Universidad Nacional de La Plata, Plaza Rocha 137, http://biblio.unlp.edu.ar.

Mölders, N. (2012) Land Use and Land Cover Changes - Impact on Climate and Air Quality, Springer, New York.

Mooi, E. y Sarstedt, M. A. (2011) *Concise Guide to Market Research- The Process, Data and Methods Using IBM SPSS Statistics,* Springer, Heidelberg.

Moore, D.J. (1969) The distributions of surface concentrations of sulphur dioxide emitted from tall chimneys, *Transactions of the Royal Society,* 265.

Motta-Garcia, J.R., Vieira-Monteiro, A.M., Duarte-Coelho dos Santos, R. (2012) Visual Data Mining for Identification of Patterns and Outliers in Weather Stations' Data, *XII Workshop de Computação Aplicada – WORCAP*, São José dos Campos, Brasil.

Moustafa, R.E. (2011) Andrews' Curves, *Computational Statistics,* 3: 373-382.

Mu, Y. y Mu, X. (2013) Energy conservation in the Earth's crust and climate change, *Journal of the Air & Waste Management Association,* 63(2): 150–160.

Necco, G.V. (1980) *Curso de cinemática y dinámica de la atmósfera*, EUDEBA, Ediciones Previas, Bs. As.

Negrin, M.; Del Panno, T.; Ronco, A. (2007) Study of bioaerosols and site influence in the La Plata area (Argentina) using conventional and DNA (fingerprint) based methods, *Aerobiologia,* 23: 249–258.

NIST (2012) *Engineering Statistics Handbook*, NIST- Sematech. Obtenido de: http://www.itl.nist.gov/div898/handbook

Nitiu D.S. (2006) Aeropalynologic analysis of La Plata City (Argentina) during 3-year period, *Aerobiologia,* 22: 79- 87.

NU (2009) *HOME* (Cine documental dirigido por Yann Arthus- Bertrand y producido por Luc Besson y la participación de Naciones Unidas, Nueva York).

NU (2013) *Objetivos de Desarrollo del Milenio.* Informe Anual 2013, Naciones Unidas, Nueva York.

Oke, T.R. (1987) *Boundary Layer Climates,* 2nd Edition, Routledge, London.

Olcese, L.E. y Toselli, B.M. (2002) Some aspects of air pollution in Córdoba, Argentina, *Atmospheric Environment,* 36 : 299–306.

OMS (2006) *Guías de calidad del aire de la OMS relativas al material particulado, el ozono, el dióxido de nitrógeno y el dióxido de azufre,* Actualización mundial 2005. WHO/SDE/PHE/OEH/06.02, Ginebra.

OPS (2005) *Evaluación de los efectos de la contaminación del aire en la salud de América Latina y el Caribe.* ISBN 92 75 32598 7, Organización Panamericana de la Salud, Washington D.C.

Orte, M.A. (2011) *Estudio y análisis de la contaminación atmosférica mediante técnicas físicas y químicas en los alrededores del Polo Petroquímico de La Plata,* Tesina de Grado de la carrera de Licenciatura en

Tecnología Ambiental, Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil.

Orte, M.A., Coman Lerner, J., Gutiérrez, M., Elordi, L., Matamoros, N., Reyna Almandos, J., Porta, A. (2015) Estudio de hidrocarburos aromáticos policíclicos asociados al material particulado y en fase gaseosa en la ciudad de La Plata y alrededores, *Libro de Actas de PROIMCA*, Universidad Tecnológica Nacional. Obtenido de: http://www.utn.edu.ar/secretarias /pp(Memorias)

Ortega Dato, J.F. (2001) *Notas sobre estadística robusta*, Documentos de Trabajo de la Facultad de CC. Económicas y Empresariales de Albacete, Universidad de Castilla-La Mancha (España). Obtenido en 2009 de: http://uclm.es/ab/fcee/documentostrabajo.html

Orton, P.M., McGillis, W.R. and Zappa, C.J. (2010) Sea breeze forcing of estuary turbulence and air-water CO_2 exchange, *Geophysical Research Letters,* 37: L13603.

Otero, L.A., Ristori, P.R., Dworniczak, J., Vilar, O., Quel, E.J. (2002) Nuevo sistema lidar de seis longitudes de onda en el CEILAP, *Actas de la 86ta Reunión AFA,* Vol. 18: 282- 285.

Otero, L.A., Ristori, P.R., Pawelko, E.E., Pallota, J.V., Quel, E.J. (2011) Six-year evolution of multiwavelength lidar system at CEILAP, Special Section: *V Workshop on Lidar Measurements in Latin America, Optica Pura y Aplicada,* 44 (1): 13-18.

Otero, L.A., Ristori, P.R., Pallota, J.V., Pawelko, E.E., D´Elia, R. y Quel, E.J. (2012) Volcán Puyehue-Cordón Caulle: medición de las cenizas en Buenos Aires, Argentina, durante junio 2011, Pyroclastic Flow, *Journal of Geology,* (2) 2: 11- 17.

Palmer, C.L. (2001) *Work at the Boundaries of Science.* Information and Interdisciplinary Research Process, Springer, Dorderecht.

Pande, S.R., Sambare, S.S. Thakre, V.M. (2012) Data Clustering Using Data Mining Techniques, *International Journal of Advanced Research in Computer and Communication Engineering,* 1(8):494- 499.

PAR (2012) *Plan Ambiental de Rosario.* Calidad de Aire y Ruido, Municipalidad de Rosario, Santa Fe. Obtenido en Diciembre de 2014 de: http://www.rosario.gov.ar/sitio/.

Peña, D. (2002) *Análisis de Datos Multivariantes,* McGraw Hill- Interamericana de España, S.A.U., España.

Perevochtchikova, M. (2009) La situación actual del sistema de monitoreo ambiental en la Zona Metropolitana de la Ciudad de México, *Estudios Demográficos y Urbanos,* 24 (3):513-547. Obtenido en Octubre de 2014 de: http://www.redalyc.org

Petcheneshsky, T., Gravarotto, M. C., Benitez, R., De Titto, E. (1998) *Gestión de la Calidad de Aire Urbano-Industrial. Situación del Monitoreo de la Calidad del Aire (GEMS- AIRE) en la República Argentina.* Departamento de Salud Ambiental del Ministerio de Salud y Acción Social de La Nación, AIDIS, Buenos Aires (1- 12).

Piegorsch, W.W. y Bailer, A.J. (2005) *Analyzing Environmental Data,* John Wiley & Sons, Ltd, Chichester, England.

PILP (2015) Parque Industrial La Plata. Información obtenida de: parqueindustrial@laplata.gov.ar.

Planchon, O., Damato, F., Dubreuil, V. and Gouery, P. (2006) A method of identifying and locating sea-breeze fronts in north- eastern Brazil by remote sensing, *Meteorological Applications,* 13: 225- 234.

Platt, U., Perner, D., Patz, H.W. (1979). Simultaneous measurement of atmospheric CH_2O, O_3 and NO_2 by differential optical absorption, *Journal of Geophysical Research,* 84: 6329–6335.

Platt, U. y Stutz, J. (2008) *Differential Optical Absorption Spectroscopy. Principles and Applications.* Springer, Heidelberg.

PLN (2004) *Régimen de Libre Acceso a la Información Pública Ambiental*, Poder Legislativo Nacional (PLN), República Argentina.

PLP (2015) Puerto de La Plata. Información obtenida de http://puertolaplata.com

PNUMA (2004) *Geo Argentina. Perspectivas del Medio Ambiente de la Argentina*, Programa de las Naciones Unidas para el Medio Ambiente (PNUMA) y Secretaría de Ambiente y Desarrollo Sustentable de la República Argentina (SAyDS).

PNUMA (2007) *Perspectivas del Medio Ambiente Urbano: Geo San Miguel de Tucumán*, http://www.pnuma.org/deat1/urbanas.html

PNUMA (2010) *Perspectivas del Medio Ambiente Urbano: Geo Córdoba,* http://www.pnuma.org/deat1/urbanas.html

PNUMA (2012) *Proyecto Geo Ciudades PNUMA*, www.pnuma.org

PNUMA-OMS (2002) Manuales de Metodología de GEMS/Aire. Volumen 1. *Aseguramiento de la calidad en el monitoreo de la calidad del aire urbano.* United Nations Environment Programme (UNEP) Global Environment Monitoring System, Programme Activity Centre (GEMS PAC), Kenia y World Health Organization (WHO), Prevention of Environmental Pollution (PEP), Ginebra.

Prieto Méndez, J.M. (2013) *Derechos de la Naturaleza*, *Fundamento, contenido y exigibilidad jurisdiccional,* Centro de Estudios y Difusión del Derecho Constitucional- Corte Constitucional del Ecuador, Quito.

Prüss-Üsütun A. y Corvalán, C. (2007) How much Disease Burden can be Prevented by Environmental Interventions ?, *Epidemiology,* 18 (1): 167- 175.

Puliafito, E. (2009) Gestión de la calidad del aire en la Argentina, *Libro de Actas PROIMCA* (publicado en 2009). Obtenido de: http://www.utn.edu.ar/secretarias /pp(Memorias)

Puliafito, E., Guevar, M., Puliafito, C. (2003) Characterization of urban air quality using GIS as a management system, *Environmental Pollution,* 122: 105- 117.

Puliafito, E., Rey Saravia, F., Pereyra, M., Pagani, M. (2007) Calidad del aire en el polo petroquímico de Bahía Blanca, *Libro de Actas PROIMCA* (publicado en 2009). Obtenido de: http://www.utn.edu.ar/secretarias /pp(Memorias).

Ragosta, M., Caggiano, R., D'Emilio, M., Macchiato, M. (2002) Source origin and parameters influencing levels ofheavy metals in TSP, in an industrial background area of Southern Italy*, Atmospheric Environment,* 36: 3071–3087.

Ratto, G., Videla, F., Reyna Almandos, J. Schinca, D. (2005) Análisis preeliminar de parámetros meteorológicos y prospección para el estudio de calidad de aire en la zona del Polo Petroquímico La Plata, *Actas de la 90va Reunión AFA*.

Ratto, G., Videla, F., Reyna Almandos, J., Maronna, R., Schinca, D. (2006) Study of meteorological aspects and urban concentration of SO_2 in atmospheric environment of La Plata, Argentina, *Environmental Monitoring and Assessment,* 121: 327- 342.

Ratto, G., Videla, F., Schinca, D.C., Reyna Almandos, J. (2007) Medidas ópticas de contaminantes y de parámetros meteorológicos para el estudio de calidad de aire, *Encuentro de Optica Aplicada (EOA)*, Fac. de Ing., UBA (Universidad de Buenos Aires), Buenos Aires y CIOp (CIC- CONICET), Gonnet. Poster.

Ratto, G., Videla, F., Maronna, R. (2009) Analyzing SO_2 concentrations and wind directions during a short monitoring campaign at a site far from the industrial pole of La Plata, Argentina*, Environmental Monitoring and Assessment,* 149: 229- 240.

Ratto, G., Videla, F., Maronna, R., Flores, A., De Pablo, F. (2010a) Air pollutant transport analysis based on hourly winds in the city of La Plata and surroundings, Argentina, *Water Air and Soil Pollution,* 208: 243-257.

Ratto, G., Maronna, R., Berri, G. (2010b) Analysis of wind roses using hierarchical cluster and multidimensional scaling analysis at La Plata, Argentina, *Boundary Layer Meteorology,* 137: 477- 492.

Ratto, G. y Nico, A. (2012a) Preliminary wind analysis regarding different speed ranges in the city of La Plata, Argentina, *Revista Brasileira de Meteorologia,* 27 (3), 281 – 290.

Ratto, G., Maronna, R., Repossi, P., Videla, F., Nico, A., Reyna Almandos, J. (2012b) Analysis of Winds Affecting Air Pollutant Transport at La Plata, Argentina*, Atmospheric and Climate Sciences,* 2, 60-75.

Ratto, G., Videla, F., Maronna, R., Reyna Almandos, J. (2012c) Calm analysis using a robust method. Argentina y Ambiente 2012, *Primer Congreso Internacional de Ciencia y Tecnología Ambiental.* Mar del Plata, 28 Mayo- 1 Junio de 2012, Argentina.

Ratto, G., Berri, G., Maronna, R. (2014a) On the application of hierarchical cluster analysis for synthesizing low-level wind fields obtained with a mesoscale boundary layer model, *Meteorological Applications,* 21: 708–716.

Ratto, G., Videla, F., Reyna Almandos, J. (2014b) Analysis of the Homogeneity of Wind Roses' Groups Employing Andrews' Curves, *Atmospheric and Climate Sciences,* 4: 447-456.

Rehwagen, M., Müller, A, Massolo L., Herbarth, O., Ronco, A. (2005) Polycyclic aromatic hydrocarbons associated with particles in ambient air from urban and industrial areas, *Science of the Total Environment,* 348: 199– 210.

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R. (2008) *Statistical Data Analysis Explained: Applied Environmental Statistics with R*, John Wiley & Sons, Chichester.

Rencher, A.C. (2002) *Methods of Multivariate Analysis,* Second Edition, John Wiley & Sons, Canada.

Reyna Almandos, J., Videla, F., Schinca, D., Ratto, G., Ragaini, J.C., Sacchetto, V., Rosato, M., Arrieta, N., Bazán, J. (2007) Métodos ópticos aplicados al monitoreo de contaminantes atmosféricos. Poster y Libro de *Actas PROIMCA* (publicado en 2009). Obtenido de: http://www.utn.edu.ar/secretarias /pp(Memorias).

Rigby, M., Timmis, R., Toumi, R. (2006) Similarities of boundary layer ventilation and particulate matter roses, *Atmospheric Environment,* 40 (27), 5112–5124.

Ritter, G. (2015) *Robust Cluster Analysis and Variable Selection,* CRC Press, New Jersey.

Romesburg, C. (2004) *Cluster Analysis for Researchers,* Lulu Press, North Carolina, USA.

Ronco, A., Müller, A., Rehwagen, M., Massolo, L., Tueros, M., Porta, A., Franck, U., Herbarth, O. (2001). Influence of industrial, traffic and domestic emissions in the air quality of La Plata (Argentina) and Leipzig (Germany) and the potential risk associated with respiratory diseases and allergies. *Proceedings of II Mercosul Chemical Industry Congress and VII Brazilian Petrochemical Congress, IBP 13001*.: IBP—Brazilian Petroleum and Gas Institute, Río de Janeiro.

Rosato, M.E. y Reyna Almandos, J. (1996) Métodos Opticos para medición de contaminantes atmosféricos, *3er Congreso Argentino de Seguridad, Trabajo, Medio Ambiente y Comunidad.* Proyectos y Modelos para la Mejora Continua y Exposición Paralela Seguridad' 96 Argentina y el Mercosur, Buenos Aires.

Rosato, M.E., Reyna Almandos, J., Ratto, G., Flores, A., Sacchetto, V., Rosato, V. G., Ripoli, J., Alberino, J.C., Ragaini, J.C. (2001) Mesure de SO_2 à La Plata, Argentine, *Pollution Atmosphérique,* 169: 85- 98.

Rosenfeld, E., Discoli, C., Ferreyro, C., San Juan, G., Martini, I., Barbero, D., Domínguez, C., Brea, B., Melchori, M., Dicroce, L. (2005) Desarrollo de una metodología y aplicación para la elaboración de un atlas energético-ambiental para la región del Gran La Plata. *Avances en Energías Renovables y Medio Ambiente*

Vol. 9 (Reunión Nacional de ASADES- Asociación Argentina de Energías Renovables y Ambiente). Obtenido de: http://www.cricyt.edu.ar/asades/

Rosenzweig, C., Solecki, W.D., Hammer, S.A. and Mehrotra, S. (2011) *Climate Change and Cities. First Assessment Report of the Urban Climate Research Network,* Cambridge University Press, Cambridge, UK.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics,* 20: 53-65.

Rousseeuw, P.J. y Leroy, A.M., (1987) *Robust Regression and Outlier Detection,* Wiley, New York.

Rousseeuw, P.J. y Van Zomeren, B.C (1990) Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, 85 (411): 633- 651.

Rousseeuw, P.J. y Van Driessen, K. (1999) A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics,* 41 (3) 3, 212-223.

Rousseeuw, P.J. y Hubert, M. (2011) Robust statistics for outlier detection In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 1: 73- 79.

Roux, I. (2008) *Application of Cluster Analysis and Multidimensional Scaling on Medical Schemes Data,* Master Science Thesis. Department of Statistics and Actuarial Science, Stellenbosch University, Science Series, Jossey-Bass, San Francisco.

Sajesh, T.A. y Srinivasan, M.R. (2013) An Overview of Multiple Outliers in Multidimensional Data, *Sri Lankan Journal of Applied Statistics,* 14: 87- 120.

Salas- Cárdenas, S.M y Sánchez- Gonzalez, D. (2014) Envejecimiento de la población, salud y ambiente urbano en América Latina- Retos del Urbanismo gerontológico, *Contexto,* 8(9): 31-49. Obtenido en Diciembre de 2014 de: http://www.redalyc.org/

San Juan, G., Díscoli, C., Martini, I., Ferreyro, C., Rosenfeld, E., Barbero, D., Brea, B., Melchiori, M., Dicroce, L., Dominguez, C., Stange, S. (2006) Estructura de un atlas urbano-ambiental para la región del Gran La Plata. Sistematización de las variables intervinientes, *Avances en Energías Renovables y Medio Ambiente,* Vol.10. (Reunión Nacional de ASADES- Asociación Argentina de Energías Renovables y Ambiente). Obtenido de: http://www.cricyt.edu.ar/asades/

Sánchez- Triana, E., Kulsum, A., Yewande, A. (2007) *Prioridades ambientales para la reducción de la pobreza en Colombia.* Un análisis ambiental del país para Colombia. Banco Internacional de Reconstrucción y Fomento/Banco Mundial, Washington. Banco Mundial y Mayol Ediciones S.A., Bogotá.

Seber, G.A.F. (1984) *Multivariate Observations,* John Wiley and Sons, New Jersey.

Seibert, P., Beyrich, F., Gryning, S.E., Sylvain, J., Rasmussen, A., Tercier, P. (2000) Review and intercomparison of operational methods for the determination of the mixing height, *Atmospheric Environment,* 34: 1001- 1027.

Seinfeld, J.H. y Pandis, S.N. (2006) *Atmospheric Chemistry and Physics. From Air Pollution to Climate Change,* Second Edition, John Wiley & Sons, New Jersey.

Sharan, M., Kumar Yadav, A., Singh, M.P., Agarwal, P., Nigam, S. (1996) A mathematical model for the dispersion of air pollutants in low wind conditions, *Atmospheric Environment,* 30: 1209- 1220.

Sharma, S. (1996) *Applied Multivariate Techniques,* John Wiley and Sons, Chichester.

Shepard, R.N.(1980) Multidimensional Scaling, Tree-fitting and Clustering, *Science,* 210 (4468): 390- 398.

Shevlyakov, G.L. y Vilchevski, N.O. (2000) *Robustness in data analysis: criteria and methods*, De Gruyter Ed., The Netherlands. http://www.geocities.ws/gshevlyakov

Sicard, M., Perez, C., Rocadenbosch, F., Baldasano, J.M., Garcia- Vizcaino, D. (2006) Mixed-layer depth determination in the Barcelona coastal area from regular LIDAR measurements: methods, results and limitations, *Boundary Layer Meteorology,* 119: 135–157.

Simpson, J.E. (1994) *Sea breeze and local wind,* Cambridge University Press, Cambridge, UK.

Sigrist, M. (1994) *Air Monitoring by Spectroscopic Techniques,* John Wiley and Sons, New York.

Smith, K.R., Corvalán, C.F., Kjellström, T., (1999). How much global ill health is attributable to environmental factors ?, *Epidemiology,* 10 (5): 573- 584.

Smith, R.L. (2001) *Environmental Statistics,* University of North Carolina, Chapel Hill. Obtenido de: http:/www.stat.unc.edu/postcript/rs/envnotes.ps

Smith, L.I. (2002) *A tutorial on Principal Components Analysis*, Obtenido en Octubre de 2006 de: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Smith, S.J., Aardenne J. van, Klimont, Z., Andres R., Volke A., Delgado Arias, S. (2010) Anthropogenic sulfur dioxide emissions: 1850–2005, *Atmospheric Chemistry and Physics Discussion,* 10, 16111–16151.

SMN (1971) *Estadísticas Climatológicas,* Servicio Meteorológico Nacional 1961-1970, SMN, Buenos Aires.

SMN (1981) *Estadísticas Climatológicas,* Servicio Meteorológico Nacional 1971-1980, SMN, Buenos Aires.

SMN (1992) *Estadísticas Climatológicas,* Servicio Meteorológico Nacional 1981-1990, Serie B, Nº 37. SMN, Buenos Aires.

SMN (2001) *Estadísticas Climatológicas,* Servicio Meteorológico Nacional 1991-2000, SMN, Buenos Aires.

SMN (2011) *Estadísticas Climatológicas,* Servicio Meteorológico Nacional 2001-2010, SMN, Buenos Aires.

Smook, R.A.F. (1998) Chapter 62 European sustainable cities: the chanllenge of citylife: being exposed to an air polluted urban environment. En: Schneider, T. *Air Pollution in the 21st Century: Priority Issues and Policy*, Elsevier, Amsterdam.

Sneath, P.H.A y Sokal, R.R. (1973) *Numerical Taxonomy,* Ed. W.H. Freeman and Company, San Francisco.

Sokal, R.R. y Rohlf, F.J. (1962) The comparison of dendograms by objective methods, *Taxon,* 11: 33- 40.

Sosa, B.S. (2015) *Contaminación ambiental por material particulado y compuestos orgánicos volátiles en la ciudad de Tandil, Provincia de Buenos Aires*, Tesis Doctoral, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

SPA (2007) Exp.2145-7007/06 Secretaría de Política Ambiental de la Pcia. de Buenos Aires, La Plata, Argentina. (Ref.: *Solicitud de información ambiental de La Plata y alrededores según los beneficios de la Ley 25.831/04* "Régimen de libre acceso a la información publica ambiental").

Spencer, N.H. (2003) Investigating Data with Andrews Plots, *Social Science Computer Review,* 21: 244-249.

Sportisse, B. (2008) *Fundamentals in Air Pollution. From Processes to Modeling.* First Edition in English, Springer, Heidelberg.

Steinley, D. (2004) Standarizing Variables in k- means clustering. En: Studies in Classification, Data Analysis and Knowledge Organization, *Proceedings of the Meeting of the International Federation of Classification Societies (IFCS)*, Banks, D., House, L., McMorris, F.R., Arabie, P., Gaul, W. Eds., Springer, Heidelberg.

Stull, R.B. (1988) *An Introduction to Boundary Layer Meteorology,* Kluwer Academic Publishers, The Netherlands.

Suggar, C.A., Lenert, L.A., Olshen, R.A. (1999) *An application of cluster analysis to health services research: empirically defined health states for depression from the SF-12,* Technical Report Nº 203, Stanford University, California.

Takahashi, K., Mirua, T., Shioya, I. (2007) Hierarchical Summarizing and Evaluating for Web Pages *Proceedings of the 1st Workshop on Emerging Research Opportunities for Web Data Management (EROW, 2007) collocated with the 11th International Conference on Database Theory* (ICDT, 2007) Barcelona, Spain, January 13, 2007. Edited by Marcelo Arenas, Pontificia Universidad Católica de Chile, Chile - Jan Hidders, University of Antwerp, Belgium.

The MathWorks (2002) *Curve Fitting Toolbox for use with Matlab*, User's Guide Version 1, The MathWorks, Inc.

Theodoridis, S. y Koutroumbas, K. (2003) *Pattern Recognition,* 2nd Edition, Elsevier, San Diego.

Thode, H.T. Jr. (2002) *Testing for Normality,* Marcel Dekker Inc., New York.

Tibshirani, R., Walther G. and Hastie, T. (2001) Estimating the Number of Clusters in a Dataset via the Gap Statistic, *Journal of the Royal Statistical Society Series B,* 63 (2), 411–423.

Tibshirani, R. y Walther, G. (2005) Cluster Validation by Prediction Strength, *Journal of Computational and Graphical Statistics*, 14(3): 511–528.

Timm, N.H. (2002) *Applied Multivariate Analysis,* Springer- Verlag, New York.

Tukey, J.W. (1977) *Exploratory Data Anlalysis,* Addison- Wesley Publishing, Company Inc. Massachussets.

Ulke, A.G., Longo, K.M., Freitas, S.R., Hierro, R.F. (2007) Regional pollution due to biomass burning in South America, *Ciência e Natura,* 10, 201.

Unal, Y., Kindap, T., Karaka, M. (2003). Redefining the climate zones of Turkey using cluster analysis, *International Journal of Climatology,* 23: 1045–1055.

UNEP (2010) *Geo Cities Manual - Guidelines for Integrated Environmental Assessment of Urban Areas,* EECCA Region, United Nations Environment Programme, UNEP-DEWA/GRID-Europe.

UNEP (2014a) Justicia ambiental y desarrollo sostenible: un simposio mundial sobre el estado de derecho ambiental, *Asamblea de las Naciones Unidas sobre el Medio Ambiente del Programa de las Naciones Unidas para el Medio Ambiente*, UNEP/EA.1/CW/CRP.1, Primer período de sesiones, 23 a 27 de junio de 2014, Nairobi.

UNEP (2014b) Plan de Acción Regional de Cooperación Intergubernamental en materia de Contaminación Atmosférica para América Latina y el Caribe, *XIX Reunión del Foro de Ministros de Medio Ambiente de América Latina y el Caribe,* 11- 12 de Marzo de 2014, UNEP/LAC-IGWG.XIX/7 Final, Los Cabos, México.

Unwin, A. (2008) Good Graphics? In: Chen, C., Hardle, W. and Unwin, A., Eds., *Handbook of Data Visualization,* Springer, Heidelberg, 57.

UN-HABITAT (2012) *State of the World's Cities. Prosperity of Cites.* United Nations Human Settlements Programme, Nairobi.

Urbina Soria, J. y Martinez Fernandez, J. (2006) *Más allá del cambio climático. Las dimensiones psicosociales del cambio ambiental global.* Primera edición. Instituto Nacional de Ecología (INE-Semarnat), Universidad Nacional Autónoma de México (UNAM), Facultad de Psicología, www.ine.gob.mx.

USAC- MAG (2012) *Monitoreo del aire en la ciudad de Guatemala. Informe anual 2011.* Universidad de San Carlos- Ministerio de Ambiente y Recursos Naturales, Guatemala.

US ATSDR (1998) *Toxicological Profile for Sulphur Dioxide,* Chapter 5. Agency for Toxics Substances and Disease Registry - Public Health Service: Science International Inc. Editors, Georgia.

Vallero, D. (2008) *Fundamentals of Air Pollution,* 4th edition Academic Press, California.

Varmuza, K. y Filzmoser, P. (2009) *Introduction to Multivariate Statistical Analysis in Chemometrics,* CRC Press, Taylor & Francis Group, Boca Raton.

Velleman, P.F. y Hoaglin, D.C. (2004) *Applications, Basics and Computing of Exploratory Data Analysis,* The Interntet- First University Press (republished), Cornell University, NY.

Veltkamp, R.C. y Lateecki, L.J. (2006) Properties and Performance of Shape Similarity Measures, pp. 47-56. In: *Data Science and Classification,* Edited by: Batagelj, V. et al., Springer, Heildelberg.

Videla, F., Schinca, D., Ratto, G., Ragaini, J.C. (2006) Desarrollo de equipos ópticos para medir SO_2 en chimeneas y aire ambiente. Presentación de resultados de mediciones de SO_2 y parámetros meteorológicos utilizando equipamiento comercial en el área de La Plata, Tecnologías e instrumentos para su evaluación integral, Sección: *La calidad del ambiente urbano. Libro de Actas LINTA*.

Wais de Badgen, I.R. (1998) *Ecología de la Contaminación Ambiental,* 1ra Edición. Ediciones Universo, Buenos Aires.

Wang, L., Zhang, Y., Feng, J. (2005) On the Euclidean Distance of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27 (8):1334-1339.

Wang, S., y Serfling, R. (2012) *On Masking and Swamping Robustness of Leading Outlier Identifiers for Univariate Data*. Educational Report. Disponible en: www.utdallas.edu/_serfling.

Wanta, R.C. (1968) Meteorology and Air Pollution In: *Air Pollution (Stern, A.)* Vol. 1 Chapter 7, Second Edition, New York Academic Press, New York.

Wark, K., Warner C., Davis, W. (1998) *Air Pollution. Its Origin and Control,* 3rd Edition, Addison Wesley Longman, Berkeley.

Weisberg, S. (2005) *Applied Linear Regression,* Third Edition John Wiley & Sons, Inc., New Jersey.

Weitkamp, C. (2005) *Lidar. Range resolved optical remote sensing of the atmoshpere,* Springer, Singapore.

Whichmann, F.A., Müller, A., Busi, L.E., Cianni, N., Massolo, L., Schlink, U., Porta, A., Peter Sly, D., (2009). Increased asthma and respiratory symptoms in children exposed to petrochemical pollution. *Journal of Allergy and Clinical Immunology,* 123: 632- 638.

WHO (1980) *Analysing and Interpreting Air Monitoring Data,* Report Nº 51, Geneva.

WHO (1998) *La Salud en las Américas,* Vol. 2, Publicación Científica Nº 569, Washington.

WHO (2000a) *Guidelines for Air Quality*, World Health Organization, Geneva. Disponible en: http://www.who.int/peh/

WHO (2000b) *Air quality guidelines for Europe,* 2nd ed. Copenhagen, World Health Organization Regional Office for Europe, WHO Regional Publications, European Series Nº 91.

WHO (2005) *Effects of Air Pollution on Children's Health and Development- A Review of The Evidence,* Word Heath Organization, European Centre for Environment and Health Bonn Office, Bonn.

WHO (2006) *Planning to protect children against hazards*, Word Heath Organization, Europe. Disponible en: http://www.euro.who.int/eehc.

WHO (2013) *Health risks of air pollution in Europe –HRAPIE project,* Regional Office for Europe, Copenhagen.

WHO (2014) Comunicado de Prensa: *7 millones de muertes cada año debidas a la contaminación atmosférica*. Disponible en: http://www.who.int/mediacentre/news/releases/2014/air-pollution/es.

Wieringa, J. (1980) Representativeness of Wind Observations at Airports, *Bulletin of the American Meteorological Society,* 61: 962- 971.

Wieringa, J. (1996) Does representative wind information exist?, *Journal of Wind Engineering & Industrial Aerodynamics,* 65: 1- 12.

Wikipedia (2011) Información obtenida de: https://www.wikipedia.org.

Wilcox, R.R. (2005) *Introduction to Robust Estimation and Hypothesis Testing,* Second Edition, Elsevier Academic Press, Oxford.

Wilks, D.S. (2006) *Statistical Methods in the Atmospheric Sciences,* Second Edition Elsevier, New York.

Wish, M. y Carroll, J.D. (1982) Multidimensional Scaling and its applications En: *Handbook of Statistics* Vol. 2, Krishnaiah, P.R. y Kanal, L.N. Eds., North Holland, Amsterdam.

WMO (1983) *Guide to Climatological Practices,* Nº 100. World Meteorological Organization, Switzerland.

WMO (2008) *Guide to Meteorological Instruments and Methods of Observation,* WMO-Nº 8., World Meteorological Organization, Switzerland.

Wolter K. (1987) The southern oscillation in surface circulation and climate over the Tropical Atlantic, Eastern Pacific and Indian Oceans as captured by cluster analysis, *Journal of Climatology and Applied Meteorology,* 26: 540–558.

Xu, R. y Wunsch, D. C. (2009) *Clustering,* John Wiley & Sons, Inc., Hoboken, New Jersey.

Yeung, K.Y. y Ruzzo, W.L. (2001) Principal Component Analysis for clustering gene expression data, *Bioinformatics,* 17: 763- 774.

Young, F.W. (1987) *Multidimensional Scaling: History, Theory and Applications,* Hamer, R.M. (Ed.), Hillsdale, NJ: Lawrence Erlbaum.

Yu, K.N., Cheung, Y.P., Cheung, R.T., Henry, C. (2004) Identifying the impact of large urban airports on local air quality by nonparametric regression, *Atmospheric Environment,* 38: 4501–4507.

Zoras, S., Triantafyllou, A.G., Evagelopoulos, V. (2008) Aspects of year-long differential optical absorption spectroscopy and ground station measurements in an urban street canyon near industrial pollution sources, *Atmospheric Environment,* 42: 4293–4303.