UNIVERSIDAD NACIONAL DE LA PLATA

Exploración de la confluencia entre Agroinformática, IoT, Grandes Datos y Extracción del Conocimiento

Trabajo Final presentado para obtener el grado de
Especialista en Ingeniería de Software

Facultad de Informática
Universidad Nacional de La Plata

Autor Javier Ernesto Matarrese Director Dr. Alejandro Fernandez

2020

Índice

1 Introducción

El sector Agrícola es de vital importancia en muchos países, sobretodo en aquellos que están en vías de desarrollo donde una gran cantidad de población depende económicamente de los resultados de este sector. La agricultura es fuente de alimentos, de bienes de intercambio a nivel nacional e internacional y de recaudación a través de impuestos para los gobiernos, entre otras cosas. La mejora en el desempeño de este sector requiere de recursos e innovación, de forma de mejorar el acceso a nuevos mercados e incrementar la capacidad productiva.

En las últimas décadas, la evolución de las tecnologías de la información y las comunicaciones ha sido notable. Se observa la aplicación de estas a casi todos los dominios, ha permitido moverse en dirección de los objetivos de estos y obtener resultados muy positivos para la humanidad.

En un contexto donde la predicción del índice de crecimiento de la población mundial supera al de producción de alimentos, y donde el cambio climático y la disminución de la superficie arable complican aún más las cosas, muchos países del mundo están poniendo el foco en temas como el desarrollo de cultivos de alto rendimiento y de ciclo más corto, desarrollo de fertilizantes más efectivos, cultivos resilientes al cambio climático, manejo efectivo de la cadena de suministros y reducción del desperdicio, entre otros.

Tradicionalmente la infraestructura agrícola y la mano de obra eran consideradas como los motores del progreso agrícola, sin embargo, hace ya algunos años se visualizó la vital importancia de la información (de la mano de las TICs) como motor de crecimiento y expansión del sector agrícola.

La cadena de producción agrícola es una candidata ideal para la aplicación combinada de tecnologías como Internet de las cosas, grandes datos y extracción del conocimiento, dado que esta contiene diversos procesos que requieren ser controlados y administrados donde contar con información precisa, concisa, oportuna y completa lleva a mejorar la planificación y el proceso de toma de decisiones de propietarios, administradores y generadores de políticas públicas.

Es posible aplicar Internet de las cosas en distintos pasos de la cadena agro-industrial Talavera et al. (2017). Puede evaluar variables de campo como estado del suelo, condiciones atmosféricas y biomasa de plantas o animales. Durante el transporte puede utilizarse para asegurar y controlar variables como temperatura, humedad, vibración y golpes de la producción. En el almacenamiento puede monitorear el estado del producto.

La traza generada en la cadena puede suministrar información al consumidor/usuario final acerca del producto y sus propiedades.

La utilización de la información recolectada por sensores a lo largo de la cadena, sumada a fuentes de información externas, puede alimentar procesos de grandes datos y extracción del conocimiento, generando predicciones que apoyen los procesos de planificación y toma de decisiones sobre utilización del terreno, fertilizantes, herbicidas, fungicidas e insecticidas, por ejemplo, en cultivos. Permite predecir las necesidades de transporte, almacenamiento y demanda, permitiendo planificar la producción contando con un panorama completo de la cadena de valor agrícola en un entorno de alto riesgo e incertidumbre.

La innovación en agricultura es, además, el resultado de la interacción entre una variedad de actores como: agentes gubernamentales, firmas privadas, asociaciones de agricultores, organizaciones no gubernamentales, etc. Es aquí donde las TICs puede contribuir generando una red donde puedan actuar de forma integrada que haga a cada entidad individual más productiva.

2 Objetivos del Trabajo

El presente trabajo pretende alcanzar los siguientes dos objetivos:

- Realizar un compendio del estado del arte de los tópicos principales: Internet de las cosas (IoT), Grandes Datos, Extracción del conocimiento y Agroinformática
- Explorar la intersección de los estados del arte de cada tópico mencionado con el foco puesto sobre las aplicaciones en el Agro. Nombrando, de esta forma, algunas de las líneas de trabajo exploradas en la actualidad, así como potenciales por explorar.

3 Organización del Trabajo

Para alcanzar los objetivos planteados, el presente trabajo se organizará de la siguiente manera:

En el capítulo 4 se explorarán los estados del arte de lo que denominamos "tópicos principales", para cada uno de ellos, haremos una descripción general donde se incluirán los conceptos más importantes, problemáticas, aplicaciones y desafíos, entre otros. Tendremos, entonces: la sección 4.1 donde trataremos Internet de las Cosas (IoT), sección 4.2 que se dedicará a Grandes Datos (Big Data), sección 4.3 donde abordaremos Extracción del Conocimiento y finalmente 4.4 dedicada a la aplicación de TICs en el agro o agroinformática.

En el capítulo 5, exploraremos las intersecciones existentes entre cada uno de los tópicos principales, partiendo desde cada una de las temáticas planteadas para agroinformática (4.4). Nombraremos aquí, algunas de los puntos específicos tratados en la bibliografía utilizada, tratando de dejar a la vista posibles líneas a seguir en próximos trabajos.

Finalmente el capítulo 6 estará destinado a las conclusiones del presente trabajo y el capítulo 7 al índice de la bibliografía utilizada.

4 Tópicos principales

4.1 Internet of Things (IoT)

La primera utilización del término Internet of Things se atribuye a Kevin Ashton, Director Ejecutivo del AutoId Center del Massachusetts Institute of Technology (MIT). Según Ashton utilizó este término relacionando la utilización de tecnología RFID con Internet (que era el tópico más caliente del momento) intentando de esta forma captar la atención de los ejecutivos de Procter & Gamble (P&G) en una presentación que les ofreció en el año 1999 («That 'Internet of Things' Thing - 2009-06-22 - Page 1 - RFID Journal», 2009).

Existe un sinnúmero de definiciones de IoT, dado el amplio espectro y la cantidad de tecnologías relacionadas, encontrar una definición que no sea ambigua es realmente difícil. La definición de la IERC sostiene que IoT es "una infraestructura de red dinámica y global con capacidades de auto-configuración basada en protocolos de comunicación estándares e interoperables donde las cosas físicas y virtuales tienen identidades, atributos físicos y personalidades virtuales, usan interfaces inteligentes y se integran a la perfección en una red de información" Vermesan & Friess (2014)

Internet de las cosas es la última revolución en Internet. Las cosas se exponen a si mismas se hacen auto-descubribles y cobran inteligencia, ejecutando o habilitando a las toma de decisiones relativas al contexto gracias al hecho de que pueden comunicar información sobre sí mismas. Apoyan este cambio el surgimiento de las tecnologías en la nube y la transición de Internet hacia IPv6 que permite el acceso a una capacidad de direccionamiento ilimitada. Patel, Patel, & Professor (2016)

Según la empresa Gartner Inc. en 2017 la cantidad de dispositivos conectados ascenderá a 8.4 Billones, esto representa un aumento del 31% con respecto al año anterior, y se espera que para 2020 la cantidad ascienda a 20.4 billones («Gartner Says 8.4 Billion Connected», 2017).

Las soluciones de IoT, en cuanto a los sectores funcionales, suelen estar asociadas al concepto de "inteligencia" o "rapidez", usualmente con nombres que contienen la palabra "smart" en inglés: Smart Wearable, Smart Health, Smart Homes, Smart Cities, Smart Environment, Smart Farming, Smart Enterprise, Smart Energy & Smart Grid, Smart Mobility & Transport, son solo algunos ejemplos.

Las características fundamentales de un ambiente IoT son, según Vermesan & Friess (2014): - Interconectividad: con respecto a IoT, todo puede estar interconectado mediante una infraestructura de comunicación global. - Servicios relacionados con las cosas: IoT es capaz de brindar servicios relacionados con las cosas, dentro de las restricciones de estas, como ser protección de la privacidad y consistencia semántica entre la cosa física y su contraparte virtual asociada. - Heterogeneidad: Los dispositivos de IoT son heterogéneos, basados en distintas plataformas de hardware y red. - Cambios dinámicos: El estado de los dispositivos cambia dinámicamente, así como cambia de la misma manera el contexto al que están relacionados. - Gran escala: la cantidad de dispositivos que estarán conectados serán al menos un orden de magnitud mayor a la cantidad que están conectados a internet hoy en día.

Siendo IoT un concepto tan amplio, no existe un acuerdo en cuanto a un único modelo de arquitectura de un sistema IoT. Existen varios modelos propuestos Group et al. (2015) ("European FP7 Research Project", "ITU Architecture", "IoT Forum Architecture", "Qian Xiaocong, Zhang Jidong Architecture", "Kun Han, Shurong Liu, Dacheng Zhang and Ying Han's (2012)'s Architecture", entre otros). A continuación describimos uno («Gartner Says 8.4 Billion Connected», 2017) que nos parece lo suficientemente amplio como para cubrir un gran área de aplicación. Según el mismo una arquitectura de IoT puede descomponerse, para su estudio, en cuatro capas:

- **Capa de dispositivos y sensores inteligentes**: Es la capa más baja, la componen los sensores que interrelacionan el mundo físico con el mundo digital. Los sensores son capaces de realizar mediciones de elementos físicos (posición, movimiento, temperatura, presión, calidad de aire, etc.) y convertirlas a valores digitales. Algunos pueden tener incluso algún nivel de memoria que les permite juntar una colección de mediciones para luego transmitirlas de una forma más eficiente o incluso puede incorporar un nivel bajo de inteligencia que les permite optimizar tempranamente el pasaje a las siguientes capas.
- **Gateways y Redes**: Dado que los sensores suelen producir una gran cantidad de información se requiere de una infraestructura de comunicación robusta (ya sea conectada o wireless) que cumpla la función de medio de transporte. Los diferentes tipos de redes disponibles actualmente (que usualmente utilizan protocolos de comunicación diversos) se utilizan para formar redes machine-to-machine (M2M network). Se requiere que estas redes, que sirven a una cantidad de servicios y aplicaciones de IoT, funcionen en una configuración heterogénea satisfaciendo requerimientos específicos de ancho de banda, latencia y seguridad.
- **Capa de gestión de servicios**: La capa de gestión de servicios hace que el procesamiento de información sea posible mediante analytics, controles de seguridad, modelado de procesos y administración de dispositivos.
  - Una parte importante de esta capa son los motores de reglas de negocios y procesos. La información en IoT tiene forma de eventos o información contextual, muchas veces requiere ser filtrada o enrutada hacia sistemas de

post-procesamiento, en otras requiere respuesta a situaciones inmediatas. Los motores de reglas soportan la generación de lógicas de decisión y disparan procesos (interactivos o automáticos) que hacen el sistema de IoT más responsivo.

  - En cuanto a analytics, se utilizan diversos métodos para poder extraer información relevante de la gran cantidad de eventos que arriban a esta capa en forma cruda y a altas velocidades. Técnicas de ***in-memory analytics*** permiten obtener información más valiosa a la hora de tomar decisiones y evitan el guardado de los eventos crudos en disco para ser luego procesados, mejorando los tiempos de respuesta. En esta línea, también son utilizadas técnicas de ***streaming analytics*** que tratan los eventos como un flujo de datos sin fin y permiten obtener información en tiempo real.
  - Sobre la gestión del flujo de información, en esta capa se administra cómo la información es accedida, integrada y regulada evitando que la capa más alta (capa de aplicación) sea recargada con el ingreso de información innecesaria o información que atente contra políticas de revelado de información privada. Se aplican técnicas de filtrado, agregación, sincronización (de eventos de varias fuentes) y ofuscación.

- **Capa de aplicación** Esta capa es la más cercana al dominio del sistema. Es la que aporta el grado de inteligencia en relación con el entorno físico de aplicación (smart transport, smart cities, supply chains, emergency response, smart agriculture). Es la encargada de entregar los servicios finales contando con la información preprocesada de la capa anterior y es encargada de proveer una visión funcional a los consumidores de los servicios del sistema.

4.1.1 Características principales IoT

El éxito en la implementación de un sistema de IoT está estrechamente relacionado con la arquitectura seleccionada en el momento del diseño del mismo. Al evaluar una arquitectura de IoT se debe tener en cuenta que cubra las características básicas de IoT y que soporte futuras extensiones. Estas características pueden ser enumeradas de la siguiente manera Abdmeziem, Tandjaoui, & Romdhani (2016):

- Distributividad Por su naturaleza, un sistema IoT es altamente distribuido. Para ser efectivo necesita contar con un número de sensores, actuadores, puertas de enlace, dispositivos de borde (computación de niebla), procesamiento en la nube, solo por nombrar algunos. De esta forma, los datos son obtenidos de distintas fuentes, y procesada en distintos lugares. Estos componentes, radicados físicamente en lugares dispares, forman una plataforma distribuida por definición.

- Interoperatividad Es necesario que los componentes (muchas veces de distintos fabricantes) se entiendan y puedan interactuar en pos de un objetivo común. Por esto, es importante la interoperatividad, Deben tenerse en cuenta características de comunicación, protocolos, formatos de dato, etc.
- Escalabilidad En los próximos años se espera un crecimiento exponencial de los objetos conectados a la red con propósitos de constituir sistemas de IoT. La proliferación de objetos conectados, así como el crecimiento de datos generados por los mismos debe ser previsto desde el diseño del sistema.
- Escasez de recursos Se debe soportar un marco en el que el acceso a los recursos de red sean esporádico y los requerimientos energéticos bajos, recursos de computación escasos y el tamaño físico de los dispositivos el mínimo posible. En muchas aplicaciones, sería inviable no satisfacer algunos de estos requisitos.
- Seguridad Los datos recolectados por los sensores, en la mayoría de los dominios de aplicación, son sensibles en cuanto a privacidad y autenticidad. Es deseable encontrar un balance entre seguridad y utilización de recursos (que suelen ser limitados por lo expuesto en el punto anterior).

4.1.2 Tecnologías Habilitantes:

Podemos dividir las tecnologías habilitantes de IoT en las siguientes tres categorías:

1. Tecnologías que habilitan a los “cosas” a adquirir información contextual
2. Tecnologías que habilitan a las “cosas” a procesar información contextual
3. Tecnologías que mejoran la seguridad y privacidad.

Las dos primeras categorías se consideran como un requerimiento (o conjunto de ellos) funcional que dota de inteligencia a las cosas. La tercer categoría es un requerimiento no funcional que, si no estuviere contemplado, reduciría en gran medida la generalización de la aplicación de IoT.

4.1.3 Desafíos IoT:

Observando el estado actual de IoT podemos destacar algunos desafíos, que de ser abordados adecuadamente, permitirían cubrir áreas hasta ahora no cubiertas y mejorarían la velocidad de adopción de esta práctica Pundir, Sharma, & Singh (2016).

- Estándares e interoperabilidad Un ecosistema de gran cantidad de dispositivos diversos y altamente conectados trae aparejado una alta heterogeneidad (entre otras cosas) que es necesario manejar. Crece la necesidad, en este contexto, de contar con estándares que propicien una convivencia efectiva en pos de los objetivos perseguidos. Desde el punto de vista del mercado, los estándares favorecen la interacción entre los fabricantes y evitan que los consumidores queden atrapados en un único fabricante. Mientras que los estándares ayudan a conseguir interoperabilidad técnica y sintáctica (dimensiones asociadas a la interacción entre máquinas), la interoperabilidad semántica, por su naturaleza, debe ser gestionada en capas superiores relacionadas con el nivel de aplicación y el dominio de los datos. Se encuentran dentro de las dificultades a superar el hecho de que estándares tengan en cuenta las normas y regulaciones vigentes en la región de utilización del sistema. Por ejemplo, distintos países tienen distintas regulaciones de la utilización del espectro radioeléctrico (en cuento a frecuencias de operación y potencias permitidas) que deben ser satisfechas.
- Seguridad La proliferación de los dispositivos inteligentes y la adopción masiva de IoT generan un sistema complejo donde aumentan los puntos expuestos a operaciones maliciosas. La cuestión de la seguridad tiene diferentes aristas en cada nivel lógico de la arquitectura de IoT (descrita previamente). Por ejemplo, en la capa más baja, capa de sensores y dispositivos inteligentes, nos encontramos con una gran cantidad de dispositivos con que operan con restricciones de procesamiento, energía, etc. que exponen al sistema y generan el desafío de mitigarlas con soluciones que tengan en cuenta las restricciones mencionadas (encriptación liviana, control de acceso a nodos, trust anchors y attestation).
- Autenticación y privacidad Dada la naturaleza de la información con la que trabaja IoT, en general datos privados capturados con sensores remotos, asume gran importancia el control de acceso y propiedad de dicha información. Es importante, además, que los dispositivos implementen mecanismos de autenticación e integridad para evitar incidentes de seguridad, atento a lo expuesto en el punto anterior. El cumplimiento de estos temas se torna sumamente importante en algunas áreas como medicina y aplicaciones para vida asistida, donde una desviación podría tener consecuencias graves. Existen marcos de trabajo que guían a las organizaciones en el proceso de aseguramiento de seguridad en IoT (security compliance frameworks), cuya evolución está tomando importancia.
- Complejidad y problemas de integración En un ambiente heterogéneo, donde conviven dispositivos con distintas características de hardware provenientes de fabricantes dispares, diferentes protocolos de comunicación, redes de variadas características, solo por nombrar algunas cuestiones, la integración es un desafío a considerar.
- Arquitecturas en desarrollo, guerras de protocolo y estándares en competencia. La falta de estandarización en las capas más bajas de la arquitectura, generan una demanda adicional de recursos de desarrollo de software en las capas posteriores

de forma de homogeneizar los resultados. A su vez, este fenómeno se reproduce en las capas siguientes potenciando generando retrabajo hasta llegar a la capa más alta. Por ejemplo, las modificaciones en las APIs son comunes para responder a cambios de hardware en los sensores o en las redes requiriendo modificaciones en el software que las consume.

- Casos de uso concreto y proposiciones convincentes Si bien existe una gran cantidad de teoría y propuestas aplicables, para extender la adopción de IoT el mercado requiere, no solo las descripciones de ciertos componentes que proveen los fabricantes de dispositivos, sino casos de éxito en sistemas completos.

4.1.4 Enfoque Interdisciplinario

Las áreas de aplicación de IoT son tan variadas como "cosas" que pueden beneficiarse gracias a estar conectadas formando una red de información de características dinámicas. Solo por dar algunos ejemplos podemos mencionar: wearables, salud, monitoreo de tránsito, gestión de flotas, agricultura, hotelería, smart grid y ahorro de energía, suministro de agua, mantenimiento industrial, etc. A cada área de aplicación aportan múltiples disciplinas relacionadas con el dominio, como un burdo ejemplo, para el área de salud podríamos mencionar: medicina, enfermería, crujía, neurología, radioterapia, etc. Cuando pensamos en aplicaciones de IoT, tenemos que incluir, además de las disciplinas relacionadas con el dominio de aplicación, las otras que están relacionadas con el funcionamiento del sistema de IoT en sí: telecomunicaciones, sistemas de automatización y control, ciencias de la computación, ingeniería. Como podemos desprender de lo mencionado en los dos párrafos anteriores, para la investigación en IoT es requisito excluyente trabajar con un enfoque interdisciplinario donde las disciplina, de acuerdo a los vínculos establecidos por el objetivo común, desarrollan acciones conjuntas. Del trabajo "Top Fifty Highly Cited Publications on the Internet of Things," Jayasekara & Abu (2018) donde se seleccionan las cincuenta publicaciones IoT más citadas en el año 2017, nos parece interesante extraer un listado de las categorías en las que Web Of Science clasifica a cada uno de los trabajos, estas categorías refieren a las disciplinas involucradas en cada trabajo. Reproducimos abajo el listado de categorías ordenado por cantidad de publicaciones:

- Telecomunicaciones
- Ciencias de la computación, Sistemas de información
- Ingeniería Eléctrica y Electrónica
- Ciencias de la Computación, hardware y arquitectura
- Ciencias de la Computación, Ingeniería en Software
- Ciencias de la Computación, aplicaciones interdisciplinarias
- Ciencias de la Computación, teoría y métodos
- Ingeniería Industrial
- Automatización y Sistemas de Control

- Ciencias de la Computación, inteligencia artificial
- Ciencias multidisciplinarias

Yendo más a detalle, puede ser interesante identificar los tópicos o subdisciplinas, dentro de las disciplinas que contribuyen a IoT. Identificar estos puntos puede ser de interés a la hora de buscar perfiles para armar el equipo de investigación según los requerimientos de la temática a trabajar. Olagunju & Khan (2016)

Disciplina	Subdisciplina / Tópicos
Ciencias de la computación	• diseño de programas • estructuras de datos • diseño de bases de datos • lenguajes de bajo nivel • sistemas operativos
Ingeniería de sistemas e ingeniería electrónica	• sistemas embebidos • circuitos electrónicos • microcomputación • automatización y control • sensores • campos y ondas electromagnéticas • procesamiento digital de señales • circuitos de comunicación
Ingeniería de Software	• extracción del conocimiento • sistemas de infraestructura
Sistemas de Información	• redes de computadoras • tecnología de sensores • seguridad

4.1.5 Aplicaciones IoT

Las áreas de aplicación de IoT son muy amplias, casi cualquier situación de la vida diaria, donde intervengan dispositivos con cierta inteligencia, puede beneficiarse de la aplicación de esta técnica. La interacción de estos dispositivos en una red dinámica y global dotan al sistema, como un todo, de posibilidades hasta ahora impensables. Es

por ello, que enumerar las posibles aplicaciones sería imposible, nos limitamos entonces a describir algunas de las más comunes:

Ciudades inteligentes El crecimiento de la aplicación de los conceptos agrupados bajo el nombre de Ciudad Inteligente está propulsado por varias razones, entre ellas el crecimiento exponencial de la disponibilidad de dispositivos que pueden participar en una infraestructura IoT, a la vez la capacidad de procesamiento de la información generada por estos sensores también está creciendo de forma similar. Otra de las razones es el crecimiento en la población mundial y la tendencia de urbanización previstos por las organizaciones especializadas. Dado que las ciudades consumen el 75% de la energía mundial, producen el 80% de las emisiones de gases con efecto invernadero, utilizan el 60% del agua destinada a uso humano, para nombrar alguno de los factores, se hace imperioso desarrollar mecanismos que permitan que las ciudades crezcan de manera sustentable Theodoridis, Mylonas, & Chatzigiannakis (2013). Dados esto factores, la visión de Ciudades Inteligentes pretende solucionar, o al menos disminuir, una variedad de problemas de los que nombraremos algunos solamente para ejemplificar.

- Estacionamiento inteligente: colabora en la gestión de los espacios destinados a estacionamiento en la ciudad mediante la instalación de sensores ferromagnéticos inalámbricos que proveen información de ocupación de las plazas destinadas al estacionamiento, esta información es provista mediante una aplicación que ayuda al ciudadano a localizar un lugar disponible de manera rápida y segura, disminuyendo las distancias recorrida por el vehículo (por ende disminuyendo la contaminación y el tránsito) y proveyendo de una gestión óptima de los espacios.
- Monitoreo: provee valiosa información de gestión obtenida de una red de sensores que registran distintos parámetros como: polución ambiental, nivel de ruido, entre otros. Estos sensores se instalan en lugares como postes de iluminación, o incluso en puntos móviles como ser el transporte público, vehículos de policía, vehículos municipales, etc. Esta información permite tomar acciones correctivas como restricciones horarias de circulación en ciertas zonas, por ejemplo.

- Sensado participativo: los ciudadanos colaboran utilizando una aplicación en sus dispositivos móviles que recoge información mediante sus sensores físicos como posición (mediante GPS, brújula), datos de medioambiente como nivel sonoro y temperatura, etc. Además pueden enviar información sobre eventos espaciales (incendios, accidentes de tránsito, etc). Esta información alimenta un mapa al que los usuarios pueden suscribirse de forma de recibir los eventos relevantes que ocurren en su zona de circulación/permanencia.
- Irrigación de precisión de espacios públicos y jardines: el sistema provee información en tiempo real de los mecanismos de irrigación instalados, así como humedad en el suelo y temperatura, entre otros, que genera acciones que permiten una utilización óptima del agua, así como un mejor mantenimiento del sistema de irrigación cuando situaciones irregulares son detectadas.

- Medición de consumo inteligente: la información provista por cada dispositivo sobre su consumo energético, sumada a los sensores ambientales (temperatura, humedad, iluminación) instalados en los edificios y espacios públicos, permiten una gestión más eficiente de la energía, por ejemplo regulando la calefacción en espacios cerrados y la iluminación de acuerdo a las necesidades de las personas que los transitan.
- Recolección de residuos inteligente: dotar a los contenedores hogareños de basura de sensores que transmitan el nivel de llenado y centralizar esta información de forma de poder general rutas de recolección óptima permiten ahorrar combustible y otros recursos valiosos, así como planificar y controlar el tratamiento y la disposición en los depósitos finales de los residuos.

Edificios / casas inteligentes

- Edificios inteligentes: los esfuerzos en hacer las casas y edificios inteligentes se centran en bajar los costos operativos, reflejados en gastos de energía por ejemplo en calefacción, aire acondicionado, iluminación, entre otros. Esto se logra mediante la instalación de una red de sensores y controladores inteligentes que se despliegan a lo largo de la edificación. Los datos provenientes de los sensores junto con fuentes externas, permite la toma de decisiones autónomas que mejoran la eficiencia en el uso de la energía, a la vez que mejoran el confort de los habitantes.
- Seguridad y vigilancia: los sistemas de seguridad actual se basan en características físicas que disparan alarmas. La visión en este caso es agregar inteligencia en las decisiones mediante el agregado a los sensores físicos de sistemas de procesamiento de imagen y procesamiento con algoritmos inteligentes capaces de decidir en base a condiciones actuales e históricas, y puedan planificar y ejecutar acciones de mitigación y alerta.

Agricultura

- Irrigación de precisión: Con el objetivo de optimizar la cantidad de agua utilizada para irrigación (utilizando de forma efectiva este recurso y minimizando la energía necesaria para bombeo) se instalan diversos sensores de humedad y temperatura del suelo, humedad y temperatura ambiental, velocidad del viento e intensidad de luz que permiten al sistema tomar decisiones en forma autónoma para controlar válvulas electromecánicas que regulan el flujo del líquido.
- Iluminación inteligente: Se aplica especialmente en cultivos de invernadero con el objetivo de optimizar la cantidad de energía eléctrica utilizada para la iluminación de los cultivos controlando automáticamente las fuentes de luz artificial de acuerdo a parámetros como intensidad de luz ambiente, temperatura y humedad provistos por una red de sensores instalados en el invernadero. Los datos de rendimiento del cultivo se combinan con los datos históricos de las decisiones tomadas de forma de

generar un sistema re-alimentado que se estabilice en un óptimo entre desempeño y costos de luz aplicada.
- Cadena de distribución: Desde la siembra de frutas y verduras, pasando por el transporte, almacenamiento y comercialización, hasta el consumo, la recolección de distintos parámetros como temperatura y humedad (entre otros) sumado a la capacidad de procesamiento de los mismos, genera un valor que permite actuar dinámicamente dotando de calidad, eficiencia y seguridad al sistema en su conjunto.

Industria

- Cadena de abastecimiento: mediante el uso de etiquetas RFID, código de barra y distintos tipos de sensores, se dota a las empresas de transporte y logística de capacidades de control y monitoreo en tiempo real que les permite realizar una planificación detallada reduciendo tiempos de transporte, de almacenamiento y stocks. Los consumibles industriales pueden ser localizados y trazados a través de toda la cadena de abastecimiento, incluyendo manufactura, empacado, distribución y venta en tiempo cercano al real

- Prevención y diagnóstico de fallas: cada componente de la industria, en el mundo físico, tiene un modelo en el mundo virtual que es alimentado con información proveniente de redes de sensores que actúan sobre el y de información recolectada de forma manual. Algoritmos inteligentes procesan la información en el modelo y lo nutren con las condiciones de salud del componente, rendimiento y riesgos asociados en tiempo real. La inspección del modelo permite generar, ya sea de forma automática o manual, planes de mantenimiento y respuestas ante contingencia.
- Industrias energéticamente eficientes: los productos "verdes" son aquellos que fueron fabricados utilizando la menor cantidad de energía posible. Cada vez más, los consumidores, toman nota de esto al decidir la compra de un producto. En este sentido, las capacidades de censado inteligente (instalado en máquinas, líneas de producción y edificios) tienen un rol fundamental al permitir que estos datos se integren en las decisiones y prácticas de gestión aumentando la eficiencia de la fábrica en general y del proceso de fabricación del producto en particular.

Salud

- Cadena de suministro de medicamentos: la utilización de tecnologías como el RFID (entre otras) permite el seguimiento en todas las etapas que van desde la elaboración de los medicamentos, su transporte, almacenamiento y expedición dotando de trazabilidad a través de la cadena de producción y suministro, eliminando, o al menos reduciendo, los riesgos que apareja el consumo de medicamentos falsificados, adulterados o cuyos requerimientos de conservación no

han sido respetados durante alguna etapa de la distribución o el almacenamiento.

- Monitoreo remoto de pacientes: el uso de IoT en el cuidado de la salud trae grandes beneficios en especial en pacientes con enfermedades crónicas que requieren de supervisión constante. El monitoreo remoto transforma el cuidado de salud de pacientes mejorando la calidad del servicio, a la vez que baja los costos y mejora el foco en dotar al paciente de una mejor calidad de vida. La aplicación de inteligencia al monitoreo permite obtener una respuesta en tiempo casi real a situaciones de emergencia, a la vez que provee de información valiosa para el diagnostico y seguimiento de los pacientes. Incluso existen interfaces de administración de medicamentos que se integran al sistema, aumentando la seguridad y confiabilidad en todo el proceso.
- Weareables: Wiot (por las siglas en ingles de weareable internet of things) se define como una infraestructura tecnológica que interconecta sensores adosados al cuerpo de forma de permitir el monitoreo de factores como la salud, el bienestar, comportamiento y otros datos útiles para mejorar la calidad de vida de los individuos Hiremath, Yang, & Mankodiya (2014). Gracias a la interrelación entre los sensores adosados al cuerpo y la infraestructura clínica se permite a los médicos realizar una evaluación de los pacientes mientras estos permanecen en sus casas. Por ejemplo el monitoreo de pacientes con Parkinson permite la intervención remota de los profesionales en tiempo cercano al real.

Energía

- Redes eléctricas inteligentes: podemos definir Smart Grid como un sistema de electricidad que usa información, tecnologías de comunicación e inteligencia computacional de forma integrada a lo largo de todo el espectro del sistema de energía eléctrica, abarcando desde la generación hasta los puntos de consumo. La integración de todos los componentes en un único sistema de gestión permite tomar acciones en forma dinámica en tiempo casi real con el objetivo de realizar un uso eficiente y racional de la energía eléctrica.
- Energía inteligente: Tradicionalmente en los sistemas Smart Grid el foco está puesto en el sector de la Electricidad (generación y distribución), el enfoque de energía Inteligente (Smart Energy en inglés) es más amplio combinando el sector eléctrico, el sector térmico (refrigeración/calefacción) y el sector del transporte con distintas opciones de almacenamiento temporal de forma de obtener la flexibilidad necesaria para integrarlos con distintas formas de generación de energías renovables fluctuantes por naturaleza. La transición entre el uso de los sistemas de generación de energías no renovables a la generación renovable requiere repensar el sistema energético en su totalidad tanto en la generación como en el consumo, la IoT puede proveer la infraestructura necesaria para habilitar esto.

Transporte

- Vehículos de conducción asistida / autónoma: mediante la integración de dispositivos incorporados en el vehículo entre sí y con los propios de otros vehículos, sumado a los incorporados en las vías de tránsito y distintos actores intervinientes, puede formarse una red dinámica de conocimiento que permite tomar decisiones que asistan o, incluso, reemplacen al conductor en su tarea. Creando así un entorno de transporte más eficiente y seguro, optimizando incluso la utilización de las vías de tránsito y generando un importante ahorro de combustible (con el conllevado impacto positivo en el medioambiente).
- Sistema Ferroviario: La aplicación de sensores, actuadores y capacidad de procesamiento en el sistema ferroviario promete mejorar la confiabilidad, disponibilidad, mantenibilidad y seguridad (RAMS) del servicio de trenes. Por ejemplo, la información recolectada sobre la utilización de los componentes de un tren junto con la proveniente de otros componentes del sistema, puede ser utilizada para mantenimiento preventivo de sus componentes (agregando confiabilidad al sistema y seguridad) confeccionando, y adaptando en forma dinámica, planes de operación y mantenimiento. La incorporación de fuentes de datos externas, y sensores propios, más toda la información histórica permite tomar decisiones en tiempo casi real para optimizar tiempos y recorridos (reduciendo o aumentando la velocidad si las condiciones de clima y el grado de mantenimiento de los componentes del tren lo requieren, por ejemplo).

4.2 Grandes Datos

En los últimos años se han producido grandes cambios en los modelos utilizados por las empresas para hacer análisis de sus registros históricos del desempeño de sus negocios (datos) y obtener, de esa forma, conocimientos que alimenten sus procesos de planificación (acciones). Los conjuntos de datos utilizados anteriormente, se limitaban a los que podían ser capturados durante los procesos estándares de negocios: transacciones, encuestas, registros de actividad de usuario en los sistemas, entre otros. La cantidad de datos en estos sistemas, si bien era grande, podía ser manejada de forma de cumplimentar los objetivos propuestos a los mismos. El advenimiento de Internet de las Cosas (IoT) hizo que, tanto la cantidad de datos que alimenta el proceso haya aumentado exponencialmente, como que las características de los mismos haya aumentado en complejidad (se conjugan datos de tipo semi-estructurados y desestructurados con los tradicionales datos estructurados). A las fuentes de datos tradicionales, se le suman: información de diversa índole generados por dispositivos móviles (como teléfonos inteligentes, tablets, sistemas de GPS, lectores de RFID, entre otros), información del usuario proveniente de redes sociales, datos de marketing online, datos de seguimiento del usuario en sitios de internet, solo por dar unos ejemplos. La confluencia de estas fuentes generan un volumen y una heterogeneidad de datos que exceden ampliamente las capacidades de tratamiento que poseen las herramientas

tradicionales, se requieren, por lo tanto, nuevas estrategias para poder satisfacer estos requerimientos.

4.2.1 Características principales Grandes Datos

La primera mención del término Big Data se atribuye a Doug Laney quien, en el año 2001, mientras trabajaba para la empresa META group (adquirida por Gartner Inc. cuatro años más tarde) lo utilizó en su reporte "3-d data management: controlling data volume, velocity and variety" Laney (2001) reflexionando sobre los desafíos relacionados con el tratamiento de grandes cantidades de datos en el campo del eCommerce donde reconoce la existencia de tres dimensiones: volumen, velocidad y variedad. El modelo de Laney, que se conoce como modelo "3Vs", ha sido ampliamente utilizado para describir Grandes Datos hasta la actualidad. Brevemente, podemos enumerar los atributos que componen el modelo como:

- Volumen: con la generación y colección de datos masivos, los mismos escalan hasta volverse considerablemente grandes
- Velocidad: la recolección y análisis de los datos debe realizarse de manera acotada en el tiempo, debe ser rápida y oportuna de manera de poder aprovechar al máximo el valor comercial generado.
- Variedad: indica que se opera con variedad de datos, tanto semiestructurados y desestructurados (audio, video, texto, etc.) como los datos estructurados tradicionales.

Si bien el reporte de Laney no define explícitamente Grandes Datos, fue utilizado en numerosos trabajos tanto académicos, como comerciales, que sí intentaron definir con más exactitud el concepto. Una de las definiciones más conocidas es la de International Data Corporation IDC (uno de los participantes más conocidos del campo) quienes en 2011 definieron "Grandes Datos describen una nueva generación de tecnologías y arquitecturas diseñadas para extraer, de manera económica, **valor** de grandes **volúmenes** de amplia **variedad** de datos, habilitando a la captura, el descubrimiento y el análisis en alta **velocidad**" (Gantz & Reinsel (2011) IDC iview pagina 6). Esta definición sintetiza las características de Grandes Datos en: Volumen, Variedad, Velocidad y Valor (agregando una "V" adicional a las 3V de Laney, el Valor). Esta definición es ampliamente reconocida ya que agrega una arista fundamental a la cuestión, el propósito y la necesidad de la actividad: el valor generado.

Algunos autores Arun & Jabasheela (2014) agregan, incluso, una característica adicional a las 4V mencionadas, se trata de la quinta V: Veracidad. La veracidad de los datos se refiere a los sesgos, el ruido y anormalidades en la información. Es el indicador de que el dato que se procesa sea significativo para el problema que se está analizando.

A modo de conclusión, coincidimos con el análisis hecho "A formal definition of Big Data based on its essential features,"De Mauro, Greco, & Grimaldi (2016) en que no

existe un consenso en cuanto a una definición única de Grandes Datos, las mismas se pueden agrupar en cuatro categorías de acuerdo donde esté puesto el foco, a saber: atributos de los datos, necesidades tecnológicas, superación de umbrales e impacto social. Retomando, nos parecen suficientemente abarcativas a efectos del presente trabajo, las cinco características que puntualizamos: Volumen, Variedad, Velocidad, Veracidad y Valor.

La empresa DOMO elabora anualmente un reporte llamado "Data never sleeps" («Data Never Sleeps 6 | Domo», 2018) en el que muestran el volumen y la velocidad de los datos generados en línea por minuto en el planeta. Año tras año, en cada reporte, se comprueba una tendencia marcadamente ascendente. Según el reporte del año 2018, para el año 2020 se espera que cada habitante de la tierra genere 1,7 mb de datos por segundo. Solo para dar unos ejemplos, durante un minuto del 2018: los usuarios de twitter envían 473.430 mensajes, se envían 12.986.111 mensajes de texto, se suben a instagram 49.380 fotos, los habitantes de EEUU usan 3.138.420 GB de información (N. de T. se presume que el término "americans" se utiliza en ese sentido). La cantidad de personas conectadas a Internet aumentó de 2.5 a 3.8 billones. El volumen de datos actual, y el crecimiento en él esperado, genera un gran desafío a la hora de recolectar, almacenar y analizar los datos con el propósito de obtener información valiosa para la toma de decisiones.

El aumento en las capacidades necesarias para procesar el volumen de datos de Grandes Datos, comparado con los volúmenes manejados tradicionalmente, exceden las capacidades que podría proveer el escalamiento vertical de los sistemas de almacenamiento, ya sea de única instancia como en cluster. Las bases de datos tradicionales imponen limitaciones adicionales para asegurar consistencia transaccional en uno o más servidores de base de datos (cuando operan en cluster), esto puede ser crítico para algunos sistemas, pero cuando el tamaño del conjunto de datos escala los sistemas de base de datos tradicionales dejan de ser adecuados requiriéndose enfoques alternativos para el almacenamiento y recuperación. Es por esto que, cuando la cantidad de datos escala, para superar estas limitaciones se requiere utilizar arquitecturas específicas para el tratamiento de Grandes Datos.

Ciclo de operación en Grandes Datos El procesamiento de Grandes Datos puede contextualizarse como un proceso de cinco pasos que permite extraer entendimiento y generar información accionable. Los cinco pasos mencionados se configuran en forma de ciclo, estos son:

- **Recolección**: los datos se recolectan desde fuentes dispares como sensores, archivos de bitácora de operación, registros de actividad en redes sociales, registros de clicks, fuentes de terceras partes, etc. Este paso suele ejecutar actividades tradicionales de ETL (por las siglas en inglés de extracción, transformación y carga), pero en una escala grande. Si las fuentes entregan grandes ráfagas de datos, los mismos deben ser rápidamente almacenados para su

posterior procesamiento. La estrategia de almacenamiento es crítica, debe tenerse en cuenta que el paso de procesamiento necesitará acceso rápido a los datos.

- **Procesamiento**: El procesamiento puede llevarse a cabo por lotes (tomando una "foto" de los datos) o directamente considerando los datos como un continuo (cada día es más común el requerimiento de que los datos sean procesados de manera cercana a tiempo real, esto se vuelve especialmente complejo cuando la velocidad de los datos es alta). Entre los patrones más utilizados en procesamiento de Grandes Datos se encuentra MapReduce, que funciona procesando subconjuntos de los datos de forma distribuida y combinando luego el resultado. De acuerdo a la calidad de los datos y el dominio del problema, en esta fase suelen incluirse actividades de limpieza, integración y cambios de representación.
- **Extracción**: En este paso se realizan consultas y actividades de análisis sobre los datos procesados con el fin de encontrar métricas y KPIs (por la siglas en inglés de indicadores clave de rendimiento). Este paso puede realizarse en un paso o requerir varias pasadas sobre los datos procesados. Existen técnicas de análisis en tiempo real que se pueden utilizar en los datos como un continuo, son complejas pero tienen la ventaja de permitir obtener resultados intermedios aún cuando los datos no terminaron de arribar.
- **Visualización/Interpretación**: El propósito de esta etapa es entregar una representación intuitiva y efectiva del conocimiento generado, habilitando su utilización por parte del usuario final. Es una tarea especialmente difícil dado el volumen y la cardinalidad de los datos, es importante ya que aporta usabilidad y valor al sistema.
- **Accionado**: Una vez que se obtuvo información a partir de los datos analizados, se pueden tomar acciones (toma de decisiones). Las mismas suelen afectar el sistema en observación, de forma que modifican los datos que ingresan al primer paso, constituyendo de esta manera un ciclo realimentado.

Tipos de análisis en Grandes Datos Al describir los ciclos de operación hablamos del Accionado. Es en este punto donde se obtienen las "conclusiones" y se plantean las acciones posibles. Según el modo en que la información obtenida pasa a tener sentido, podemos clasificar el proceso de análisis de Grandes Datos en cuatro tipos (los dos primeros tienen que ver con acciones a futuro, los otros con el pasado):

- Prescriptivo: el análisis prescriptivo se ofrecen varios cursos de acción posibles, y se detalla el que y el como se llevaron a cabo los pasos en el análisis. En análisis prescriptivo se presta especial atención a preguntas específicas.
- Predictivo: el análisis se lleva a cabo para ganar información valiosa que permita predecir resultados futuros y tendencias. Es un procedimiento que no asegura certeza, pero pronostica lo que podría pasar en el futuro.
- Diagnostico: el resultado del análisis permite entender porqué pasó alguna cosa y porque fueron elegidos los pasos tomados. Provee un mejor entendimiento de

situaciones específicas y permite contestar algunas preguntas relacionadas con ellas.

- Descriptivo: el proceso permite descubrir patrones ocultos estudiando conjuntos de datos del pasado y del presente. Describe los datos crudos y presenta algo interpretable por el usuario.

Arquitectura para Grandes Datos No existe una arquitectura única para el manejo de Grandes Datos, la misma varía de acuerdo al tipo de problema o dominio de aplicación y a las restricciones existentes (como tecnologías disponibles, recursos materiales, capacidades del equipo, etc). Sin embargo, en general, se encuentran, definidas en mayor o menor medida, las siguientes capas (ordenadas de menor a mayor nivel):

- **Capa de Infraestructura** Es el nivel más bajo, está compuesto por los equipos que proporcionan capacidad de cómputo, así como medios de comunicación y dispositivos de almacenamiento. Puede estar compuesta por equipos físicos y/o virtualizados (ya sea Sistemas en la Nube privada o pública), o una combinación de ambos. La virtualización no es un requisito sine qua non, pero es muy interesante ya que provee capacidades de escalamiento y permite reasignar mejor los recursos de forma dinámica.
- **Capa de Almacenamiento** En la capa de almacenamiento, los sistemas de Grandes Datos, suelen incluir una combinación de tecnologías de almacenamiento, entre otras: almacenamiento de archivos planos para datos no estructurados (imágenes, sonido, video, etc.), almacenamiento semi-estructurado como bases de datos NoSql y almacenamiento estructurado (usualmente bases de datos relacionales). Esta capa, por su velocidad de respuesta comparada con las otras, suele ser el cuello de botella de todo el sistema. Se recomienda ubicar los dispositivos de almacenamiento "cerca" de la capa de procesamiento, de forma tal de reducir el impacto de las transferencias de datos por la red.
- **Capa de Procesamiento** La capa de procesamiento es la encargada de proveer soporte para consultas, análisis de datos y flujo de trabajo. Entre sus responsabilidades, suelen encontrarse: ejecución de consultas y/o código de análisis de datos hecho a medida; frameworks de procesamiento de canales de datos (encargados de guiar los datos durante operaciones como captura, transformación, normalización, validación y almacenamiento); frameworks de flujo de trabajo que coordinan múltiples trabajos y fuentes de datos en un resultado único (aquí se suele decidir entre patrones de integración mediante orquestación o coreografía). Esta capa es donde ocurre la mayor parte de la exploración y análisis de datos, generando conjuntos de datos que son posteriormente utilizados por otras tareas que realizan análisis más extensos.
- **Capa de Análisis Interactivo** Mientras que la capa de procesamiento requiere al usuario conocimientos y habilidades de bajo nivel (usualmente en manos de programadores o analistas de datos), la capa de análisis interactivo puede ser operada por usuarios que no poseen esos conocimientos. Esta capa, que puede

existir o no, expone la información a usuarios finales a través de visualización de datos, consultas adhoc, exportación de conjuntos de datos a almacenamientos de baja complejidad (como hojas de datos, por ejemplo).

- **Capa de Aplicación** Expone los resultados del análisis en forma básica o bien ejecuta directamente las acciones (cuando esto aplica). Las salidas suelen ser: tableros de control, gráficos y herramientas de exploración de alto nivel, reportes sumarizados. Las aplicaciones suelen ser construidas a medida (ya sea dentro de la empresa o por proveedores externos), u ofrecidas en la nube bajo el esquema de Software Como Servicio (SAAS), por ejemplo.

4.2.2 Desafíos Grandes Datos

Si bien las posibilidades ofrecidas por Grandes Datos son muy amplias, existe una brecha importante entre los potenciales de esta técnica y los beneficios que se están obteniendo en la actualidad. Esto se debe a que hace falta avanzar en la resolución de ciertos problemas que frenan su implementación, o que, por lo menos, disminuyen grado de avance de la misma. Estos problemas, que afectan a una o más fases del ciclo de operación en Grandes Datos (que describimos en la introducción del capítulo), son: la heterogeneidad de los datos, la escala, la oportunidad de la información obtenida, la complejidad y la privacidad Agrawal et al. (2012).

Heterogeneidad y datos incompletos Como mencionamos anteriormente, durante la recolección de datos se cuenta con fuentes, generalmente, dispares introduciéndose datos heterogéneos dentro del flujo de procesamiento. Dado que los algoritmos de análisis requieren operar sobre datos homogéneos se constituye un desafió el estructurar y uniformizar las entradas (recordamos que los datos provenientes de las fuentes de entrada pueden ser estructurados, semi-estructurados o desestructurados). Además de lo mencionado, el trabajo con fuentes semi-estructuradas y desestructuradas tiene connotaciones en cuanto a la representación, transmisión, acceso que deben ser tenidas en cuenta, dado que en los sistemas informáticos estos atributos suelen ser más eficientes cuando se cuenta con datos estructurados.

Otras de las cuestiones que debemos tener en cuenta, dado la naturaleza de las fuentes de datos con las que operamos, son los datos incompletos o, directamente faltantes. Es común que algún atributo que integra un registro no haya podido ser recolectado (por ejemplo, una sonda averiada momentáneamente que no entrega una de las mediciones que componen un registro de la línea temporal) o que directamente ese registro sea integrado en el ciclo de procesamiento. Si bien existen procedimientos, que usualmente se ejecutan al inicio de la fase de procesamiento, que limpian los datos y corrigen errores, usualmente siguen existiendo datos incompletos después de aplicarlos. Es por esto que se hace necesario la creación de algoritmos de análisis que pueden operar con datos datos incompletos, faltantes o erróneos.

Escalamiento El manejo de grandes volúmenes de datos fue tradicionalmente un desafío en sí, aunque el aumento constante en capacidad de procesamiento y almacenamiento (se estima que los ambos se duplican cada dos años, siguiendo la ley de Moore) servía como mitigación al problema. La diferencia que se plantea con Grandes Datos es que la taza de aumento del volumen de datos es mucho mayor, se puede estimar la misma como de orden exponencial, la mitigación mencionada dejó entonces de ser efectiva, planteando nuevos desafíos.

Sobre este punto, una de las cuestiones a tener en cuenta es la forma que está tomando la evolución tecnológica sobre el aumento de capacidad de procesamiento. Si bien, en décadas anteriores el aumento en la frecuencia de reloj de los procesadores a seguido las previsiones de la ley de Moore, en la última década se ha alcanzado valores cercanos al límite y dado los avances en cuanto a densidad de integración, se ha dado paso al aumento de cantidad de núcleos en un solo procesador. Esto ha requerido, del software, el avance en algoritmos que optimicen la gestión multi-núcleo. En la actualidad, están empezando a tener importancia las restricciones de consumo eléctrico en los sistemas informáticos (se piensa que en un futuro por estas restricciones no será posible la utilización de toda la capacidad de procesamiento instalada, a esto se denomina dark silicon), dando importancia a la generación de algoritmos que actúen activamente sobre el consumo eléctrico optimizando la utilización de recursos.

Otra cuestión a tener en cuenta es el pasaje desde la utilización de recursos de calculo propios a la utilización de recursos de computación en la nube. Este tipo de procesamiento permite repartir carga de procesamiento en grandes clusters y sumarizar los resultantes. Dado que el costo depende del volumen y velocidad de procesamiento utilizados, los algoritmos deberían variar estos parámetros eficientemente de acuerdo a las metas de tiempos de respuesta y costos aplicadas. Además de esto, los algoritmos deben ser lo tolerantes a los fallos que, dado que la infraestructura es mucho más compleja, son de ocurrencia más frecuente.

La última cuestión relativa al escalamiento tiene que ver con la evolución de los dispositivos de almacenamiento, tradicionalmente se utilizaron medios magnéticos que tienen un buen desempeño en las lectoras/escrituras secuenciales y un desempeño pobre en los accesos aleatorios. Como consecuencia, todos los algoritmos de almacenamiento y consulta operaban bajo esas premisas. La utilización masiva de dispositivos de estado sólido, que tienen un excelente desempeño en lecturas/escrituras aleatorias, esto es un beneficio que en la actualidad no es aprovechado por los algoritmos mencionados, marcando otro desafío en la ruta a la utilización eficiente de recursos.

Oportunidad El aumento en volumen procesado trae aparejado una variable adicional en cuanto al tiempo que se demora en procesar los datos de entrada y obtener información accionable.

Muchas aplicaciones imponen restricciones en cuanto a el tiempo que puede transcurrir entre que el dato ingresa y la decisión se toma. Cuando operamos con grandes datos,

no es posible, hacer un análisis total de los registros históricos relacionados con cada dato de entrada, de forma de obtener una salida. En este caso, es posible utilizar datos parciales pre-computados y realizar un calculo incremental en tiempo real, de forma de obtener una respuesta oportuna.

Dado que examinar todo el conjunto de datos para encontrar algunos elementos que satisfagan un criterio específico sería impracticable, el enfoque utilizado es pre-computar estructuras de índice que luego permitan acceder directamente a los datos buscados. La limitación de este enfoque es que las estructuras de índice deben ser diseñadas para satisfacer criterios de problemas específicos. Dado que la complejidad de dichos problemas aumenta constantemente y los tiempos de respuesta requeridos son cada vez menores, la creación de nuevas estructuras de índice cada vez más complejas resulta un gran desafío.

Privacidad La privacidad es un tema de por sí complejo en cualquier sistema informático, pero al hablar de Grandes Datos se vuelve aún más complicado. En principio las soluciones implementadas deben, básicamente, cumplir las regulaciones legales del contexto donde se utilicen. Además, y no menos importante, deben respetar las implicaciones sociales en dicho contexto (existen temores en las personas sobre la recolección de información personal y, sobre todo, su utilización, estos temores están, muchas veces fundados en ciertos temas que los medios de información divulgaron sobre el uso que hacen grandes empresas de la información que poseen sobre sus usuarios). Como puede verse, el desafío de la privacidad excede lo técnico.

Existen, además, desafíos específicos de acuerdo al dominio de aplicación. Por ejemplo, en los servicios basados en localización, es difícil garantizar el anonimato de las localizaciones enviadas. Aunque no se envíe la identidad del usuario, la misma podría ser inferida ya que se ha mostrado que existe una correlación entre la identidad y las localizaciones frecuentes.

Adicionalmente, existe un balance entre limitar la cantidad de información privada que se comparte (de forma de garantizar confidencialidad) y la utilidad de la información compartida. Encontrar la forma de limitar la cantidad de información privada necesaria para generar decisiones útiles, es particularmente un tema de estudio desafiante. Paradigmas como el de la privacidad diferencial, constituyen un avance interesante en esta cuestión, aunque su aplicación reduce en gran medida la utilidad de los datos recolectados, es necesario encontrar enfoques alternativos que mejoren el balance citado anteriormente.

Asistencia humana Si bien los avances en extracción del conocimiento son muy grandes, y estos procesos pueden operar de forma automática, en ciertos dominios es interesante, o muchas veces necesaria, la interacción con el humano dado que este posee habilidades para detectar patrones difíciles de encontrar con métodos automáticos. Es

entonces interesante incluir la asistencia humana dentro del flujo de Grandes Datos, al menos en la capa de procesamiento y extracción.

Existen propuestas que avanzan en este sentido. Por ejemplo en Análisis Visual donde se trata de incluir al experto humano en las etapas de modelado y análisis. Está pendiente ampliar esta a participación a las otras etapas del proceso (donde este sea valioso de acuerdo al dominio de aplicación particular).

En algunos casos es fructífera la colaboración de expertos de múltiples disciplinas. El sistema de análisis debería poder gestionar la colaboración de varios expertos de distintas ramas (que muchas veces, están incluso dispersos físicamente), permitirles la exploración de resultados e incorporar sus decisiones dentro del lazo realimentando el proceso.

4.2.3 Aplicaciones Grandes Datos

La posibilidad de tomar decisiones informadas basadas en procedimientos de Grandes Datos resulta muy atractiva para muchas áreas de aplicación, incluso fuera de las áreas estrictamente económicas/comerciales (Bancos, Inversiones, Marketing) donde su potencial es obvio. Áreas de implicancia social como educación, salud, gobierno y comunicaciones se ven altamente beneficiadas, también, de su utilización.

En esta sección mencionaremos algunos ejemplos de aplicación de Grandes Datos, en algunas áreas específicas, el espectro de aplicación de esta nueva tecnología es tan amplio así como disruptiva ha sido su aparición, por tanto las aplicaciones posibles siguen siendo aún insospechadas.

Comportamiento Social Como fuente de datos, una de las más extensamente disponibles y más estructurada son los textos, o sea, palabras. La idea es aplicar un análisis masivo que permita encontrar, en grandes volúmenes de palabras, conocimiento que permita predecir acciones o actividades.

La fuente de datos por excelencia, a la hora de buscar conocimiento en palabras, son las redes sociales. En la actualidad, las actualizaciones de estado de Facebook y Twitter son la fuente más común, seguida por blogs y foros en línea. Son obvias las ventajas de la toma informada de decisiones de Marketing o Planificación de ventas mediante las técnicas de análisis de Grandes Datos aplicados a redes sociales, aunque existen un sinnúmero de otras aplicaciones no tan evidentes. Por ejemplo, (O'Reilly Radar, 2011) Team (2011) describe que mediante el análisis de tendencias en buscadores de Internet pudo anticiparse la gripe porcina dos semanas antes de que hubiera información del Centro de Control de Enfermedades de Estados Unidos, lo mismo para el brote de Dengue en Brasil (donde la información oficial tardó, incluso, mucho más en llegar).

Usar estas técnicas como fuente de conocimiento para el al campo de las Ciencias Humanas genera posibilidades ni siquiera imaginadas hasta ahora, tomemos un ejemplo:

utilizando una Red Social como Instagram se combinan las imágenes, con la interacción entre los usuarios que han generado (comentarios), con la metadata que el usuario creó (etiquetas), con información geoespacial (donde fue creada) y otras informaciones transaccionales como la fecha de creación. El resultado de explotar esta información y ponerlo sobre una línea de tiempo, es infinitamente más grande que toda la herencia cultural digitalizada hasta el momento.

Al utilizar redes sociales como fuente de datos, debe tenerse en cuenta la diferencia existente entre el comportamiento social en el mundo digital y la vida real de las personas. En general, las personas al utilizar redes sociales tienden a construir la imagen que muestran en esos ámbitos. Eso construye un sesgo con respecto a los deseos y pensamientos reales que esas personas poseen, que debería ser tenido en cuenta al momento de obtener conocimiento de estos datos.

Seguimiento de localizaciones La ubicuidad de los dispositivos de comunicación móvil (equipados con sistemas de posicionamiento ya sea basados en GPS o con servicios de localización aproximada) da como resultado una fuente de datos extraordinaria sobre movilidad humana. La aplicación de análisis de Grandes Datos es la técnica por excelencia para explotar este volumen de datos de forma efectiva.

La información obtenida de esta manera ha demostrado ser efectiva para predecir movimientos de poblaciones humanas tanto en situaciones estables (como movilidad diaria o migraciones) como en situaciones de emergencia como desastres naturales (permitiendo a los gobiernos elaborar planes de acción en esas situaciones, por ejemplo).

En el área de la salud, la posibilidad de conocer los patrones de movilidad de las personas se puede aplicar al análisis de la diseminación de enfermedades infecciosas, pudiendo generar planes de respuesta y control de las mismas.

El análisis en tiempo cercano al real permite obtener una fotografía del tráfico en grandes centros urbanos permitiendo diseñar sistemas dinámicos de gestión del mismo, o responder ante sucesos como accidentes de tránsito o eventos inesperados.

En prevención del crimen, se han realizado experiencias combinando distintas fuentes como datos geoespaciales y demográficos, información estadística histórica sobre hechos delictivos y datos de redes sociales, obteniendo predicciones sobre actividades delictivas en centros urbanos.

Reducción de la incertidumbre en la Naturaleza La incertidumbre se define como las situaciones que implican información imperfecta o faltante. La naturaleza es una fuente de incertidumbre en si. El análisis de Grandes Datos tiene potencial de ser aplicado para reducir la incertidumbre en la naturaleza. Una perforación petrolera no exitosa trae a la empresa que la realiza una perdida valuada en varias decenas de millones de dólares. Mediante la aplicación de análisis de Grandes Datos en fuentes de sismología por

reflexión de más de 50 terabytes, la empresa Chevron pudo aumentar las posibilidades de éxito hasta 1 en 3 en las perforaciones en el Golfo de México.

El cambio climático es uno de los mayores desafíos de la humanidad en este siglo. Por su tamaño las fuentes de datos geoespaciales exceden las posibilidades de procesamiento tradicionales siendo el análisis de Grandes Datos el principal candidato para su tratamiento Lee & Kang (2015). El estudio del cambio climático requiere de la confluencia de investigadores de distintas disciplinas de la ciencia quienes han encontrado en la ciencia de datos una herramienta que les permite obtener información sobre distintos escenarios futuros posibles de cambio climático. Distintas agencias gubernamentales y no gubernamentales (como la US EPA, NASA, entre otras) procesan y distribuyen información valiosa sobre el tema pudiendo generarla y sintetizarla gracias a técnicas de grandes datos. El estudio de la relación entre cambio climático y biodiversidad, y cambio climático y salud no había sido posible hasta el momento, y ha avanzado notablemente gracias a la aplicación de estas técnicas.

Interacciones sociales Según Helbing y Balietti Helbing & Balietti (2011) las transacciones digitales son las huellas omnipresentes de la interacción social. El análisis de transacciones comerciales es, sin duda, uno de los aspectos más explotados de Grandes Datos. Igualmente, la aplicación de esta tecnología, excede en gran medida el objetivo de optimizar la rentabilidad comercial. El análisis de las interacciones en sistemas socioeconómicos en Observatorios de Crisis (que operan con datos dispares como transacciones económicas y financieras, conflictos, diseminación de enfermedades, etc) pueden predecir crisis o identificar debilidades sistémicas, y ayudar a evitar o mitigar el impacto de una crisis.

Mediante la utilización de Grandes Datos en transacciones digitales ha logrado disminuirse en gran medida el tiempo requerido en disponer de información, que habitualmente, tardaría semanas, meses o hasta podría estar disponible cuando ya no tiene valor (oportunidad de la información). Por ejemplo, mediante el análisis de las búsquedas en un popular motor de búsquedas Web (como Google) pueden estimarse la aparición y dispersión geográfica de epidemias de Gripe en tiempo cercano al real («Google Flu Trends», 2014), mientras que aguardar las estadísticas tradicionales de agencias gubernamentales podría demorar semanas (en el mejor de los casos).

En cuanto a aplicaciones de Marketing, es conocido el caso de la empresa Target que, entrenó un modelo con datos sobre las ventas en sus locales y los registros de nacimientos (para sus clientes). Con este modelo pudo, según los comportamientos de los clientes en sus tiendas, predecir cuales de sus clientes estaban en proceso de embarazo (incluso antes de que sus allegados lo supieran, tal es así que trascendió en la prensa el caso de un padre que se entera que su hija adolescente está embarazada porque recibió una carta de promoción de Target) y su fecha estimada de parto. Target fue capaz, de esta manera, de dirigir promociones para esos clientes en especifico.

La empresa GT Nexus concentra las transacciones relacionadas con la Cadena de Suministro (ordenes de compra, pagos, manufactura, logística, entre otras cosas) de cien empresas que están entre las más grandes del mundo (cada una de estas empresas, tiene a su vez, más de mil empresas relacionadas con sus transacciones) Hardy (2012). La aplicación de Grandes Datos permite a esta empresa ofrecer una abstracción de la actividad a nivel mundial (algo así como una imagen actualizada en tiempo cercano al real) que permite a sus clientes (mediante una plataforma en la nube con posibilidades avanzadas de visualización y consulta) acceder a conocimiento estratégico sobre el mercado global. Mediante la aplicación de análisis prescriptivo una empresa podría, por ejemplo, planificar el ciclo completo de un nuevo producto optimizando con precisión cada punto de la cadena de suministro.

Seguimiento del comportamiento La incorporación a la vida cotidiana de dispositivos conectados que permiten capturar el comportamiento de las personas (teléfonos inteligentes, wearables, relojes inteligentes, sensores de IoT, etc), sumado a la capacidad que, gracias a el escalamiento tecnológico, han adquirido los sistemas informáticos de uso habitual, para registrar y compartir información sobre su uso (bitácoras de accionamientos en sitios web, bitácoras de uso de juegos línea, etc.), hacen de Grandes Datos un excelente candidato para integrar esos registros y obtener conocimiento (generar información valiosa para distintas aplicaciones) de ellos. Se hace interesante, en este contexto, el foco sobre el Comportamiento Anormal, siendo este definido como el comportamiento que se separa del comportamiento colectivo promedio.

En el área de salud, los avances realizados en sensores miniatura, microelectrónica y procesos de fabricación, sumados a la alimentación eléctrica inalámbrica, han permitido avanzar en la creación de sensores implantables en el cuerpo humano. Los sensores implantados se integran en una red de sensores corporales, que sumada a los sensores wearables/teléfonos inteligentes y el sensado ambiental, generan conjuntos de datos que son imposibles de tratar mediante técnicas tradicionales. Dado que sensado fisiológico es continuo y de largo plazo, impone desafíos en cuanto a su interpretación clínica (las prácticas actuales se basan en mediciones tomadas en forma esporádicas durante las visitas de los pacientes a los centros de salud). Por ejemplo, mediante la observación continua de la presión sanguínea del paciente puede construirse un perfil de comportamiento. Si bien, interpretar las señales obtenidas trae aparejado ciertas complicaciones (las mediciones deben ser evaluadas según el contexto en el que fueron tomadas) la obtención del significados fisiológicos ocultos permite entender los casos de hipertensión fuera de control, o mejorar los esquemas actuales de tratamiento de la hipertensión. Y yendo más lejos aún, los implantes inteligentes podrían incluso reaccionar suministrando drogas o actuando como estimuladores cerebrales Andreu-Perez, Poon, Merrifield, Wong, & Yang (2015).

Los juegos en línea del tipo multijugador crean mundos virtuales donde las personas interactúan. El registro del comportamiento de esas personas en el mundo virtual genera una fuente de conocimiento sobre dinámica social, en ella se reflejan las acciones entre

los usuarios a lo largo de las sesiones de juego permitiendo encontrar comportamientos sociales positivos y/o negativos entre los jugadores. Por ejemplo, mediante técnicas de grandes datos sobre estas fuentes pueden generarse teorías sobre Comportamientos Tóxicos que permiten detectar, prevenir y contrarrestar comportamientos como cyberbullying (acoso virtual), griefing, mischief y cheating Kwak, Blackburn, & Han (2015). En general, la práctica del acoso virtual se asocia con depresión, ansiedad y se ha observado que puede terminar en acciones drásticas como el suicidio. Además, las víctimas de estas prácticas suelen sentir efectos emocionales que persisten en el mundo real.

4.3 Extracción del conocimiento

El interés por la Inteligencia Artificial no es algo reciente, desde 1950 se comienzan a desarrollar grandes expectativas en la sociedad que llevan a la fantasía de creer que las computadoras podrían alcanzar niveles de inteligencia humanos mediante esta disciplina. Estas expectativas, claramente, no son satisfechas como tal, y a lo largo de las siguientes tres décadas la Inteligencia Artificial, como disciplina, ingresa en una fase de amesetamiento. más recientemente, gracias al éxito práctico en campos como Machine Learning y Extracción del Conocimiento la disciplina vuelve a estar en auge, con metas más concretas esta vez Holzinger, Kieseberg, Weippl, & Tjoa (2018). Una buena cantidad de casos de éxito pueden ser vistos hoy por el público general en distintos dominios de aplicación de la vida diaria, igualmente muchos de los científicos que trabajan en el campo siguen sin estar cómodos con las implicaciones del término “inteligencia artificial”.

En la extracción del conocimiento de grandes datos se suelen utilizar algoritmos cuya aplicación ha sido exitosa previamente, en minería de datos. Estos algoritmos han evolucionado y se han adaptado a los nuevos requerimientos impuestos por esta disciplina donde se espera de ellos altas velocidades de procesamiento con costos mínimos en altos volúmenes de información, altas velocidades de entrada y gran variedad de datos.

El aprendizaje automático es una rama de la Inteligencia Artificial. Según Tom M. Mitchell, el aprendizaje automático es el estudio de los algoritmos que permiten que los programas de informáticos mejoren automáticamente a través de la experiencia Mitchell (1997). Definición: se dice que un programa informático aprende de la experiencia E con respecto a una clase de tareas T y una medida de desempeño P, si su medición de desempeño P en tareas T mejora con experiencia E.

4.3.1 Características principales EC

Tipos de aprendizaje Existen varias maneras de agrupar o categorizar los métodos de extracción de datos, una de ellas, y quizás la más general es según el tipo de aprendizaje, aquí tenemos:

Aprendizaje supervisado: En este tipo de aprendizaje se cuenta con un conjunto de variables de entrada (a las que llamaremos X) y una variable de salida (la llamaremos Y - también conocido como atributo de clase). Se utiliza un algoritmo para obtener la función de mapea entra las entradas y la salida. El objetivo en esta fase del proceso (entrenamiento) es obtener una aproximación, tan precisa como sea posible, a esta función Y=f(X). Para esto se trabaja con un conjunto de datos etiquetado (clasificado) previamente, conocido como conjunto de datos de entrenamiento. Cada objeto del conjunto de entrenamiento posee un atributo (Y) que indica su clase, el valor de este atributo puede haber provenido de el ingreso manual o haber sido obtenido de información histórica recolectada previamente. Dado que se conoce el valor de salida, se puede ejecutar el algoritmo sobre el conjunto de entrenamiento y evaluar la función de mapeo hasta que esta alcance un nivel de confianza aceptable. Es por esto que este tipo de aprendizaje se denomina supervisado, ya que en la fase de entrenamiento se cuenta con información como para evaluar las predicciones obtenidas contrastándolas con las respuestas correctas.

Una vez superada la fase de entrenamiento del modelo, el mismo puede ser aplicado a objetos que carecen de valor en el atributo de clase con el propósito de obtenerlo (o sea, clasificarlos). Obviamente la eficacia del proceso está estrictamente vinculada a la cantidad de objetos clasificados que contenga el conjunto de entrenamiento, y por ende a la cantidad de trabajo, en general manual, aplicado a su construcción.

Dentro de esta categoría, se delinean dos subcategorías de algoritmos de aprendizaje supervisado:

. Clasificación: Se realiza clasificación cuando el valor de la variable de salida es una variable categórica (el valor pertenece a una o más categorías sin que exista un orden intrínseco entre ellas).

. Regresión: El valor de la variable de salida pertenece al dominio de los números reales, o sea representa un punto dentro de un continuo (como por ejemplo una distancia, un costo en una medida monetaria, etc.).

Aprendizaje no supervisado: Por otro lado, la idea detrás del aprendizaje no supervisado es que la computadora aprenda encontrando patrones complejos que modelen la estructura de los datos de entrada, pero esta vez sin contar con la ayuda de un conjunto de entrenamiento con datos previamente clasificados (como es el caso del aprendizaje supervisado). Es un tipo de aprendizaje automático, en el sentido de que no requiere de intervención de los especialistas de dominio. En este caso, se da más un proceso de minería de datos que de aprendizaje en sí. En este tipo de actividad no existen las respuestas correctas o incorrectas, se trata de un procedimiento que luego de su ejecución tendrá como resultado el conocimiento o los patrones hallados entre los datos ingresados. En esta categoría encontramos los algoritmos de:

. Clustering: En este tipo de problemas se trata de descubrir agrupamientos inherentes a los datos (por ejemplo reconocimiento de agrupación entre personas en una red social).

. Reglas de asociación: Son problemas en los que se intentan descubrir reglas de asociación que describen los datos (por ejemplo los compradores del producto A luego compran el producto B).

Aprendizaje semisupervisado Por último, el aprendizaje semisupervisado se encuentra entre el aprendizaje supervisado y el no supervisado, tomando un conjunto de datos etiquetados utilizándolos para reforzar la operación en un conjunto más grande de datos no etiquetados. Dentro de esta categoría tenemos el denominado aprendizaje por refuerzo (reinforcement learning) donde se utiliza un mecanismo de recompensas reforzando la información de entrada en los casos donde el algoritmo encontró la un buen resultado en una situación particular. Una buena parte de algoritmos utilizados en aplicaciones prácticas caen dentro de esta categoría, sobretodo cuando en el dominio de aplicación resulta costoso categorizar la información, y como consecuencia, se cuenta con un conjunto pequeño de datos etiquetados.

Enfoque Generativo y enfoque Discriminativo De acuerdo al enfoque interno de los algoritmos, podemos visualizar dos enfoques: el discriminativo y el generativo. Dado un conjunto de entrenamiento donde cada instancia posee variables observadas X y la variable objetivo Y (o variable clase). Se tiene:

. Enfoque Discriminativo: En modelos probabilísticos, buscan encontrar la distribución de probabilidad condicional P(Y/X), En modelos no probabilísticos, lo que se busca es encontrar la función que mapea las variables observadas X a la variable objetivo Y. Se denominan discriminativos porque se enfocan en encontrar una probabilidad condicional que permita luego discriminar entre clases.

. Enfoque Generativo: Buscan obtener un modelo de como los datos se generan. En este sentido buscan encontrar la distribución de probabilidad conjunta P(X,Y) desde los datos de entrenamiento. En otras palabras tratan de modelar la distribución D desde la cual los datos se generaron.

Los algoritmos generativos aprenden como se estructuran y distribuyen los datos, en este sentido tienen un abordaje más general. En cambio los discriminativos aprenden a categorizar una variable objetivo Y de acuerdo a los datos de entrada X, en este sentido atacan una cuestión particular del tema.

4.3.2 Métodos / Algoritmos de extracción de conocimiento

Como mencionamos anteriormente, existen varias formas de clasificar los algoritmos utilizados para extracción de conocimiento, adicional al agrupamiento por tipo de

algoritmo (que describimos anteriormente), vamos a presentar en esta sección un agrupamiento que tiene en cuenta similitud en el funcionamiento de los algoritmos. Aunque esta forma de agruparlos es más específica que la anteriormente expuesta, hay algoritmos que clasifican en más de una categoría de las que presentaremos (esto es natural, dado que algunos están inspirados en más de un algoritmo "padre" que cae en grupos diferentes).

Algoritmos basados en Regresión Los algoritmos de extracción de conocimiento basados en regresión pertenecen a la categoría de aprendizaje supervisado, ellos intentan modelar la relación existente entre las variables de entrada y las de salida (variables objetivo del aprendizaje automático) refinándose iterativamente usando una medida de error de las predicciones hechas por el modelo. El modelo hallado permite predecir el valor de las variables objetivo para cada vector de variables de entrada. En el campo de la estadística los métodos basados en regresión son bien conocidos y han sido adoptados para su uso en aprendizaje automático desde los comienzos de esta disciplina.

Mas generalmente, dado el vector de variables de entrada x perteneciente a Xn (observaciones del conjunto de entrenamiento) con la variable objetivo t, lo que se busca es la probabilidad condicional p(t|x) que expresa el grado de incertidumbre que existe sobre el valor de t (reflejando de esta forma el hecho de que se trata de una predicción).

El modelo más simple de regresión lineal, es la combinación lineal de las variables de entrada. Se trata de una función de la forma $y(x, w) = w_o + w_1 x_1 + ... + w_d x_d$donde w0…wd son los parámetros del modelo y xi las variables de entrada. Siendo aquí el objetivo del proceso de entrenamiento encontrar los valores de los parámetros wi.

Los modelos de regresión lógica, trabajan sobre ideas similares, pero el resultado de la predicción es un atributo binario obtenido mediante la transformación de la salida de la regresión mediante una función no lineal conocida como función lógica (una función que transforma un valor real en uno binario de acuerdo a una distribución de probabilidad definida).

Dado que la combinación lineal simple genera un modelo demasiado restrictivo, suele extenderse este modelo a la combinación lineal de funciones no-lineales fijas quedando el modelo con una forma: $y(x, w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x)$ donde las funciones ϕ_j(x) se conocen como funciones base, aunque estas funciones base no son lineales y por lo tanto y(x,w) tampoco lo es, el modelo se considera lineal en el sentido estadístico dado que es una combinación lineal en cuanto a los parámetros w_j. Existen múltiples propuestas para las funciones base: en regresión polinómica estas toman la forma de potencias de x, función base gaussiana, función base sigmoidal, función base de Fourier (útil en aplicaciones que impliquen espectros de frecuencias), entre otras.

La regresión lineal de cuadrados mínimos (OLSR), toma su nombre del método que se utiliza para estimar los valores de los parámetros de la combinación lineal mencionada

en el párrafo anterior, esto es: método de los cuadrados mínimos. En este método los parámetros desconocidos se estiman minimizando la suma de los cuadrados de la desviación entre el modelo y los datos de entrenamiento. El método emplea algunas propiedades del álgebra lineal para reducir el sistema de ecuaciones planteado en uno más manejable, es posible resolver este sistema, entonces, de forma de obtener finalmente los parámetros buscados.

Algoritmos de esta familia:

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

Métodos Basados en Instancias Se trata de un método estadístico, basado en la idea de que en un conjunto de datos en general las instancias se encuentran en proximidad de otras con propiedades similares Maglogiannis, Karpouzis, & Wallace (2007). Si las instancias tienen asignado un valor en el atributo de clase, entonces el valor de las instancias que no lo tengan asignado puede ser inferido a partir de aquél asignado a su/s vecinos más próximo/s. A esta familia de algoritmos pertenece el kNN (por las siglas en inglés de k nearest neighbor) que opera seleccionando los k vecinos más próximos a la instancia que se intenta clasificar y determina el valor de clase seleccionando el valor de clase más frecuente entre ellas.

La idea detrás de estos algoritmos es representar cada instancia en el espacio n-dimensional (donde cada dimensión corresponde a un atributo de clase) y evaluar la proximidad relativa entre cada uno de los puntos representados (no es aquí importante la posición absoluta de cada punto, sino la distancia relativa entre ellos). Dada la importancia que tiene en el resultado final, la elección de la medida de distancia a utilizar es un tema clave, en general se busca que la medida utilizada minimice la distancia entre instancias con clasificación similar mientras maximice la distancia entre instancias con clasificaciones diferentes. Algunos algoritmos utilizan esquemas de peso que alteran la medida de distancia según criterios establecidos mejorando los resultados notablemente.

Los algoritmos de aprendizaje basados en instancia son del tipo perezoso (lazy-learning), o sea, retrasan la inducción o generalización hasta que se realiza el proceso de clasificación. Esta característica hace que tengan un bajo consumo de recursos en la etapa de aprendizaje, y un consumo mayor durante la clasificación (con el conllevado efecto sobre los tiempos en ambas fases). En línea con esto, un aspecto clave en la definición del modelo es la selección de los atributos a incluir en la instancia (mas atributos, significa más dimensiones, por lo tanto más requerimiento de recursos y menos velocidad de respuesta).

Algunos aspectos a tener en cuenta: grandes requerimientos de almacenamiento; la elección de la función de similitud es crítica; la selección del valor para k (los procedimientos de validación de este valor tienen un costo computacional alto).

Algoritmos de esta familia:

- Nearest Neighbor
- k-Nearest Neighbor (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

Algoritmos de Regularización Uno de los aspectos que suelen degradar la precisión de los modelos es el sobre ajuste (overfitting). Este fenómeno se da cuando, el modelo está sobre adaptado al conjunto de entrenamiento y debido a esto perdió capacidad de generalización. Uno de los causantes son los valores atípicos (outliers) o ruido en el conjunto de entrenamiento, si el entrenamiento incorpora estos valores (muestras aleatorias que no representan a los datos reales) el modelo resultante representará un conjunto de datos más complejo y menos general (con una varianza elevada).

Los algoritmos de regularización son, típicamente, una extensión de métodos de regresión a los cuales incorporan un tratamiento diseñado específicamente para reducir los efectos de la multicolinealidad sobre las capacidades predictivas del modelo. Los estimadores mínimos cuadráticos ordinarios presentan varianza mínima en la clase de los estimadores insesgados, cualquier estimador que presente menor varianza y sea lineal será sesgado. El procedimiento Riedge propone introducir una constante k>0 que permite aumentar levemente el sesgo logrando una reducción en la varianza, de forma de mejorar el error cuadrático medio de la estimación. En la práctica existen varias maneras de obtener el valor de k, las más usadas son: trazas riedge, validación cruzada, validación cruzada generalizada (GCV), etc.

Otro algoritmo de regularización muy utilizado es el LASSO (por las siglas de Least Absolute Shrinkage and Selection Operator). A diferencia de la regresión Ridge, donde el modelo final incluye las n variables del modelo original, la regularización LASSO, cuando se utilizan valores elevados del parámetro ϕ, lleva prácticamente a 0 algunos coeficientes de la estimación produciéndose una suerte de selección de las variables del modelo final (mejorando su interpretabilidad, entre otras ventajas) Por último, Elastic Net, en cierta forma, combina los dos métodos mencionados anteriormente reduciendo el impacto de las variables que no son tan influyentes. Y, LARS (Least-Angle Regression) que funciona mediante un principio similar a los de regresión por pasos (forward selection, backguard elimination, all subsets) con resultados más precisos y mejores tiempos de ejecución (una versión menos ávida - greedy - de los mismos).

Algoritmos de esta familia:

- Ridge Regression

- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)

Arboles de Decisión La familia de algoritmos Arboles de Decisión entran en la categoría de Aprendizaje Simbólico. Los arboles de decisión clasifican las instancias organizándose de acuerdo a los valores de las propiedades de las mismas. Cada nodo de este árbol representa un atributo y cada arco un valor posible del mismo. De esta forma, cuentan con un nodo raíz y el proceso de decisión queda modelado nodo por nodo hasta llegar al valor del atributo de clase.

La construcción de un árbol de decisión óptimo es un problema NP-Completo es por esto que se ha dedicado mucho esfuerzo en la generación de heurísticas que puedan llegar a un optimal.

Idealmente, en la raíz del árbol, debería ubicarse el atributo que mejor divida las instancias pertenecientes al conjunto de entrenamiento. Existen diversos métodos para seleccionar el nodo raíz. Los estudios muestran que según las características del conjunto de entrenamiento, un método puede ser preferido a otro. Una vez elegido el método de selección se lo aplica recursivamente al resto de los atributos hasta que no existan particiones posibles, utilizando un enfoque dividir y conquistar.

Un aspecto a tener en cuenta es el sobre ajuste (overfitting) que se da cuando, dado un árbol de decisión h existe uno h' cuyo error es más grande que el de h al ser evaluado contra el conjunto de entrenamiento, pero más chico cuando es evaluado contra el conjunto de datos completo, o sea el modelo no generaliza. Existen diversas formas de tratar el sobre ajuste, las más comunes son: detener la generación del árbol antes de llegar al árbol completo; ejecutar una poda sobre el árbol completo. Es oportuno nombrar el criterio Occam Razor que dice que dados dos modelos con errores de generalización similares, se debería preferir el más simple. Los árboles de decisión son naturalmente univariados, las divisiones se abordan de a un atributo por vez. Aunque existen algunos métodos que permiten construir arboles multivariados.

Una de las características más importantes de los árboles de decisión es que son descriptivos. Una persona puede comprender porque el árbol clasificó de determinada manera una instancia.

Algoritmos de esta familia:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

Redes Bayesianas Al igual que los algoritmos de aprendizaje basados en instancias (descritos anteriormente) las redes Bayesianas comparten un abordaje estadístico de clasificación, que permite predecir la probabilidad de que a una instancia le corresponda una clase. Las Redes Bayesianas son grafos dirigidos acíclicos que representan la relación de probabilidad entre las variables de instancia. Cuentan con un nodo por variable, y cada arco representa la influencia causal entre ambas variables (ambos nodos) mientras que la ausencia de ellos muestra independencia condicional entre variables.

El entrenamiento de una Red Bayesiana se divide normalmente en dos etapas: la obtención de la estructura del grafo dirigido acíclico (DAG) y la determinación de sus parámetros (usualmente las probabilidades condicionales son almacenados en matrices, a razón de una matriz por variable, esta matriz se denomina tabla de probabilidad condicional).

En la etapa de obtención de la estructura del grafo, se tienen dos posibilidades: se conoce la estructura del grafo o la misma es desconocida. La primera de las posibilidades, usualmente, se da cuando la estructura del grafo es suministrada por un experto y es fija, este es el caso más sencillo. Cuando la estructura no es conocida de antemano, se debe proceder a inducirla. Una alternativa es utilizar una función de puntuación que pueda medir que tan ajustado está el modelo con respecto a el conjunto de datos de entrenamiento, de esta manera se puede buscar dentro de un universo de grafos posibles cual es el que mejor puntuación tiene, existen diferentes propuestas en cuanto a heurística a utilizar para esta búsqueda. Otra alternativa, es utilizar prueba de hipótesis (chi-cuadrado, mutual information) para obtener las relaciones de independencia condicional entre las variables y utilizar esta información como restricciones a utilizar en la construcción del grafo.

En la etapa de determinación de los parámetros probabilísticos del grafo se generan las matrices de probabilidad condicional (una matriz por variable, donde se especifica la probabilidad condicional de esa variable condicionada a cada uno de sus padres). Una vez fijada la estructura del grafo, se utiliza un procedimiento para estimar el valor de los parámetros a partir del conjunto de datos de entrenamiento utilizando el enfoque de máxima probabilidad (suele aplicarse el algoritmo expectation–maximization EM). La probabilidad conjunta puede ser reconstruida simplemente multiplicando estas tablas.

Finalmente utilizando la red bayesiana y un conjunto de atributos X1,X2,…,Xn el algoritmo retorna el valor del atributo de clase c que maximiza la probabilidad a posteriori P(c|X1,X2,…,Xn).

Una de las ventajas más interesantes de las redes de bayes es que hace un buen uso de la información conocida a priori (incluida en el conjunto de entrenamiento) gracias al conocimiento adquirido sobre la relación estructural entre los atributos, aunque su principal desventaja es que la complejidad aumenta exponencialmente de acuerdo a la cantidad de ellos. Una simplificación del método general descrito son las redes Naive Bayes que incorporan restricciones en la estructura del grafo (es un DAG con un padre, que es el atributo de clase) y asumen que todos los hijos son independientes con respecto

a su padre, con esto baja la complejidad en beneficio de una mejora grande en los tiempos de entrenamiento.

Algoritmos de esta familia:

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)

Clustering Esta categoría reúne a un conjunto de métodos que realizan segmentación, esto es separan las observaciones en segmentos de forma tal que se maximizan tanto la homogeneidad interna y como la heterogeneidad entre los segmentos. En otras palabras, siendo un cluster un conjunto de objetos, la tarea de clustering implica segmentarlos de forma que se maximicen la medida de similitud inter-cluster y la medida de disimilitud intra-cluster. En las afirmaciones vertidas anteriormente está implícita la necesidad de definir la noción de similitud entre observaciones, para esto se define una función que puede ser la distancia euclidiana, alguna variación de la misma o cualquier otra que pueda dar un valor que indique el grado de similitud entre dos observaciones (con ciertos requerimientos como simetría y en algunos casos desigualdad triangular).

Los métodos de clustering son no supervisados, descriptivos, ya que buscan generar una cantidad finita de categorías que permitan describir los datos. Se pueden dividir en dos subcategorías: métodos jerárquicos y no jerárquicos. Baesens (2014). En los métodos jerárquicos, tenemos el clustering jerárquico divisivo donde el procedimiento comienza desde un gran cluster que contiene a todas las observaciones y, en cada etapa, se divide los clusters disponibles en dos de acuerdo al criterio definido por la función de similitud, el proceso se detiene cuando todos los clusters son atómicos. Los métodos jerárquicos por aglomeración funcionan forma inversa, o sea se parte de todas las observaciones como clusters atómicos y en cada etapa se combinan de a pares hasta llegar a un único cluster que contiene todas las observaciones. En ambos métodos, es necesario definir un criterio que permita decidir cuando detener las iteraciones, el mismo puede estar definido en base a una cantidad deseada de clusters o una distancia intra-clusters que se desea que se supere para todos los clusters.

Entre los métodos no jerárquicos, un enfoque común es definir una función de costo sobre un conjunto parametrizado de clusters siendo el objetivo del algoritmo de cluster encontrar un conjunto de clusters que minimicen la función de costo, tenemos entonces un problema de optimización. Hay que tener en cuenta que este tipo de problemas de optimización son NP-complejo, incluso llegar a una aproximación lo es. Se asimila con este tipo de métodos el algoritmo K-means. En K-means cada cluster tiene un centroide (que no necesariamente es un punto del conjunto de datos), el problema aquí es elegir los k centroides y determinar cuales serán los objetos que pertenecen a cada cluster de

forma que se minimicen la distancias de los objetos al centroide de sus clusters, dado que este problema es NP-Complejo, este algoritmo busca un óptimo local y se suele ejecutar múltiples veces con distintos valores aleatorios de inicialización de forma de seleccionar el mejor resultado.

Algoritmos de esta familia:

- Hierarchical Clustering
- k-Means
- k-Medians
- Expectation Maximisation (EM)

Reglas de Asociación El aprendizaje basado en reglas de asociación forma parte de los métodos de análisis descriptivo, Aquí se busca describir patrones de comportamiento, al contrario del análisis predictivo donde se intenta predecir el valor de cierto atributo de instancia. Estos métodos trabajan con un conjunto transacciones D, cada transacción contiene un conjunto de items (i1, i2, ...,in) elegidos del conjunto de items I (todos valores categóricos). Una regla de asociación es una implicación del tipo X=>Y (donde X e Y están contenidos en I y la intersección entre X e Y es vacía). En este contexto se llama a X "antecedente" de la regla e Y consecuente de la regla. Las reglas de asociación son estocásticas dado que están acompañadas por una medida estadística del grado de asociación.

Básicamente los algoritmos de esta familia trabajan con dos medidas: soporte y confianza. El soporte es la fracción del total de transacciones que contienen a X e Y, por otro lado la confianza es la fracción de items que contienen X e Y en el total de transacciones que contienen X (en otras palabras es la probabilidad condicional de X dado Y). El procedimiento básico consiste en establecer valores mínimos para estas dos medidas (minsup y minconf), identificar los conjuntos cuya confianza supere minsup (a los items de este conjunto se los denomina items frecuentes) y eliminar los que su confianza no supere minconf.

Para realizar el procedimiento descrito se proponen distintos algoritmos que permiten mejorar el costo computacional de recorrer varias veces el conjunto de transacciones completo, como el algoritmo APriori. Este, se basa en que los sub conjuntos de un conjunto frecuente son también frecuentes, y que los super conjuntos de un conjunto infrecuente (aquellos conjuntos que no superan minsup) son infrecuentes. Con estas premisas, se utiliza un método botton-up, tomando los conjuntos frecuentes más pequeños y se los une entre sí, son descartados los infrecuentes y el procedimiento avanza así hasta que no se pueda extender más ninguno de los conjuntos frecuentes obtenidos. Este procedimiento es más eficiente que el de fuerza bruta, igualmente existen procedimientos más eficientes aún.

Algoritmos de esta familia:

- Apriori algorithm

- Eclat algorithm

Redes Neuronales Artificiales Las Redes Neuronales Artificiales (NNA por las siglas de Artificial Neural Networks) están basadas en la idea de perceptrón. Los perceptrones se aplican en clasificadores binarios para aprendizaje supervisado, los mismos pueden decidir si los atributos libres de una instancia pertenecen o no a una clase, o sea predecir el valor del atributo de clase de forma binaria. El perceptrón se basa en una función que mapea los atributos de entrada (X1..Xn) con un peso asociado (W1..Wn) y lo compara con un valor umbral, siendo el resultado de esta comparación el valor del atributo de clase. Entrenar el clasificador binario es encontrar el conjunto de valores para los pesos W1 a Wn (conocido como vector de predicción), para esto se asigna un valor inicial a dicho vector y se aplica el algoritmo a cada instancia del conjunto de entrenamiento corrigiendo los valores de W hasta encontrar un vector que satisfaga todo el conjunto.

Los algoritmos de clasificación binaria tipo perceptrón son efectivos para clasificar conjuntos de datos cuyas instancias sean linealmente separables (un conjunto de puntos del espacio euclídeo es linealmente separable si existe una línea tal que separe los puntos pertenecientes a una clase de los pertenecientes a la otra, la idea puede generalizarse a espacios euclidianos de mayor dimensión). Este tipo de algoritmos dejan de ser efectivos en conjuntos no linealmente separables dado que no es posible encontrar un vector de predicción que satisfaga la condición nombrada anteriormente, para resolver este tipo de problemas se utilizan Redes Neuronales Artificiales. Las ANN se componen de una cierta cantidad de neuronas artificiales conectadas entre si formando capas o niveles. La organización del clasificador tipo perceptrón comentada anteriormente podría verse como una ANN de una sola capa, siendo un caso particular de arquitectura de una ANN. Las capas se nombran como capa de entrada, capa de salida y las que están entre ellas se denominan capas ocultas. Las ANN más comunes son las feed-forward, que permiten el paso de la información en un solo sentido desde la entrada a la salida. El proceso de entrenamiento es similar al descrito para los perceptrones, una vez que se cuenta con los pesos, se los fija como constantes y se puede utilizar el modelo para predecir el atributo de clase. El funcionamiento de una red depende de tres factores: la función de activación, la arquitectura de la red y los pesos asociados a las conexiones; con los dos primeros parámetros fijos (ya que fueron seleccionados en tiempo de diseño) el comportamiento de la red queda, en gran manera, dirigido por los pesos asociados a las conexiones.

En el proceso de entrenamiento de las redes feed-forward suele utilizarse el algoritmo back-propagation o alguna de sus variantes. El inconveniente aquí es que suele ser demasiado lento para ciertas aplicaciones. Existen variadas propuestas para mejorar la velocidad de entrenamiento, en general se basan en la idea de utilizar algún mecanismo para calcular el valor iniciar de dichos parámetros, en lugar de iniciar el ciclo de entrenamiento con valores de pesos aleatorios, obteniendo de esta forma una mayor velocidad de convergencia hacia los optimales,

Algoritmos de esta familia:

- Perceptron
- Back-Propagation
- Hopfield Network
- Radial Basis Function Network (RBFN)

Aprendizaje Profundo En esta categoría están incluidas una amplia variedad de algoritmos basados en la idea general de red neuronal artificial (descrita en el punto anterior) con la particularidad de que aquí se utilizan varias capas de procesamiento no lineal de naturaleza jerárquica (hablamos aquí de una cantidad de las capas ocultas que se describieron para ANN) donde las características de alto nivel se definen mediante las de menos nivel a medida que se recorre las capas. Este tipo de redes han tomado gran auge en las últimas dos décadas hasta separarse de las ANN y ser tomadas como una familia específica. Son aplicadas en modelos tanto de aprendizaje supervisado como no supervisado, en una gran variedad de metodologías y arquitecturas.

De acuerdo al propósito de aplicación y/o métodos a ser utilizados, se pueden categorizar las técnicas de deep learning en tres grupos Deng (2014):

. Redes profundas para aprendizaje no supervisado o generativo: se destinan a capturar las correlaciones de alto nivel en los datos observados con propósitos de clasificación o generación cuando no existen datos clasificados de entrenamiento. Cuando se utiliza en modo generativo puede ser utilizada para caracterizar la distribución de probabilidad condicional de los datos visibles y sus clases de forma tal que pasen a formar parte de la información visualizada, de esta forma mediante la regla de Bayes puede funcionar como un modelo discriminativo.

. Redes profundas para aprendizaje supervisado: están destinadas a realizar discriminación con el propósito de clasificar. Cuentan, en forma directa o indirecta, con datos con información de clase en el momento del entrenamiento del modelo. Pueden ser nombradas en cierta literatura como redes profundas discriminativas.

. Redes profundas híbridas: son redes profundas discriminativas que incorporan la información provista por una red profunda no supervisada o generativa. También, se incluye en esta categoría, las redes cuyo criterio discriminativo para el aprendizaje supervisado es utilizado para estimar los parámetros de una red generativa.

Algoritmos de esta familia:

- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders

Reducción de Dimensionalidad Los datos del mundo real (imagen, sonido, señales en general) suelen pertenecer a espacios de alta dimensionalidad presentando esto un desafío adicional a la hora de realizar una extracción de conocimiento en forma eficiente y trae aparejada una capacidad de generalización pobre, es aquí donde los algoritmos de reducción de dimensionalidad acercan una alternativa interesante. La reducción de dimensionalidad es una transformación de datos de alta dimensión en una representación de ellos de menor dimensión sin perder sus propiedades representativas (idealmente la dimensionalidad intrínseca de los datos) Tr (2009). La dimensionalidad intrínseca es la mínima cantidad de parámetros necesarios para representar las propiedades observadas de un conjunto de datos. La reducción de dimensionalidad habilita el procesamiento de los datos en algunos dominios donde, sin ella, sería impracticable en cuanto a requerimientos de capacidad computacional y utilización de memoria.

El método análisis de componente principal (PCA) es el más difundido de los miembros de esta familia. En él tanto las tareas de compresión y recuperación se realizan mediante una transformación lineal. El método encuentra la transformación lineal con la cual se minimiza la diferencia entre cada vector recuperado y su vector original según sus cuadrados mínimos. Si x_1,...,x_m son m vectores de R^d se reducirá la dimensionalidad de este vector a R^n mediante la matriz $W \in R^{n,d}$ usando una transformación lineal x->Wx. A su vez, una segunda matriz $U \in R^{d,n}$ se utilizará para recuperar a cada vector x desde su forma comprimida. La versión recuperada de x, a la que llamaremos $\tilde{x}$ pertenece al espacio dimensional R^d. Los valores de las matrices W y U se calculan de forma tal que minimicen la suma de las distancias cuadráticas totales entre los vectores originales y sus versiones recuperadas. Puede demostrarse que, siendo A=$\sum_{i=1}^{m} x_i x_i^T$ y u_1,...,u_n los n eigenvalores de A, la solución a este problema de optimización es la matriz U cuyas columnas son u_1,...,u_n y la matriz $W = U^T$

Tradicionalmente la reducción de dimensionalidad se llevó a cabo mediante métodos lineales como el análisis de componente principal (PCA) aunque, con el tiempo, varios modelos no lineales han sido incorporados a la práctica. En contraste con los métodos lineales, las técnicas no lineales obtienen mejores resultados en campos donde los datos son complejos y no lineales. Entre los algoritmos no lineales más conocidos, se pueden nombrar Sammon Mapping, Isomap y Laplacian Eigenmaps entre otros.

Algoritmos de esta familia:

- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)

- Flexible Discriminant Analysis (FDA)

Algoritmos de ensamble Bajo la familia de Algoritmos de Ensamble, lo que encontramos, en realidad, son meta algoritmos que combinan distintos algoritmos de aprendizaje automático en un único modelo predictivo con el fin de mejorar los resultados que se obtendrían aplicando los algoritmos de forma individual. Algunos ensambles utilizan algoritmos de base del mismo tipo, generando un ensamble homogéneo, aunque se obtienen resultados interesantes combinando diferentes tipos de algoritmo base.

Según la configuración en que se aplican los algoritmos de base utilizados, esto es, como son combinados los resultados de cada paso, se pueden observar dos patrones de aplicación diferentes:

. Ensamble secuencial: la motivación principal de aplicar los algoritmos base en forma secuencial es explotar la dependencia entre ellos. De esta forma, el desempeño general puede incrementarse aumentando el peso de las etiquetas aplicadas a las observaciones que no han sido correctamente etiquetadas por el algoritmo base aplicado anteriormente. Por ejemplo: Boosting, este algoritmo se caracteriza por ajustar una serie de modelos débiles (aquellos que apenas son capaces de superar la clasificación al azar) asignando un peso a los resultados y ajustar este peso en los pasos sucesivos (aumentando los pesos de las observaciones mal clasificadas en pasadas anteriores).

. Ensamble paralelo: en esta configuración, se aplican en paralelo más de un algoritmo base, de forma de explotar la independencia entre ellos, y se promedian los resultados de forma de reducir el error general. Por ejemplo: Bootstrapped Aggregation, donde se espera reducir la varianza promediando varias estimaciones, esto es entrenar M modelos con M sub conjuntos de entrenamiento diferentes y luego utilizar un procedimiento de votación (para clasificación) o promedio (para regresión).

Algoritmos de esta familia:

- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (blending)
- Gradient Boosting Machines (GBM)
- Gradient Boosted Regression Trees (GBRT)
- Random Forest

4.4 Agroinformática

4.4.1 Características principales Agroinformática

Son escasas las actividades que han quedado fuera de los avances registrados en ciencias de la información en las últimas de décadas, y la agricultura no escapa a esta realidad. En este marco, entendemos por agroinformática a la aplicación de las ciencias de la información a las actividades agrícolas.

La agricultura sostenible requiere una rápida y precisa adaptación a las cambiantes condiciones naturales y económicas que la rodean. En el último siglo se registraron cambios fundamentales en las actividades agrícolas, pasando el conocimiento de estar en el dominio exclusivo del especialista que conoce el campo a sofisticados sistemas de monitoreo y toma decisiones asistidos por computadora. El cambio de escala, en la actividad también es notable, pasando de cultivos de tamaño pequeño a grandes cultivos, muchas veces imposibles de ser recorridos por el especialista donde la aplicación de insumos en forma homogénea sería totalmente ineficiente.

La utilización de modernas maquinarias con sensores geolocalizados, en muchos casos sistemas autónomos tipo robot que registran datos automáticamente, hacen que los volúmenes generados sean muy grandes y requieran sistemas avanzados de tratamiento para su explotación, generando conocimiento e información accionable, en una suerte de sistema realimentado de lazo cerrado.

4.4.2 Agricultura de precisión:

La evolución de la agricultura en los últimos dos siglos ha sido muy importante, primero, en forma de técnicas de mecanización durante la revolución industrial en los países de Europa Occidental (en los siglos 18-19), y luego con la aplicación de los procedimientos de producción industrial en Estados Unidos y otros países desarrollados (entre los años 1930-1950). Los grandes cambios en las herramientas técnicas, que dejó este proceso, no tuvieron su contraparte en los métodos de operación. Tradicionalmente la unidad básica de operación era el campo, en la actualidad, aunque con herramientas de mayor escala, sigue siendo el campo en forma uniforme. En la delimitación del campo se tienen en cuenta dos factores: que la uniformidad y homogeneidad de condiciones ecológicas sea la mayor posible (esto suele implicar tamaños de campos reducidos) y que la operación de la maquinaria sea lo más eficiente posible (esto requiere campos lo más extensos posibles). Aún donde estos factores tienen un balance ideal, existen, dentro del campo, variaciones (topológicas, de composición del suelo, el riego, la temperatura, humedad, etc) que hacen que, incluso cuando la solución agrotécnica esté perfectamente implementada, solo será óptima en algunos sectores del campo, a la vez que en los restantes será subóptimo y en otros supraóptimo. Esta situación lleva a una disminución en la cosecha y/o en la calidad, mayor consumo de energía e insumos (fertilizantes, abonos, plaguicidas, semillas) que

disminuyen las ganancias, incrementan el estrés ambiental y la polución, entre otros efectos negativos.

El movimiento hacia un proceso de agricultura de precisión permite dejar de lado el tratamiento uniforme del campo mediante la aplicación de valores cercanos al óptimo en cada sección del campo. La agricultura de precisión, más que la aplicación de nuevas tecnologías de la información a la agricultura, significa un cambio de paradigma hacía un proceso agrícola dinámico basado en información y conocimiento Pepó (2013).

Para el productor, existe una relación económica entre los resultados obtenidos y la aplicación de técnicas de agricultura de precisión bastante evidente. Desde este punto de vista la adopción, este tipo de sistemas se considera parte de los costos fijos, con el potencial de disminuir los costos variables y aumentar los beneficios económicos.

En esa línea, E. Mantovani; F. de A. de Carvalho Pinto y D. Marçal de Queiroz definen Agricultura de precisión Bongiovanni (2006) como "un conjunto de técnicas orientado a optimizar el uso de los insumos agrícolas (semillas, agroquímicos y correctivos) en función de la cuantificación de la variabilidad espacial y temporal de la producción agrícola. Esta optimización se logra con la distribución de la cantidad correcta de esos insumos, dependiendo del potencial y de la necesidad de cada punto de las áreas de manejo"

Variabilidad espacial y temporal Es importante resaltar, en la definición anterior, dos conceptos: variabilidad espacial y variabilidad temporal. La primera se refiere a la variación de los parámetros agrotécnica evaluados en distintos puntos geográficos dentro de un mismo periodo de tiempo. Mientras que la segunda refleja la variación de dichos parámetros en distintos momentos para el mismo punto geográfico. La medición de estas variaciones no es una novedad en el mundo del campo, si son novedosas las posibilidades que ofrece la informática en cuanto a su explotación.

Objetivos Entre los objetivos a los que la agricultura de precisión intenta contribuir podemos encontrar:

- Reducir la utilización de insumos y el consumo de energía.
- Mejorar el rendimiento (dentro de las condiciones ecológicas y agrotécnicas dadas)
- Mejorar la calidad del cultivo.
- Mejorar la rentabilidad.
- Disminuir los desperdicios.

En el campo, la aplicación de métodos de agricultura de precisión, permite tener el control sobre parámetros tales como:

- Mecanización del suelo (superficie, profundidad)

- Administración de nutrientes (nutrientes orgánicos, fertilizantes)
- Siembra (densidad de siembra, profundidad).
- Protección de plantas (control de malezas, herbicida, tratamientos con fungicidas e insecticidas)
- Riego (áreas, cantidades)
- Cosecha (área, fecha de cosecha)
- Manejo de residuos de la cosecha

Tipos de procesamiento El modo de procesamiento de un sistema de agricultura de precisión puede ser:

- En línea o tiempo real: Se obtienen ciertos datos del terreno (usualmente) mediante sensores (humedad del suelo, temperatura, daños causados por pestes, enfermedades, etc.) y, en el mismo momento (en el campo), se toman en consecuencia las acciones agrotécnicas (altura de siembra, provisión de nitrógeno, insecticidas, etc.). El ciclo de operación está compuesto de las siguientes fases (según Pepó (2013)): detección, procesamiento de la información, control, implementación de la acción agrotécnica. En la operación en línea el concepto de variabilidad espacial cobra especial importancia.
- Fuera de línea: las acciones agrotécnicas se basan en información (usualmente almacenada en una base de datos) recolectada con anterioridad a la operatoria. La base de datos suele contener estudios realizados al suelo, datos climáticos, resultados en campañas anteriores, etc. Estos datos, luego de ser procesados, generan conocimiento que permite tomar acciones posteriormente. El ciclo de operación está compuesto de las siguientes fases (según Pepó (2013)): recolección de datos, procesamiento de la información, toma de decisiones, implementación de la acción agrotécnica. Aquí puede observarse el concepto de variabilidad temporal definido con anterioridad.
- Híbrido entre los dos anteriores. Utiliza una combinación de fuentes generadas fuera de línea con datos de terreno del instante en que se aplica el control al sistema. Explota las variabilidades temporales y espaciales simultáneamente.

Tecnologías habilitadoras Las tecnologías abajo enumeradas, se encuentran dentro de los factores habilitadores de esta práctica Bongiovanni (2006). No es, posible verlas de manera independiente, sino como un sistema que permiten implementar Agricultura de Precisión, estas son:

- Sistema de Posicionamiento Global GPS: Dada la importancia de la gestión espacial, en agricultura de precisión, se puede observar el GPS como una de las tecnologías facilitadoras más importantes en este disciplina. La tecnología GPS (acrónimo del inglés, Global Positioning Satellite System) nace como una tecnología militar y, desde su liberación para uso civil, es utilizada en una diversas aplicaciones, entre ellas agricultura. Esta tecnología permite calcular la posición (latitud, longitud y altura) del receptor GPS de acuerdo a la intensidad de recepción de señales radioeléctricas emitidas por un sistema de satélites geoestacionarios. En este sentido, el GPS, permite conocer la localización precisa (un equipo GPS trabaja con un error de 5 a 20 metros, aunque mediante el adicionado de un sistema de compensación DGPS el error baja a unos pocos centímetros) y constituye la base de la aplicación de acciones en el terreno. Algunos GPS incorporan un barómetro que incrementa la precisión de la altura.
- Sistemas de Información Geográfica (SIG): son sistemas de información que permiten gestionar datos procedentes del mundo real que están vinculados a una referencia espacial, permitiendo incorporar aspectos del tipo sociales, culturales, económicos o ambientales, facilitando los procesos de toma de decisión. Las herramientas de software SIG permiten combinar información espacial con información alfanumérica en un modelo del tipo base de datos, permitiendo realizar análisis espaciales o territoriales resolviendo consultas que combinan criterios alfanuméricos y espaciales.
- Percepción remota: una definición abarcativa de percepción remota MARTELLOTTO, MENDEZ, VON MARTINI, & BIANCHINI (2007) es “un grupo de técnicas para recolectar información sobre un objeto o área si tener que estar en contacto físico con el objeto o área”. Para esto se utilizan sensores (radiómetros) capaces de registrar variaciones electromagnéticas en alguna de las bandas del espectro electromagnético (la banda utilizada delimita el área de aplicación de la medición). De acuerdo a la distancia a la que se encuentre el sensor del objeto en estudio, en aplicaciones del agro, se utilizan sensores terrestres (usualmente instaladas en maquinaria agrícola), aerotransportados (aeronaves, tripuladas o no) u orbitales (satélites). Los sistemas de percepción registran la energía electromagnética reflejada por los objetos, este hecho permite dos modos de operación: sensores activos (generan una señal que al ser reflejada por el objeto es registrada por el sensor) o sensores pasivos (registran las ondas electromagnéticas generadas por fuentes externas, en general la radiación solar).
- Tecnologías de dosis variable: La Tecnología de Dosis Variables (VRT) es la tecnología que habilita la aplicación de dosis de insumos de acuerdo a las necesidades específicas del punto de aplicación, de esta forma se implementa la principal diferencia con la agricultura tradicional donde se aplican dosis homogéneas (en general el promedio) a todo el terreno. Existe un actuador que controla la cantidad de insumos a aplicar de acuerdo, o bien, a la posición del punto de aplicación (siguiendo una decisión tomada fuera de línea) o de acuerdo a

una decisión tomada en base a sensores en el momento mismo de la aplicación (en línea). La actuación sobre artefactos como pulverizadores de líquidos, dosificadores de fertilizantes granulados, esparcidores de aire, sembradoras y arados; permiten modificar según las necesidades los patrones de siembra, riego, fertilización, aplicación de pesticidas y herbicidas, entre otros.

- Análisis de datos georeferenciados: la aplicación de técnicas de extracción del conocimiento a datos de múltiples variables recogidos en el campo y georeferenciados permite, entre otras cosas, obtener información que permita tomar decisiones óptimas para cada punto geográfico específico. Luego, mediante las tecnologías de aplicación variable y, con la ayuda de los sistemas de posicionamiento global, la información obtenida es aplicada como una acción concreta en el terreno. La incorporación de información espacial a las variables de entrada del proceso de extracción de conocimiento, es un desafío adicional ya que los algoritmos utilizados tradicionalmente carecen de posibilidades de contemplar la autocorrelación espacial entre datos de sitios vecinos.

Zonas de manejo Según Doerge se define zona de manejo Doerge (1999) como una subregión del campo que muestra una combinación funcionalmente homogénea de factores limitantes del rendimiento para la cual es apropiada una dosis específica de insumos o un tratamiento particular. El concepto de zona de manejo es útil en la aplicación de tecnologías de dosificación variable VRT (densidad de siembra, fertilizantes, pesticidas), soporte de la decisión de tipos de cultivo o aplicación de tratamientos para mitigar problemas particulares (malezas, enfermedades o infestaciones de insectos). Para la delimitación de las zonas de manejo se realiza un análisis de la variabilidad espacial y temporal (ambos conceptos definidos en párrafos anteriores) del campo. Existen diferentes estrategias para medir la variabilidad, las cuales serán analizadas en más profundidad en este trabajo, pero para el presente objetivo podemos agruparlas en dos categorías D. E. Clay et al. (2017b): muestreo directo (ejecución de mediciones en puntos específicos del campo definidos a priori, conocido como metodología de grilla) y mapas o mediciones complementarias (utilización de información espacial disponible para el campo como mapas de rendimiento, mapas de conductividad eléctrica del suelo, elevación y tipo de drenaje, mapas de sensores remotos, entre otros). En cuanto al procedimiento a utilizar, existen varios software especializados y enfoques estadísticos que pueden aplicarse. La información y el procedimiento utilizado para delimitar las zonas de manejo depende de problema particular a atacar, incluso pueden definirse en un mismo campo diferentes zonas de manejo con fines distintos.

4.4.3 Monitoreo de rendimiento y Mapas

Dotar a las máquinas agrarias de capacidades de cuantificación de su actividad y ubicación geográfica de la misma, permite generar datos invaluables para distintas

actividades tales como agricultura de precisión. De esta forma, las mediciones de cantidad cosechada con en un instante junto con la ubicación de la medición, permiten, por ejemplo estimar la cantidad de nutrientes removidos por la cosecha, analizar la eficacia de tratamientos de fertilización o simplemente, conocer el rendimiento de un área específica del campo. Este procedimiento permite establecer la variabilidad temporal y espacial del campo, constituyéndose esto en un punto fundamental a la hora de optimizar la utilización de insumos agrícolas en cada punto del campo. Es una herramienta fundamental para la confección de mapas de rendimiento.

Tradicionalmente, el control que permitía conocer el rendimiento del campo era la medición de volumen o peso en el momento de venta del producto. Con el paso del tiempo, y a medida que los avances tecnológicos se acercan al campo, esto fue siendo desplazado por la utilización de sistemas de monitoreo que estiman el flujo del producto a través de la cosechadora e integran esta información con la posición geográfica de la medición. Fulton et al. (2018)

Existen en el mercado una amplia variedad de equipos de monitoreo de rendimiento, adecuados para los distintos tipos de cultivo (se diferencian en la técnica utilizada para la estimación del peso de granos, volumen de algodón, o caña de azúcar, por dar algunos ejemplos). Además, es bueno aclarar, que estos equipos han mejorado su precisión con el paso de los años, y aún hoy, se continúan produciendo avances en este aspecto. La aplicación de tecnologías de percepción remota mediante fenotipado de alto rendimiento Araus & Cairns (2014) permiten tener una visión previa al cultivo muy interesante en algunas aplicaciones.

Dado que la medida de rendimiento es función de la masa y la superficie cosechada, no solo se tiene en cuenta el flujo del producto a través de la cosechadora, también debe tenerse en cuenta el ancho del cabezal de la máquina y su velocidad, que junto con la ubicación, permiten registrar el rendimiento por unidad de superficie. Es aquí donde la precisión y calibración de varios instrumentos pasa a ser importante (el sensor de flujo de masa, el sistema de posicionamiento - usualmente DGPS/GNSS - y el sensor de las medidas del cabezal de cosecha y posición de las mismas).

Sensores de flujo de masa Los sensores de flujo de masa utilizados dependen del tipo de cosecha a medir, usualmente, pertenecen a dos grandes grupos: aquéllos que cuantifican la fuerza de impacto del producto a medida que atraviesan el sensor y aquéllos que cuantifican cambios de volumen del mismo Fulton et al. (2018). La mayoría de los esquemas de medición de humedad comerciales se ajustan a una de las siguientes categorías:

- Medición de peso mediante placa de impacto: consta de elevador de los granos que los propulsa contra un sensor de impacto colocados antes del sinfín de carga, de esta forma la fuerza del impacto contra la placa es función de la velocidad (que se estima en base a la velocidad del elevador y la geometría del mismo) y la masa del producto. Dado que la estimación varía de acuerdo al tipo de grano y

sus características al momento de ser cosechado, este sistema debe ser calibrado cuidadosamente. En general la calibración se realiza al comenzar la jornada de cosecha realizando algunas iteraciones de medición con cantidades conocidas de grano (o bien, en ensayos donde se cosecha y luego se mide el peso con un sistema externo, ingresando la medición en la consola de medición de forma que pueda ajustar los parámetros necesarios). Es importante que cada una de las iteraciones de calibración sean representativas de las diferentes variaciones con las que el monitor se va a encontrar luego en el campo.

- Medición de volumen mediante sensor fotoeléctrico: consta de un emisor de un haz de luz y un receptor fotosensible, el volumen se estima midiendo la señal proveniente del receptor que varía de acuerdo al flujo de la cosecha. Existen, básicamente dos tipos de sensores fotoeléctricos los que miden la luz reflejada en los objetos y los que miden la obstrucción del haz de luz (el emisor y receptor se encuentran enfrentados, en este caso). El sensor se ubica en el conducto que conduce el material recolectado. Para el caso de las cosechas de granos el sensor se ubica en el elevador de grano recolectado de forma de detectar el paso de cada paleta que los contiene, en productos como el algodón, se ubica directamente en el conducto de transporte que lleva al algodón hasta la canasta de la cosechadora.
- Medición por microondas y radiación: La estimación de volumen por microondas y radiación trabajan de forma similar a la medición por sensor fotoeléctrico, en el caso de las microondas se mide la señal reflejada por el material en el conducto de transporte, en la medición por radiación (usualmente rayos gamma) el receptor y el emisor están enfrentados, y el material a medir circula entre ambos. La medición por microondas es ampliamente utilizada en máquinas comerciales. En cuanto a la medición por radiación tiene la ventaja de no verse afectada por la humedad de la cosecha, aunque debe tenerse que dados los efectos adversos de la radiación debe ser empleada con todos los recaudos del caso, y de hecho en algunos países no está permitida su utilización.
- Medición por flujo volumétrico: funciona midiendo la velocidad de rotación de una rueda con paletas, el material recolectado llena el espacio entre las paletas y la rueda avanza una posición, dado que el volumen que puede contener el espacio entre dos paletas es conocido, así como la cantidad de paletas de la rueda, la rotación de la misma permite estimar el volumen de material que circula por el medidor.

Dado que el peso y/o el volumen de la mayoría de las cosechas varía producto de la perdida de humedad, es este un factor importante a considerar al momento de la estimación del rendimiento. Existen diferentes procedimientos que permiten medir el grado de humedad de la cosecha al momento de realizar la misma, de forma de poder introducir esta variable en las estimaciones generadas por el monitoreo de rendimiento. En general se utilizan sensores capacitivos que miden el coeficiente dieléctrico de los granos (dado que la humedad aumenta la conductividad de los granos, el coeficiente

dieléctrico es un buen indicador de la humedad que almacenan los mismos) a medida que estos atraviesan las placas del sensor (ya sea en el sinfín de grano limpio o en el elevador de granos). Una alternativa muy utilizada es la toma manual de muestras mediante medidores portátiles. Además, es muy común que las máquinas registren otros parámetros durante la cosecha, como ser: posición del cabezal de la cosechadora (levantado/bajo), temperatura externa, velocidad del elevador de granos, velocidad del ventilador del conducto de transporte (en cosechas de algodón, por ejemplo).

Recolección y limpieza de datos de rendimiento El monitor de rendimiento es capaz de suministrar información valiosa directamente durante la ejecución de la cosecha en las pantallas instaladas en las maquinarias agrarias a tal fin (rendimiento actual, rendimiento promedio, superficie cosechada, humedad del material recolectado, entre otras). Si bien, las posibilidades de visualización en vivo son interesantes, existe una gran cantidad de aplicación para los datos generados durante el proceso de cosecha y pese a que existen monitores de rendimiento comerciales desde la década del 90 aún hoy la extracción de conocimiento de los datos generados sigue siendo un área de activo estudio. Durante la ejecución de la cosecha los datos de rendimiento son registrados y luego transferidos (mediante medios removibles, conexiones tipo CAN o conexión inalámbrica) para su procesamiento. Es de interés que los datos estén geográficamente referenciados, para esto es necesario contar con un GPS y en lo posible con corrección DGPS, de este modo se amplia el espectro de aplicación pudiéndose generar mapas de rendimiento.

Si bien muchos fabricantes tiene su formato de datos, existe ciertos lineamientos que siguen la mayoría de ellos. En general se trata de archivos de texto de valores separados por coma (CSV). Cada línea del archivo incluye Shearer, Fulton, McNeill, Higgins, & Mueller (1999): longitud, latitud, flujo de masa, hora del GPS, duración del ciclo, distancia recorrida durante ese ciclo, ancho efectivo del cabezal, humedad, estado del cabezal (bajo, levantado), número de pasada, número de serie del monitor, identificador del campo, identificador de la carga, tipo de cosecha, estado del GPS, punto de pérdida de precisión, altura. Antes de la utilización de los datos generados por el monitoreo de rendimiento, es importante realizar una limpieza sobre los mismos, dado que durante el proceso se registraron momentos que podrían generar sesgo sobre el conocimiento obtenido o bien, no son necesarios directamente. Por ejemplo, los tramos donde no se realizó cosecha (el cabezal estaba alto), los bordes del campo (donde no se puede considerar se aplicó el ancho del cabezal por completo), cambios bruscos en la marcha (existe una diferencia de tiempo entre el instante en que grano es cosechado y el instante en que es medido por el sensor, las velocidades constantes hacen despreciable este delta de tiempo, pero no los cambios bruscos). La limpieza de datos es un proceso de selección más que eliminación, los datos no seleccionados pueden ser almacenados para aplicaciones que puedan explotarlos de forma no perder información (por ejemplo, puede ser útil contar con mediciones de cantidad recolectada en los bordes del campo), aún cuando parezcan anómalos o erróneos.

Aplicación de sistemas de información geográfica Dado que los datos obtenidos están georeferenciados por naturaleza, es interesante la aplicación de un sistema de información geográfica (GIS) para la explotación de los mismos. Teniendo en cuenta que "una capa de datos (data layer) es una representación geográfica de un objeto específico.. los objetos del mundo real pueden ser mapeados a características (features) de una capa de datos.. feature es una representación de un objeto en un mapa" Brase, Shannon, Clay, & Kitchen (2018), podemos obtener, de los datos recolectados, un data layer que represente el rendimiento en un instante de tiempo (suele contarse con otras capas como: limites del campo, características obtenidas por sensores remotos, dosis de fertilizante aplicados, etc). Una capa de datos está conformada por dos componentes: un mapa y una tabla de atributos, en este caso los componentes del mapa son polígonos obtenidos del sistema de posicionamiento utilizado durante la recolección y los valores de la tabla de atributos representan el rendimiento obtenido según la metodología comentada anteriormente. De los formatos de datos utilizados en GIS, el formato vectorial (representación mediante puntos, líneas y polígonos, siendo estos últimos los utilizados en este caso) es el más natural dadas las características de los datos obtenidos (que reflejan tramos recorridos por la cosechadora), aunque también pueden ser obtenidos datos en formato raster (una grilla en la que cada celda representa una determinada superficie del campo de acuerdo a una resolución prefijada) si es más conveniente para el software GIS utilizado.

Análisis de datos de rendimiento El siguiente paso es el análisis, donde los datos recolectados se convierten en información (convirtiéndolos, organizándolos y resumiéndolos) de forma que puedan ser utilizados para toma decisiones. En el análisis, un concepto central a tener en cuenta es la resolución. Existen tres tipos de resolución, según Brase et al. (2018):

- Resolución temática: es el detalle con que la información es categorizada para su utilización (usualmente su representación en un mapa). Los mapas temáticos permiten observar la variabilidad espacial mediante la representación de una característica según su categorización. A mayor cantidad de categorías, mayor resolución temática (existe una relación de compromiso, dado que mucha cantidad de categorías son difíciles de distinguir, y pocas significan una perdida de información importante).
- Resolución espacial: es la cantidad de detalle que tiene la característica representada en una capa de datos. En las capas vectoriales, se traduce en cantidad de vértices contenidos en el mapa, en las capas raster en el tamaño (físico) que tiene una celda de la grilla. Aquí también existe una relación de compromiso entre la cantidad de detalle y la capacidad de almacenar/procesar la información.
- Resolución temporal: es el detalle con que se cuenta para reflejar variabilidad temporal. Se utiliza para mostrar características que varían con el tiempo: se

guardan registros de manera periódica, cada registro estará etiquetado con un detalle del instante en que fue tomado. Existen distintas herramientas provistas por el sistema GIS que permiten analizar la información obtenida de los procesos de recolección y procesamiento descriptos anteriormente. Enumeramos abajo, algunas de las posibilidades mencionadas por Brase et al. (2018) que pertenecen a la intersección con Monitoreo de Rendimiento y nos parecen interesantes:

- Clasificación de capas de datos: Para que sea útil a las capacidades de interpretación humanos la información debe ser categorizada y visualizada. Una herramienta adecuada para estos fines es el mapa temático, para crearlo se seleccionará un mecanismo de clasificación temática y una resolución adecuada a los objetivos planteados. Existen dos aproximaciones a este tema: se puede basar la clasificación en categorías estándar (por ejemplo, tipos de suelo según PH) o un sistema de categorías definido por el usuario, donde el mismo selecciona los límites de cada categoría (por ejemplo, no existen categorías universales para rendimiento, se deben definir de acuerdo a la zona, tipo de cosecha, etc). Para seleccionar la resolución, debe tenerse en cuenta que utilizar poca cantidad de categorías genera una importante pérdida de información, a la vez que es impráctico (en términos de utilidad y/o de recursos) la utilización de una gran cantidad de ellas.
- Estadísticas espaciales: Dado que la información espacial es susceptible a ser analizada estadísticamente, como cualquier otro tipo de información, los sistemas GIS ofrecen la posibilidad de realizar análisis geoestadístico, que resulta de interés en la aplicación que estamos describiendo. Para esto, permiten confeccionar tablas de frecuencia que presentan ordenadamente un grupo de observaciones (según parámetros como frecuencia absoluta y relativa, número de clases, amplitud de clase, marca de clase) y calcular parámetros geoestadísticos como medida de tendencia central (media, mediana, moda), medidas de dispersión (desviación estándar, varianza, coeficiente de variación) y medidas de forma (coeficiente de curtosis, coeficiente de sesgo o asimetría). Algunos software permiten generar una matriz de correlación entre variables, por ejemplo tomando las variables rendimiento y concentración de magnesio podría estudiarse si esta última influye en el rendimiento en la zona bajo estudio.
- Cálculo de campos: Permite aplicar expresiones matemáticas de diversa complejidad a una o más variables, de forma de obtener una nueva variable. Este operador suele encontrarse con el nombre de calculador de campos (field calculator) en las herramientas de software comerciales. Dado que en las distintas capas podemos contar, además de datos de rendimiento, con datos de diversa índole recolectados en el momento de siembra, recolección o escrutinio directo, la capacidad de derivar campos adicionales mediante la aplicación de fórmulas matemáticas a estos ofrece un basto campo de aplicación. Adicionalmente estas herramientas proveen un operador llamado calculador raster (raster calculator) similar que permite derivar capas raster a partir de otras del mismo tipo.

- Normalización o Estandarización: La normalización o estandarización permite hacer comparaciones entre grupos de mediciones mediante la transformación de escala de las variables. Por ejemplo, Es muy usual querer comparar la eficacia de un tratamiento aplicado a dos zonas de administración donde existen cultivos diferentes, en este caso es necesario un proceso matemático que permite comparar la variación de rendimiento en uno y otro cultivo. Existen diversos procesos para alcanzar este fin, algunos de ellos logran su objetivo mediante el ajuste a una variable normal y otros implican alguna transformación de proporción para llevar los datos a un intervalo comparable.
- Conversión a Raster: Se trata del proceso de conversión desde una capa vectorial o bien datos vectoriales geolocalizados, hacia una capa raster (una grilla) con el fin de aplicar operaciones como clasificación, estadísticas, calculo de campos, etc. Aquí debe seleccionarse la resolución de acuerdo a los conceptos descriptos en párrafos anteriores. Esta operación suele incorporar algún procedimiento de interpolación que permite la estimación del valor de las celdas de acuerdo a los valores de puntos cercanos, esto último es muy aplicado en agricultura para generar capas raster desde información de muestreo directo en el campo (por ejemplo, muestras de humedad en el suelo).
- Variación temporal: Los cambios temporales pueden ser evaluados utilizando dos o más capas tipo raster (debidamente estandarizados e interpolados) que representen periodos de tiempo consecutivos. Dado que las celdas se encuentran alineadas, puede ser calculada la tasa de variación mediante la aplicación de un calculador raster que evalúe la diferencia entre ellas. Como resultado se obtiene una nueva capa raster que contiene los valores de variación entre los periodos de tiempo analizados. La variación temporal es muy utilizada en sistemas agro-informáticos, uno de los ejemplo es el calculo de la variación de rendimiento entre cosechas sucesivas.
- Clasificación de imágenes: Los sensores remotos utilizados en agricultura de precisión proveen imágenes que están compuestas por información de sensores capaces de captar la reflectancia en distintas bandas del espectro electromagnético (longitudes de onda del espectro visible y, muchas veces, fuera de él como los infrarrojos). Dado que, cada banda de frecuencia tiene propiedades de reflectancia distintos (y conocidos), representar una combinación de ellas es útil para mostrar características especificas de la superficie terrestre. Por ejemplo, dado que las plantas saludables reflejan mayor porcentaje de infrarrojo que las que están bajo condiciones de estrés, la información de esas longitudes de onda son útiles para evaluar el estrés hídrico, observar el estrés nutricional y la localización temprana de enfermedades o plagas de los cultivos. Con la relación entre las bandas, los software especializados, son capaces de calcular índices, el más conocido es el NDVI (normalized differential vegetable index) que se basa en el hecho de que las plantas que crecen vigorosamente absorben la luz roja y las que tienen un

estado saludable reflejan los infrarrojos, el resultado es un índice que muestra la variabilidad en vigor de los cultivos.

Mapas e interpretación Como paso final en el proceso de agricultura de precisión, podemos hablar de la interpretación de los resultados e implementación de las acciones en consecuencia. Usualmente las decisiones finales son tomadas por expertos en el dominio, o al menos guiadas por ellos, siendo una valiosa entrada al proceso de toma de decisiones los resultantes de los procesos comentados anteriormente. Nombramos a continuación solo algunas herramientas de interpretación:

- Mapas de estabilidad: los mapas de estabilidad reflejan el rendimiento de un área a medida que pasa el tiempo (reflejando de esta manera la variabilidad temporal del área). Se construyen a partir de mapas de rendimiento de ciclos consecutivos de cultivo para el mismo área (pese a que pueden calcularse con la información de dos cosechas consecutivas, el agregado de más cantidad de ellas mejora notablemente la calidad de los mismos), así mismo si el campo contiene diferentes cultivos la información deberá ser normalizada y estandarizada para poder ser integrada. Existen diferentes métodos de calculo, usualmente se itera restando cada punto en los mapas raster de ciclos adyacentes o bien se calcula el desvío estándar para cada punto de los mismos. Aunque pueden ser fácilmente generados mediante cálculo de campos, la mayoría de los paquetes GIS pueden generarlos como una función nativa. Conocer la estabilidad del campo, permite decidir, en cada ciclo: tipo de cultivo, tasa de siembra, aplicación de fertilizantes, pesticidas,
- Mapas promedio: Los mapas promedio permiten identificar el rendimiento histórico de del área. El procedimiento de construcción del mapa es similar al de los mapas de estabilidad, aunque el operador utilizado es el promedio estadístico. En cuanto a su aplicación, es interesante la intersección entre los mapas de estabilidad y promedio, ya que permiten identificar, por ejemplo, zonas con alto rendimiento histórico donde se registró variación de estabilidad en la última cosecha, y actuar en consecuencia.
- Mapas de remoción de nutrientes: Es posible estimar la cantidad de nutrientes como nitrógeno, fósforo, potasio y micronutrientes contenidos en los granos cosechados. Dado que la capa de rendimiento contiene la información de cantidad de granos cosechados, es posible entonces estimar los nutrientes removidos y construir con esto el mapa de remoción de nutrientes. Este mapa es de gran utilidad, por ejemplo, para alimentar la dosificación variable de fertilizantes.
- Análisis estadístico: Existe una multiplicidad de factores que afectan el rendimiento, es por ello difícil tener certeza de cual es la respuesta de este ante cambios en dichos factores. Es común que se intente responder la pregunta: ¿ha mejorado el rendimiento con un determinado cambio aplicado a los parámetros del cultivo? Para responder esta pregunta se pueden conducir experimentos y utilizar modelos matemáticos según Clay describe en D. E. Clay et al. (2017a).

De esta manera, los experimentos deben ser diseñados de una forma apropiada que permita luego comparaciones estadísticas válidas entre las zonas en las que se aplicarán las opciones bajo estudio (existen diferentes estrategias de muestreo, las más comunes son: ejecutar los ensayos en disposición de grilla -equidistantes- o simplemente al azar dentro de una zona en estudio, o bien hacerlo en diferentes zonas de administración). Habiéndose realizado el diseño de experimentos, se propone una hipótesis (por ejemplo la mejora en rendimiento al aplicar cierto tratamiento) y esta hipótesis se somete al proceso de prueba de hipótesis estadístico. La conclusión se obtendrá en relación a la hipótesis nula, la que podrá ser rechazada en favor a la hipótesis alternativa o bien no rechazada (en este caso no se afirma que la hipótesis nula es verdadera, solo que no hay suficiente evidencia en contra de la hipótesis nula que permita afirmar que la hipótesis alternativa es verdadera). Continuando con el ejemplo la prueba dará como resultado si el tratamiento no afecta el rendimiento del cultivo (hipótesis nula) o si se ha modificado el rendimiento de acuerdo al tratamiento (hipótesis alternativa). Durante el experimento se recogerá información y se la analizara, basándose en la media y las varianzas de cada réplica y basándose en un criterio estadístico, de forma de verificar si la hipótesis nula es aceptada o rechazada. En la prueba de hipótesis, existen dos tipos de errores: el primero ocurre cuando se rechaza la hipótesis nula siendo esta verdadera (a este tipo de error se lo denomina falso positivo o tipo 1), y el segundo cuando se acepta la hipótesis nula siendo esta falsa (falso negativo, o error de tipo 2). Antes de ejecutar el test se selecciona un valor que indica la cantidad de error aceptado, se lo denomina valor alfa y representa la probabilidad de rechazar la hipótesis nula cuando esta es verdadera.

- Análisis de utilidad neta: Los mapas de rendimiento proveen la información necesaria para calcular ingresos, pero dado que cuando se utilizan tecnologías de dosificación variable los costos se distribuyen de forma no uniforme en el campo, es en este caso donde se puede calcular (y resulta útil hacerlo) el mapa de utilidad neta que permite visualizar la rentabilidad de la producción. Para calcular este mapa es posible utilizar la calculadora raster y generar una fórmula que tenga en cuenta varias de las capas disponibles ya sea ingresos y costos, por ejemplo costo de las semillas, fertilizantes, rendimiento. Procediendo en la información raster, cada celda de la grilla del mapa de utilidad neta contendrá este dato para el punto representado por dicha celda. De forma similar, puede ser utilizado el procedimiento descrito para calcular el mapa de retorno de inversión.

4.4.4 Percepción Remota

La percepción remota (Remote Sensing en inglés, también traducida al castellano como “Sensado Remoto”) es definida en Schowengerdt (2007) como “la medición de las propiedades de los objetos de la superficie terrestre utilizando datos adquiridos desde aeronaves o satélites .. medir algo a distancia, al contrario de In Situ. Dado que no se

está en contacto con el objeto de interés, se basa en señales propagadas de algún tipo como ópticas, acústicas, o microondas". Aunque existen variadas definiciones, todas tienen en común que se trata de una medición efectuada por un sensor que no tiene contacto físico con el objeto observado. En el área de aplicación del presente trabajo, los sensores suelen ser transportados en aeronaves (tripuladas o no, como drones) y satélites; aunque existen situaciones en que los mismos se encuentran directamente en el campo (sensores de mano, o transportados por máquinas agrarias, vehículos todo-terreno, robots , etc) y dado que estas no encuadran en la categoría de Percepción Remota, por la escasa distancia entre el sensor y el objeto destino, serán tratados en un apartado especifico sobre Percepción Cercana (Proximal Sensing en inglés).

Históricamente se puede decir que la actividad comienza a mitad del siglo 19 con las primeras imágenes tomadas desde globos aerostáticos y otras aeronaves tripuladas y no tripuladas, aunque las primeras aplicaciones relacionadas con el agro tuvieron lugar recién en la primer mitad del siglo 20 con los primeros programas gubernamentales de relevamiento de cultivos llevados a cabo por el Departamento de Agricultura de Estados Unidos (USDA) y relevamiento de utilización del suelo realizado, en el mismo país, por el Servicio de Conservación de Suelo (NRCS). En este primer estadio se utilizaban dispositivos de captura fotográficos transportados por aeronaves tradicionales (fotografía aérea). Es partir de los años 70 cuando se forman las bases de la era moderna de esta actividad, con la puesta en órbita de los primeros satélites artificiales orientados a la observación de recursos terrestres, como la misión LANDSAT del año 1972. Desde ese entonces se desarrollaron distintos proyectos de investigación relativos a procesamiento de las imágenes satelitales y su aplicación a distintos aspectos de las ciencias terrestres, entre las aplicaciones en el agro podemos mencionar los primeros modelos de estimación de cultivos y rendimiento (en las primeras etapas aplicados a cultivos de trigo). En el año 1999 se pone en órbita el primer satélite comercial IKONOS capaz de imágenes de alta resolución. También deben mencionarse la recolección de imágenes con sensores digitales ubicados en aeronaves tripuladas y más recientemente no tripuladas (como drones), que por su cercanía al objeto estudiado (dado que realizan vuelos bajos y específicamente localizados) y por su economía de aplicación, pueden general resultados muy interesantes para el sector privado.

Tipos de plataformas: por plataforma se entiende el dispositivo capaz de transportar el sensor en un desplazamiento que permita escudriñar el objeto de estudio. Al hablar de percepción remota estamos asumiendo cierta distancia entre ambos (sensor y objeto), pudiendo variar desde unos pocos metros (cuando hablamos de vehículos no tripulados, como drones), pasando por aeronaves tripuladas que vuelan desde una decena hasta algunas centenas de metros, hasta algunos kilómetros en el caso de satélites en órbita sobre la superficie terrestre.

Tipos sensor según la salida (imagen/no-imagen): los sensores con imagen de salida generan un conjunto de vectores que representa como varía la entrada a lo largo del campo de visión FoV del tipo (x,y,d), esto representa la distribución espacial de la señal medida en el campo de visión. Suelen estar compuestos por un sensor pixelado,

una línea de pixels o un solo pixel con un sistema mecánico de barrido. Por otro lado, los sensores no-imagen retornan una señal basada en la intensidad de la entrada en todo el campo de visión (la salida del sensor no representa como varía la señal de entrada a lo largo del campo). Cada muestra representa la señal percibida en todo el campo en conjunto.

Sensores activos o pasivos: los sensores pasivos utilizan la energía que existe en la naturaleza, suele medirse la energía de alguna fuente natural (por ejemplo el sol) reflejada por el objeto en estudio, o bien la energía emitida directamente por el objeto (por ejemplo el calor). Por otro lado, los sensores activos generan su propia energía que es emitida sobre el objeto observado y se realiza una medición de la energía reflejada por el mismo (por ejemplo, los radares emiten pulsos radioeléctricos en diferentes direcciones, siendo capturada la señal reflejada por los objetos distribuidos en el campo de medición pudiendo de esta forma estimarse las características físicas de los mismos).

Sensores de energía reflejada o emitida: La mayoría de los sensores, tanto activos como pasivos, contienen elementos sensibles a la energía reflejada por el objeto bajo medición (siendo la fuente de energía un elemento natural en los sensores pasivos, o una fuente controlada en el caso de los activos). Los sensores de energía emitida, se basan en que los objetos terrestres emiten alguna clase de energía que puede ser medida, por ejemplo energía térmica que puede ser captada por sensores infrarrojos calibrados específicamente para tal fin. Las características de las imágenes obtenidas (que dependen totalmente de las características del mecanismo utilizado para obtenerlas) puede ser definida mediante cuatro parámetros utilizados en las descripciones de los sensores, estos parámetros reflejan la calidad del sensor en cuatro "resoluciones" que nombramos abajo:

Resolución espacial: se refiere a la cantidad de detalle que posee la imagen adquirida, esto es la superficie terrestre representada por un pixel de imagen (que es la unidad mínima de la imagen). Se denota con el tamaño de un lado del cuadrado representado por un pixel sobre la superficie terrestre. A menor superficie representada por cada pixel, mayor es la resolución espacial del sensor.

Resolución radiométrica: es la cantidad de niveles con los que el sensor puede representar una medición. Usualmente medida en cantidad de bits, por ejemplo una resolución radiométrica de 16 bits permite representar 65535 niveles en cada medición.

Resolución espectral: cantidad e identificación de las bandas espectrales soportadas por el sensor.

Resolución temporal: se refiere a la frecuencia con la que el sensor puede obtener imágenes del mismo área. Se aplica claramente a las plataformas satelitales donde aplica el concepto de revisita, que es el tiempo que tarda el satélite en recorrer la misma órbita. Dado que los sensores utilizados en percepción se remota detectan y clasifican los objetos de la superficie terrestre basándose en la medición de señales propagadas, y que en la gran mayoría de estos se el tipo de señales con las que se trabaja es radiación electromagnética, definiremos abajo algunos conceptos relativos a ella.

La energía electromagnética es una entidad física que combina campos eléctricos y magnéticos oscilantes, se manifiesta mediante dos concepciones: el ondulatorio y el corpuscular. El aspecto ondulatorio interpreta la radiación como un campo eléctrico y otro magnético oscilando en planos perpendiculares. La periodicidad en el espacio del fenómeno ondulatorio determina la longitud de onda (es la distancia entre dos puntos en los que el campo electromagnético alcanzó la misma amplitud). El aspecto corpuscular, concibe la radiación como un haz de corpúsculos llamados cuantos de radiación o fotones que se desplazan a la velocidad de la luz en una dirección determinada. Estas dos concepciones se consolidan mediante la Relación de Planck, que permite interpretar a una radiación de determinada frecuencia como un flujo de cuantos de determinada energía.

El espectro electromagnético se organiza en bandas de frecuencia (por similitud) de acuerdo a la longitud de onda de la energía electromagnética. Las bandas más bajas corresponden a los rayos X (con longitudes de onda del orden de los 0,0001um) y las más altas corresponden a las ondas de radio (con longitudes de onda de varias decenas de metros). De las bandas que lo componen, las más utilizadas en aplicaciones de percepción remota para agricultura son las listadas en la siguiente tabla:

Nombre de la banda	Longitud de onda	Fuente de radiación
Ultravioleta extremo (UVC)	0,1-0,2μm	solar
Ultravioleta cercano (UVA/UVB)	0,2-0,38μm	solar
Visible (V)	0,4-0,7μm	solar
Infrarrojo cercano (NIR)	0,7-1,1μm	solar
Infrarrojo de onda corta (SWIR)	1,1-1,35μm / 1,4-1,7μm / 2-2,5μm	solar
Infrarrojo de onda media (MWIR)	3-4μm / 4,5-5μm	solar / térmica
Infrarrojo de onda larga (LWIR o TIR)	8-9,5μm / 10-14μm	térmica
Microondas / Radar	1mm - 1m	térmica / artificial

En la superficie del sol los fotones se generan por el desplazamiento de electrones debido a acciones y reacciones atómicas y moleculares. Estos viajan a la velocidad de la luz desde la fuente (el sol, en este ejemplo) al destino (la superficie terrestre). Al chocar, por ejemplo, con el follaje de una planta, algunas longitudes de onda son absorbidas, otras reflejadas y algunas transmitidas por la misma.

La atmósfera terrestre funciona como un potente atenuador de las señales obtenidas por el sensor. Teniendo en cuenta que en sensores satelitales las señales viajan desde la fuente (el sol) hasta la superficie terrestre, donde algunas son reflejadas y viajan hacia el sensor,

la señal atraviesa la atmósfera dos veces. La atenuación descripta no es pareja en todas las bandas, existen determinados rangos de longitud de onda donde la percepción remota es posible, y rangos en los que no. Es importante notar que existen complejos mecanismos de compensación que reducen este efecto, ensanchando las bandas utilizables.

Firma espectral Se puede definir firma espectral como la medida de la reflectancia que ofrece un objeto en función de las longitudes de onda de las radiaciones que lo inciden. En aplicaciones de percepción remota relativas al agro el emisor de radiación más comúnmente utilizado es el Sol y las longitudes de onda utilizadas son las visibles y cercano a visible.

La firma espectral se basa en la idea de que cada tipo de superficie interactúa con la radiación de forma distinta absorbiendo, reflejando o transmitiendo las radiaciones en forma diferente en cada frecuencia (a mayor absorción menor reflectancia). De esta forma, es posible guardar y luego reconocer objetos o características de los mismos por la diferencia entre firmas espectrales, incluso en situaciones donde la resolución espacial del sensor es baja como para identificarlos directamente por su forma.

Hay factores que afectan la adquisición de la firma digital mediante el sensor remoto, es importante tenerlos en cuenta aunque suele ser posible compensarlos (mediante el procesamiento de la salida de los sensores en crudo, muchas veces en la plataforma misma, en las estaciones de tierra, o mediante el software del usuario final) o reducir su efecto, estos son: el ángulo de incidencia de la fuente de radiación (usualmente el sol), relieve sobre el que se encuentra la superficie a medir (pendientes), modificaciones introducidas por la atmósfera (nubes y otros objetos climáticos).

Los sensores multiespectrales e hiperespectrales permiten la utilización de un cierta cantidad de bandas, siendo necesario seleccionar cuales incluir en la firma espectral. Esto se hace teniendo en cuenta: la aplicación de la misma (que objetos y/o características se van a detectar), características del sensor (resoluciones, bandas), tiempo de procesamiento requerido en el registro y en la aplicación, entre otros.

Como vimos anteriormente, según el tipo de salida los sensores se clasifican en imagen y no-imagen. Estos últimos tienen amplia aplicación (por la naturaleza de su funcionamiento) en la generación de firmas espectrales, en esta categoría están los espectroradiómetros que entregan un perfil espectral de la superficie captada (esto es una curva con la intensidad de la señal reflejada por el objeto versus la longitud de onda).

Analizando las firmas espectrales de superficies con vegetación terrestre, desde el longitudes de onda del rango visible hasta el infrarrojo, según Ferguson, Rundquist, Shannon, Clay, & Kitchen (2018) podemos observar los siguientes comportamientos de acuerdo al rango de frecuencia conocidos como propiedades de reflectancia de la vegetación:

- Región visible: el principal factor que modifica la reflectancia en el rango visible son los pigmentos, y en especial la clorofila.
- Región infrarrojo cercano: el principal factor que controla la reflectancia en este rango es la estructura celular (y la estructura del follaje)
- Región del infrarrojo medio: el contenido de agua en las hojas definen, en gran parte, la reflectancia en las longitudes de onda pertenecientes al infrarrojo medio.
- Región del infrarrojo térmico (IRT): exhibe la temperatura del follaje, está relacionada con el estado de la vegetación (se ha mostrado que la temperatura refleja el grado de estrés, por ejemplo en condiciones de falta de humedad en el suelo).

Índice de vegetación Dado que la reflectancia en distintas longitudes de onda varía de acuerdo a las características de la vegetación (propiedades de reflectancia de la vegetación), se elaboró el concepto de índice de vegetación, que es una combinación matemática o transformación de la reflectancia de la superficie terrestre en dos o más longitudes de onda destinada a resaltar alguna característica particular de la vegetación.

Los índices de vegetación permiten interpretar los datos de los sensores de percepción remota reflejando, principalmente, las diferencias y cambios en hojas expuestas del follaje de las plantas. El método de validación más comúnmente utilizado es la correlación directa entre el VI obtenido del sensor remoto y las características de interés obtenidas por observación directa en el campo (utilizando métodos invasivos o externos/sensores próximos, la validación con estos últimos, al ser similares a los utilizados en el sensor remoto, suele ser de utilidad para detectar desviaciones por la interferencia de efectos atmosféricos).

Existe un centenar de índices de vegetación (VI) en la literatura científica, de estos, unos treinta son utilizados desde hace décadas en diferentes aplicaciones del agro. Estos fueron relevados en el estudio Xue & Su (2017) de donde extrajimos los más relevantes que comentaremos a continuación (siguiendo el agrupamiento propuesto en dicho trabajo).

- Normalized Difference Vegetation Index (NDVI): Es el índice de vegetación más utilizado. Se trata de una relación normalizada entre la banda de longitudes de onda correspondiente del color rojo y la del infrarrojo cercano. Caracteriza el crecimiento y vigor del follaje.
- Índices VI básicos: El índice RVI (Ratio Vegetation Index) se basa en la idea de que las plantas absorben en relación más luz rojo que infrarroja. Es unos de los primeros VI, propuesto a fines de los 60, es muy utilizado en estimación y monitoreo de biomasa verde sobretodo en plantaciones densas, dado que en condiciones de baja densidad es sensible a condiciones atmosféricas. El índice DVI (Difference

Vegetation Index) trabaja con las mismas bandas, es sensible al suelo de fondo, y se suele aplicar al monitoreo del ambiente ecológico de la vegetación.

- Índices VI con corrección atmosférica: El índice ARVI (Atmospherically Resistant Vegetation Index) es similar al NDVI, con la diferencia de que incluye una corrección de efectos atmosféricos que se basa en la afirmación de que la mayor concentración de partículas en la atmósfera (debido a algún tipo de aerosol, lluvia, neblina o humo) aumenta la dispersión en la zona visible, y lo hace más en las longitudes de onda del azul que en las del rojo.
- Índices VI ajustados por suelo: Los índices ajustados por suelo minimizan el efecto del brillo del suelo (sin follaje) en los índices de vegetación que incluyen las zonas del rojo e infrarrojo cercano, este se da especialmente en zonas donde la densidad de vegetación no es alta, y por ende índices como el NDVI se verían distorsionados. Por ejemplo, SAVI (Soil Adjusted Vegetation Index) trabaja también con las bandas R y NRI, pero incorpora un factor L conocido como índice condicionador de suelo que mejora la sensibilidad en los casos mencionados. El valor de L se fija de acuerdo a las condiciones ambientales específicas de cada aplicación, y varía, entre otras cosas, de acuerdo al porcentaje de suelo cubierto por vegetación. Debido a esto último,el índice MSAVI utiliza una función de L en lugar un L constante. Existen otras variaciones de SAVI que incorporan compensaciones de acuerdo al tipo de suelo, entre otras.
- Transformación Tasseled Cap (GVI, YVI, and SBI): Kauth y Thomas encontraron una lógica en patrones hallados en imágenes espectrales de LANDSAT correspondientes al ciclo de vida de algunos cultivos y propusieron una transformación a la que llamaron “Tasseled Cap”. Dicha transformación va desde un el conjunto de bandas disponible en la imagen provista por el sensor a un nuevo conjunto de bandas representativas en el mapeo de vegetación (por ejemplo: grado de vegetación verde, grado de vegetación amarilla, humedad, características del suelo, etc.). Esta transformación opera, de manera similar a la técnica de análisis de componente principal (PCA), mediante una combinación lineal de las bandas de la imagen original.
- VI basados en percepción remota mediante sistemas aéreos no tripulados (UAS): Los sistemas aéreos no tripulados trabajan a baja altura ofreciendo una muy buena resolución espacial, temporal y poca interferencia de factores atmosféricos, a un costo muy competitivo. Los sensores de estas plataformas operan en general en las bandas visibles (aunque cada vez existen más sensores que incorporan bandas del infrarrojo). Por el motivo anterior, existen una variación de NDVI comúnmente utilizada en estas aplicaciones, que opera en las bandas visibles del espectro en 800 nm y 680 nm. Además, específicamente para este tipo de sensores, considerando las características de la vegetación saludable y el uso del rango visible, se propone el índice VDVI (Visible-Band Difference Vegetation Index) que incorpora las tres bandas del espectro visible RGB generando un índice con una excelente precisión.

- Índices relacionados con el estado de la vegetación: El índice WDRVI (Wide Dynamic Range Vegetation Index) es similar al NDVI pero introduce un coeficiente que permite regular la contribución del componente NIR frente al R, esto mejora el rango dinámico en ambientes de alta biomasa, condición en que NDVI tiende a saturar. Por otro lado, se propone un índice que permite estimar la absorción de clorofila mediante mediciones en tres puntos del espectro (550 nm, 670 nm y 700 nm), este índice se denomina CARI (Chlorophyll Absorption Ratio Index). Dada la relación directa existente entre el suministro de agua y el estado del cultivo, y mediante la utilización de sensores infrarrojo termal, se propone la utilización del índice CWSI (Crop Water Stress Index) que permite medir el estado de irrigación del cultivo (utilizando como entrada las mediciones del sensor infrarrojo térmico y las mediciones de temperatura del cultivo estresado y el cultivo bien regado).

Siglas del VI	Nombre del VI	Definición
NVDI	Normalized Difference Vegetation Index	$\frac{\rho_{NRI} - \rho_R}{\rho_{NRI} + \rho_R}$
RVI	Ratio Vegetation Index	$\frac{R}{NRI}$
DVI	Difference Vegetation Index	$NIR - R$
ARVI	Atmospherically Resistant Vegetation Index	$\frac{NIR - RB}{NIR + RB}$
SAVI	Soil Adjusted Vegetation Index	$\frac{(\rho_n - \rho_r)(1 + L)}{(\rho_n + \rho_r + L)}$
MSAVI	Modified Soil Adjusted Vegetation Index	$0.5*\{2R_{800}+1-\sqrt{(2R_{800}+1)^2 - 8(R_{800} -$

Siglas del VI	Nombre del VI	Definición
GVI	Green Vegetation Index	$-0.290MSS_4-0.562MSS_5+0.600MSS_6+$
YVI	Yellow Vegetation Index	$-0.829MSS_4-0.522MSS_5+0.039MSS_6+$
SBI	Soil Brightness Index	$+0.433MSS_4-0.632MSS_5+0.586MSS_6+$
NVDI	Normalized Difference Vegetation Index	$\frac{\rho_{800}-\rho_{680}}{\rho_{800}+\rho_{680}}$
VDVI	Visible-Band Difference Vegetation Index	$\frac{(2*\rho_{green}-\rho_{red}-\rho_{blue})}{(2*\rho_{green}+\rho_{red}+\rho_{blue})}$
WDRVI	Wide Dynamic Range Vegetation Index	$\frac{(\alpha\rho_{nir}-\rho_{red})}{(\alpha\rho_{nir}+\rho_{red})}$
CWSI	Crop Water Stress Index	$\frac{(T_{canopy}-T_{nws})}{(T_{dry}-T_{nws})}$

Aplicación de la percepción remota en la agricultura La agricultura es una de las áreas de explotación de la percepción remota desde hace más de cuatro décadas. Originalmente, las actividades de investigación estuvieron orientadas a la clasificación de tipos de cultivos y estimación área cultivada, y con el tiempo el foco se fue moviendo hacia la caracterización de las propiedades físicas y biológicas de los cultivos y su contexto.

Las posibilidades de aplicación se fueron ampliando, a medida que se generó o se

incorporó tecnología habilitadora, como ser: la evolución de los sensores mediante la incorporación de bandas en sensores multiespectrales, el movimiento hacia sensores hiperespectrales, el aumento en la resolución de los mismos, el uso de sensores más avanzados en aeroplanos tripulados y la más reciente incorporación de vehículos aéreos no tripulados como plataforma (y la construcción de sensores livianos específicamente destinados a estos).

Según Shanmugapriya y otros en la revisión realizada en Shanmugapriya, Rathika, Ramesh, & Janaki (2019) se agrupan las aplicaciones en agricultura en categorías que comentaremos aquí:

- Monitoreo de área de cobertura: Fue la primer aplicación de la percepción remota y, quizás, la más difundida. Provee herramientas para estimación de superficie cultivada, clasificación de cultivos y evaluación de rendimiento. En este punto se da la utilización de los llamados Índices de Vegetación (que elaboramos en un punto específico de este trabajo), estos permiten caracterizar, discriminar y modelar tanto cultivos como los parámetros físicos de los sistemas agrarios. Estos Índices de Vegetación han sido foco de investigación por décadas y han ido evolucionando, tanto para acompañar la aplicación de nuevas tecnologías (paso de sensores multiespectrales a hiperespectrales, incorporación de nuevos sensores en drones, etc), como para ser destinados a aplicaciones más específicas (por tipo de cultivo, por área geográfica).
- Evaluación de estado del cultivo: Los cultivos sometidos a algún tipo de estrés (irrigación deficiente, faltante de nutrientes o minerales) cambian sus características de reflectancia (absorción) espectral. Estas variaciones pueden ser detectadas por técnicas de percepción remota de forma de proveer información precisa y oportuna para la toma de decisiones que corrijan estas situaciones (de ser necesario) y posteriormente evalúen el resultado de las mismas. El muestreo en intervalos regulares suele ser necesario para permitir diagnosticar situaciones de estrés y estimar los efectos de estas en los resultados. En esta aplicación, también, el uso de índices de vegetación es una práctica común.
- Estimación de nutrientes y agua: La evaluación del estado de los cultivos mediante percepción remota junto con la utilización de Sistemas de Información Geográfica, habilitan la aplicación de técnicas de agricultura de precisión como dosificación variable VRT que permiten una utilización de insumos cercana al óptimo para cada sector del cultivo (tema importante, por ejemplo, en zonas áridas donde el agua es limitada).
- Evapotranspiración del cultivo: La estimación de la evapotranspiración es esencial en el diseño y evaluación del esquema de irrigación. Existen modelos que permiten estimarla a partir de imágenes de índice de vegetación proveniente de percepción remota, el más conocido de ellos está basado en NDVI.

- Discriminación y control de malezas: Dado que las malezas crecen en pequeñas áreas del campo, el control mediante técnicas de agricultura de precisión (como dosificación variable y mapas de) es preferible al tratamiento homogéneo del campo. Aunque los procedimientos basados en el análisis de diferencias de respuesta espectral son los más difundidos (índices de vegetación, firma espectral), el análisis de texturas (mientras que las técnicas espectrales se basan en el análisis de la unidad mínima del sensor - el pixel - estas técnicas se basan en la identificación de patrones dentro de las inmediaciones de pixels adyacentes) y análisis fenológico (series temporales de imágenes son comparadas, típicamente utilizando índices de vegetación) también puede ser aplicado, bajo algunas circunstancias con mejores resultados Bradley (2014). Dado las necesidades en cuanto a resolución espacial de esta aplicación, la detección de pequeñas áreas afectadas con malezas es un desafío teniendo en cuenta las características de los sensores montados en plataformas satelitales y aéreas, aunque la aplicación de vehículos aéreos no tripulados (típicamente drones) es promisorio en esta aplicación.
- Infestación de plagas y enfermedades en cultivos: Cuando las plantas están expuestas a agentes patógenos, en etapas tempranas de la infección, se activan mecanismos fisiológicos de respuesta como la reducción de proceso de fotosíntesis que induce un incremento en la fluorescencia y emisión de calor. En etapas posteriores los patógenos causan una reducción del contenido de clorofila en el follaje que incrementa la reflectancia en el campo visual y genera un corrimiento hacia la banda del rojo del espectro, además de cambiar la densidad del follaje y la superficie del mismo, lo cual puede observarse en la banda del infrarrojo cercano Martinelli et al. (2015). La percepción remota puede proveer información en ambos estadios que permite la detección, identificación y cuantificación de la infestación. Al igual que el punto anterior, la resolución espacial es un desafío y suelen verse el uso complementario de percepción remota con percepción proximal y/o inspección local.
- Pronóstico de rendimiento: Los pronósticos de rendimiento permiten al productor tomar decisiones informadas con anticipación a la cosecha (capacidad necesaria para la recolección, almacenamiento, transporte, comercialización, etc.). Para los gobiernos, además, son valiosos dado que permiten hacer predicciones a escala regional en temas tales como políticas alimentarias, recaudación impositiva, entre otros. Con este fin, varios estudios proponen esquemas donde se alimentan modelos de Cosecha con variables del estado del follaje y propiedades del suelo provenientes de Percepción Remota mediante técnicas específicas de Asimilación de Datos Jin et al. (2018).

4.4.5 Percepción Proximal de Suelos y Plantas

La principal característica de Percepción Proximal es que el instrumento de medición se encuentra cerca del objeto medido, esta situación, marca además la principal diferencia con Percepción Remota (tema de la sección anterior). El propósito principal de esta práctica es la cuantificación de la variación temporal y espacial en los cultivos y suelos.

Los sensores proveen información sobre el estado de la planta o el suelo en un instante de tiempo dado, sin embargo es importante integrar esta información de modo de obtener conocimiento sobre la dinámica del sistema de producción agrícola.

Según Viscarra Rossel, Adamchuk, Sudduth, McKenzie, & Lobsey (2011) puede definirse percepción proximal en suelos como el uso de sensores emplazados en el territorio para obtener señales del suelo estando el detector en contacto directo o cercano (a menos de 2 metros) del mismo. Los sensores proveen información acerca del suelo ya que las señales corresponden a mediciones físicas relacionadas con él mismo y sus propiedades. Esta definición puede ser extendida al otro objeto de estudio de esta sesión: las plantas.

Si bien los estudios sobre percepción proximal suelen enfocar el tema de la variabilidad del suelo desde un punto de vista más amplio que la aplicación a la agricultura (incluyendo aplicaciones como arqueología, minería, ecología y ciencias naturales) encontramos en ellos una valiosa fuente de información para el presente trabajo.

Mecanismo de medición:

- No invasivos: No existe un contacto directo entre el sensor y el objeto de la medición.
- Invasivos: Existe contacto entre el sensor y el objeto de la medición. Dentro de estos podemos nombrar dos variantes: In situ y Ex situ
  - In situ: La medición se realiza en el lugar donde se encuentra el objeto.
  - Ex situ: Un mecanismo de muestreo se utiliza para tomar observaciones y analizarlas a posteriori.

Fuente de energía:

- Pasivo: Él sensor mide algún tipo de radiación de energía ya existente en la naturaleza, ya sea la radiación del propio objeto (como en la termografía) o bien la reflectancia del objeto al ser expuesto a una fuente externa (como el son).
- Activo: El sensor posee una fuente de radiación artificial (radar, láser, fluorescencia) y mide las modificaciones de radiación que se producen por la interacción con el objeto. Los sensores activos se ven menos afectados por la incidencia de factores ambientales.

Tipo de operación:

- Móvil: el proceso de adquisición se lleva a cabo, o bien, en forma continua mientras el sensor es transportado (on-the-go) o bien, haciendo pequeñas detenciones para realizar las mediciones (stop-and-go). Esta metodología permite obtener un buen registro de la variación espacial, mientras la variación temporal debe ser cuidadosamente manejada.
- Fijo: el sensor se encuentra emplazado en un punto fijo realizando mediciones, en general, continuas. Las capacidades en cuanto a registro de la variación temporal son muy interesantes. En general este tipo de sensores trabajan de forma conjunta, a lo largo del campo, de forma de proveer un conjunto de mediciones de acuerdo a una disposición determinada.

Mecanismo de inferencia:

- Directa: La medición se basa directamente en un proceso físico que arroja como resultado la propiedad observada.
- Indirecta: Se utiliza una función de transformación (función de pedotransferencia si se trata de suelos) o modelos que permiten obtener el valor de la propiedad del objeto mediante datos auxiliares recolectados.

Tipos de sensores y aplicaciones Las técnicas de medición utilizadas en Percepción Proximal son categorizadas de forma homogénea por los autores que desarrollan esta temática, tomaremos las categorías propuestas en Adamchuk, Hummel, Morgan, & Upadhyaya (2004) y comentaremos brevemente cada una a modo de marco para la descripción de las aplicaciones en agricultura de precisión.

Sensores eléctricos y electromagnéticos Entre los métodos geofísicos de exploración, según Samouëlian, Cousin, Tabbagh, Bruand, & Richard (2005), aquellos basados en las propiedades eléctricas son especialmente promisorios ya que las mismas están fuertemente correlacionadas con los materiales del suelo y sus propiedades.

La sal que contiene el agua utilizada para irrigación persiste en el suelo mientras que el agua pura vuelve a la atmósfera mediante evaporación y transpiración de las plantas. Los efectos de la salinidad en el suelo se manifiestan en los cultivos como baja en la tasa de crecimiento, perdida de vigor y bajas en el rendimiento. La medición de la conductividad eléctrica en el suelo es una de las maneras más sencillas de cuantificar y monitorear la salinidad en el suelo. Corwin & Lesch (2003)

Pese a que existen procedimientos de laboratorio que permiten determinar la salinidad del suelo, estos son costosos, complejos y lentos, lo cual los hace poco apropiados para generar mediciones que reflejen la variabilidad espacial y temporal del suelo con miras de ser aplicadas en agricultura de precisión. Los métodos más comúnmente usados son la medición de resistividad eléctrica ER (por ejemplo mediante el método de Wenner o alguna variante) y la medición por inducción electromagnética.

El sensor ER basado en el método Wenner funciona con cuatro electrodos que son introducidos en el suelo a la misma profundidad y en línea recta, en dos de los polos (electrodos) se inyecta una tensión continua o de baja frecuencia, y en los dos polos restantes (electrodos de potencial) se mide la diferencia de potencial eléctrico. Una ecuación basada en la distancia entre los polos, su disposición geométrica y la diferencia de potencial medido permite calcular la resistividad del suelo. Las mediciones pueden ser tomadas en la superficie o a distintas profundidades (insertando los polos en el suelo), siendo esta última técnica obviamente más invasiva.

La resistividad eléctrica es función de las propiedades del suelo: naturaleza de los sólidos contenidos, distribución de los huecos, grado de saturación de agua, resistividad eléctrica de los fluidos (concentración de sustancias disueltas) y temperatura. Estos parámetros afectan la resistividad de diferentes maneras, variadas investigaciones se han realizado para establecer la relación de cada uno de ellos con la misma. (Samouëlian et al. (2005))

Por otro lado, el método de medición de conductividad en el suelo mediante inducción electromagnética EMI tiene un gran potencial ya que permite evaluar las propiedades del suelo (contenido de agua, sal, etc.) de forma no invasiva. Este tipo de sensores utilizan dos bobinas, una de ellas (la bobina transmisora) induce Corrientes de Foucault (conocidas también como corriente torbellino o en ingles Eddy-current) circulares en dirección al suelo, donde se genera un campo electromagnético secundario que es proporcional a la corriente que fluye por su interior, finalmente una fracción de la energía que circula por el secundario es inducida en la segunda bobina (receptor). Ambas señales (la emitida y la recibida) se diferencian en fase y amplitud de acuerdo a las propiedades del suelo, lo que permite deducir sus características (Sudduth, Drummond, & Kitchen (2001)).

En Sheets & Hendrickx (1995) se realiza una comparación entre EMI y el método de dispersión de neutrones (medición del contenido de hidrógeno)que muestra que el primero cuenta con una precisión aceptable, a la vez que es fácil de usar y rápido, haciéndolo ideal para mediciones no invasivas de humedad en el suelo. Por otro lado, una ventaja de los métodos EMI frente a ER es que, dado que no requieren contacto físico con el suelo permiten realizar mediciones rápidas, en movimiento y no destructivas (no afectan sus propiedades al realizar las mediciones) y son, por ello, ideales para la realización de mapeo de suelo siendo montados en plataformas móviles.

Sensores ópticos y radiométricos Espectrometría de rayos gama:

Los rayos gamma son un tipo de radiación electromagnética (por esto, compuesta de fotones o cuantos) de alta energía y longitud de onda corta emitida por isotopos naturales. Existen varios isotopos radiactivos que emiten rayos gama en la naturaleza, pero solamente las series de desintegración del Potasio 40 (K), Uranio 238 (U) y Torio 232 (Th) lo hacen con una intensidad suficiente como para ser detectados con espectroscopia de rayos gama. Estos isotopos están presentes, en distinto grado, de forma natural, en suelos y rocas. Los espectrómetros gama pasivos hacen uso de estos isotopos, los sensores activos, en cambio, hacen uso de una fuente radioactiva que

emite fotones de energía que son luego detectados por el espectrómetro. El contenido de agua y la densidad aparente del suelo son los principales factores que atenúan el paso de los rayos gama, la textura y mineralogía del suelo lo hacen en también aunque en menor grado. El análisis de suelo por radiometría de rayos gama es no invasivo y no destructivo, provee alta definición espacial, puede arrojar buenos resultados incluso cuando existe una capa de vegetación sobre el suelo en estudio. (Mahmood, Hoogmoed, & van Henten (2013))

Espectrometría Ultravioleta, Visible e Infrarrojo:

Son sensores que utilizan las ondas de electromagnéticas para detectar la cantidad de energía absorbida/reflejada por el suelo o el follaje de las plantas. Funcionan bajo principios similares a los sensores de Percepción Remota descritos anteriormente, presentando algunas ventajas como mayores resoluciones espaciales y temporales, a expensas de, generalmente, mayores costos de operación. Los sensores ópticos activos, emiten energía electromagnética en dirección al destino de la medición, el objeto irradiado absorbe una parte de la energía y otra la refleja, esta última es detectada mediante un elemento fotosensible. Los sensores ópticos pasivos, no irradian el objeto de medición, la energía proveniente de una fuente natural (como el sol) es utilizada en este caso. Los primeros funcionan en un entorno más controlado, sin verse afectados por factores externos como las nubes, ubicación de la fuente con respecto al objeto y otros fenómenos atmosféricos que distorsionan las condiciones de medición.

El desarrollo de Índices Espectrales (comentado anteriormente cuando desarrollamos el tema de Percepción Proximal) que permiten reducir las mediciones multiespectrales a una única variable ha contribuido en gran medida al desarrollo de este tipo de sensores. Como menciona Holland, Lamb, & Schepers (2012) los índices que combinan las longitudes de onda cercanas al rojo (Red) y el infrarrojo cercano (NIR) presentan buenos resultados en la estimación de la cantidad y calidad de vegetación verde. Los índices NDVI (normalized difference vegetation index) y SRI (simple ratio index), descriptos anteriormente en este trabajo, están relacionados con la pigmentación de los cultivos y por esto son adecuados en la estimación de biomasa y estrés por falta de irrigación. Algunos índices incluyen factores de compensación para evitar la incidencia del suelo y otras superficies que no son follaje en la medición, dentro de esta categoría el citado trabajo menciona los índices SAVI, NLI, MNLI y MSR.

Dentro de los sensores ópticos activos, pueden distinguirse los que operan con una fuente de luz fija y los que lo hacen con una fuente de luz modulada. Durante la medición es necesario tener en cuenta la presencia de fuentes de energía externas que irradian el objeto y afectan a la misma. Los sensores que operan con luz modulada, son capaces extraer, de la señal registrada por el sensor, las componentes iluminadas en una fase diferente a la usada para modular la luz emitida, de esta forma pueden operar de igual manera con o sin presencia de fuentes de luz externas como el sol.

Este tipo de sensores operan usualmente en las bandas visible (V) e Infrarrojo cercano (NIR), el desarrollo de nuevos emisores que trabajan en la banda del Ultravioleta e

infrarrojo medio (MWIR), y los detectores asociados, extendieron el rango de longitud de onda dotando de nuevas posibilidades en evaluación de necesidades de agua y detección de enfermedades en plantas, así como análisis de composición orgánica en suelos.

Time Domain Reflectometry (TDR):

Un sensor TDR opera con una sonda inmersa en el objeto a medir (en nuestro caso el suelo) por la cual se hace circular una señal electromagnética proveniente de un generador de pulsos, parte de la señal inyectada viajará desde un extremo de la sonda al opuesto y volverá al inicio. Mediante un mecanismo de muestreo tipo osciloscopio se mide el tiempo t de propagación de la señal electromagnética a través de la sonda. Hallado t puede ser calculado la constante dieléctrica del medio k. Dado que, de los componentes que pueden hallarse en el suelo, el agua tiene un coeficiente dieléctrico mucho mayor a los demás, la cantidad de agua en el suelo puede ser estimada dado el coeficiente dieléctrico del suelo con una buena precisión. La constante dieléctrica aparente k se ve levemente afectada por la temperatura, la salinidad, textura y densidad del suelo. Noborio (2001)

Frequency Domain Reflectometry (FDR):

Este tipo de sensores operan sobre preceptos similares a TDR. Se inyectan ondas electromagnéticas del orden de los 150 MHz en sondas insertas dentro del suelo, se miden las frecuencias de las ondas reflejadas. Las mismas varían de acuerdo a las propiedades dieléctricas del suelo de acuerdo a su capacitancia. Las sondas son capacitores (están construidas por dos cilindros montados de forma coaxial separadas por un elemento dieléctrico) que interactúan con el suelo que las circunda de forma que la capacitancia medida está relacionada con él. El determinar la capacitancia del medio, permite conocer la constante dieléctrica que está relacionada con la humedad del suelo (dado que el agua posee mayor coeficiente dieléctrico comparado con los otros componentes del suelo). Guadalupe Ramos Hernández, Gracia-Sánchez, Patricia Rodríguez-Martínez, & Adalberto Zuñiga-Morales (2019)

Ground Penetrating Radar (GPR):

El principio de funcionamiento de los sensores GPR es similar al del sonar. El radar produce pulsos de energía electromagnética de muy alta frecuencia (en el orden de 100Mhz a 1Ghz) que son transmitidos hacía la tierra, una o más antenas reciben el reflejo de dichos pulsos electromagnéticos cuando regresan a la superficie. Dado que la propagación de la energía electromagnética dentro del suelo depende de sus propiedades eléctricas, la evaluación de las señales recibidas permite generar un perfil de dichas propiedades del suelo y, dado que estas están principalmente controladas por el contenido de agua, puede obtenerse un buen indicador de humedad del suelo en varias profundidades, sumado al uso de técnicas de GPS, un mapa de humedad puede ser trazado desde una plataforma móvil en la superficie (este método es no invasivo). Aunque siempre el objetivo es determinar la velocidad de la señal y su atenuación dentro de la tierra, existen variaciones en el modo de operación del radar, siendo el más común el perfil por reflexión (reflection profiling), entre los más comunes podemos nombrar: transiluminación (common midpoint - CMP), reflexión y refracción gran

angular (wide-angle reflection and refraction - WARR), y transiluminación. Davis & Annan (1989)

Sensores mecánicos Los sensores mecánicos son aplicados normalmente para medir la densidad aparente / compactación del suelo. Este parámetro es especialmente interesante pues está relacionado con cuestiones como el crecimiento de las plantas, erosión del suelo, o incluso consumo de combustible durante la labranza. El sensor posee una herramienta que se introduce en el suelo y un mecanismo de medición de la resistencia mecánica del suelo sobre dicha herramientas mientras se aplica fuerza sobre la misma. Los penetrómetros cónicos son utilizados normalmente para medir la resistencia del suelo a la penetración (acción perpendicular a la superficie), requieren que la plataforma se detenga mientras se ejecuta la medición, hasta los mecanismos automatizados consumen mucho tiempo por este motivo. Existen también penetrómetros horizontales, que miden la resistencia del suelo mientras se arrastra una herramienta introducida en él. Es interesante la integración de medidores de tensión en arados, ya que la medición de la resistencia es directa y puede hacerse en simultáneo con la mecanización del suelo. Adamchuk et al. (2004) Es importante observar que el resultado de la medición dependerá, además de la densidad aparente del suelo, de otras características, principalmente su humedad, es por esto interesante la combinación del penetrómetro con un sensor de humedad de suelo.

Sensores acústicos Los sensores acústicos utilizan transductores capaces de registrar pequeñas variaciones de sonido que se correlacionan con variaciones en las propiedades del suelo o de las plantas. Los sensores activos inyectan una señal sonora en el medio y luego registran, mediante transductores, las reflexiones que vuelven hacia el sensor. En cambio los pasivos, no poseen un emisor y se limitan a obtener una representación de las señales sonoras (en general filtradas en bandas de frecuencia específicas) que generan las plantas en su actividad (por ejemplo, crecimiento de las raíces en Shimotashiro, Inanaga, Sugimoto, Matsuura, & Ashimori (1998)).

En Lu & Sabatier (2009), los autores describen un experimento que busca explorar la relación entre las variaciones temporales en la velocidad del sonido debidas a cambios en a características físicas del suelo (potencial hídrico del suelo, humedad, temperatura) en condiciones naturales. Para esto disponen emisores de sonido y transductores, así como sensores TDR, tensiómetros y medidores de temperatura en suelo, en una porción de tierra al aire libre en un periodo de dos años. Los autores dicen que los datos experimentales son armoniosos con los resultados esperados según la teoría. Y afirman que el estudio sugiere que las mediciones de la velocidad del sonido puede ser utilizada como una herramienta para la medición del potencial hídrico del suelo.

Existe también una relación entre la velocidad del sonido en el medio y la humedad del suelo, esta cuestión se explora en Adamo, Andria, Attivissimo, & Giaquinto (2004). Los autores muestran un modelo que permite estimar la cantidad de humedad en el suelo,

según ciertas condiciones de validez y frecuencias de sonido propuestas, para un amplio rango de suelos de interés agrícola.

La aplicación de técnicas de tomografía a este tipo de sensores (ver) permite construir tomografías del contenido de agua en dos dimensiones de forma no invasiva. En Blum, Flammer, Friedli, & Germann (2004) los autores ensayan un proceso en el que los tiempos de viaje de las señales acústicas se convierten en una distribución de velocidades mediante un algoritmo de tomografía, estas velocidades son traducidas a contenido de agua resultando en un mapa de contenido de agua en suelo (los resultados son contrastados empíricamente en el paper).

La utilización de transductores que se acoplan al arado y, mediante el procesamiento y análisis automático de las ondas sonoras, permiten la determinación sobre la marcha de la profundidad de mecanización óptima (de acuerdo a las capas de compactación detectadas) a medida que la herramienta avanza en el campo es promisoria, ya que el suelo es mecanizado solamente hasta la profundidad necesaria ahorrando energía y evitando la erosión el suelo. T. E. Grift, M. Z. Tekeste, & R. L. Raper (2005)

Sensores neumáticos Los sensores neumáticos miden la presión requerida para forzar un volumen de aire dado dentro del suelo a una determinada profundidad, dicha presión está relacionada ciertas propiedades del suelo tales como su estructura y compactación (Adamchuk & Viscarra Rossel (2011)). La compactación del suelo tiene efectos negativos tanto en la agricultura, como en el medioambiente. La compactación afecta la estructura del suelo, reduce la productividad de los cultivos, aumenta la escorrentía y la erosión, empeora la contaminación de las aguas superficiales causada por desechos orgánicos y agroquímicos, y causa el uso ineficiente del agua y nutrientes debido al drenado lento (Hemmat & Adamchuk (2008)). Este tipo de sensores constituyen mayormente un área de estudio, siendo muy escasas las aplicaciones comerciales que se han dado a esta técnica.

Sensores electroquímicos A diferencia de los sensores que se nombraron anteriormente, los cuales están orientados a relevar las propiedades mecánicas o físicas del suelo (o de los cultivos, en algunos de los mencionados), los sensores electroquímicos están orientados al relevamiento de las propiedades químicas del suelo, como ser el PH y contenido de nutrientes (nitrato, potasio y existen algunos avances en la detección de fósforo).

Un sensor electroquímico consiste de una membrana que responde de forma selectiva a iones de un determinado tipo (H+, K+, NO3−, Na+, entre otros) y un transductor que transforma dichas reacciones en una señal eléctrica mensurable. Los sensores electroquímicos más utilizados en aplicaciones de agricultura suelen ser de dos tipos: electrodo selectivo de iones (ISE) y transistor de efecto de campo sensible a iones (ISFET).

En los sensores ISE, la actividad de un ion especifico genera un cambio de potencial, con respecto a un electrodo de referencia, que es convertido en potencial eléctrico mediante el transductor. Estos sensores son capaces de medir contenido de Nitratos y Potasio en el suelo, además existen algunos avances en la detección de Fósforo. Los ISE no son aptos para medición en línea debido a que su tiempo de respuesta es de varios minutos. (Lin, Wang, Zhang, Zhang, & Chen (2008))

Los ISFET siguen los principios de los ISE, pero la membrana se encuentra montada sobre la capa aislante de un transistor FET, lo que permite que la corriente que pasa a través del transistor varíe de acuerdo a la actividad de un ion específico. Los ISFET tienen ciertas ventajas sobre los ISE, son mucho más pequeños, tienen mejor relación señal ruido, respuesta rápida y pueden ser integrados varios sensores en un solo chip. Hay aplicaciones en las que detectan amonio, nitrato, potasio. (Lin et al. (2008))

Fusión de Sensores / Sistemas multisensor Ya desde hace varios años, en el terreno de los sistemas de percepción remote, se ha puesto énfasis en la fusión de sensores, donde sensores que registran distintos fenómenos físicos se integran en una plataforma única y/o bajo un mismo proceso de adquisición de datos (Adamchuk & Tremblay (2017)).

La aplicación de estas ideas en la órbita de los sistemas de agroinformática está inspirada en el concepto de fusión de datos multisensor MDF (por las siglas de Multisensor Data Fusion). Según Huang, Lan, Hoffmann, & Lacey (2007), MDF es una tecnología que fusiona datos de múltiples sensores permitiendo realizar estimaciones más precisas del medio a través de mediciones y detección. Las áreas de aplicación de MDF son amplias, desde aplicaciones militares hasta aplicaciones civiles como vigilancia, monitoreo de maquinaria compleja, diagnostico médico, edificios inteligentes, control de calidad de alimentos y agricultura de precisión, entre otros. Las técnicas utilizadas vienen de un amplio rango de disciplinas, como procesamiento de señales, reconocimiento de patrones, estimación estadística, extracción del conocimiento y sistemas de control.

Dado que cada técnica de medición tiene sus fortalezas y debilidades, y que un solo sensor no es capaz que medir todos las propiedades del suelo (dadas las interdependencias entre las variables), la selección de un conjunto complementario de sensores para medir un juego de características del suelo es importante. La integración de varios sensores en una única plataforma puede proveer algunos beneficios operacionales sobre la utilización de un único sensor, como ser: robustez de operación, mayor confianza dado que se realizan mediciones independientes en la misma porción de suelo, cobertura de atributos extendida, y dimensionalidad del espacio de medidas incrementada. (Viacheslav I. Adamchuk, Raphael A. Viscarra Rossel, Kenneth A. Sudduth, & Peter Schulze Lammers (2011))

Existen, según Huang et al. (2007), tres enfoques en lo referente a la fusión de los datos de los sensores, cada uno de ellos está motivado por el tipo de aplicación y utiliza distintos métodos para realizar la fusión de los datos. Estos enfoques son: 1) Fusión directa de los datos; 2) Representación de los datos mediante vectores característicos, y su consiguiente

fusión; 3) Procesamiento de cada sensor hasta obtener inferencias o decisiones que serán luego combinadas.

En Liggins, Hall, & Llinas (2009) los autores enumeran tres ventajas principales de usar MDF, estás son:

- Si se utilizan varios sensores idénticos, se aprovecha una ventaja estadística al utilizar N observaciones independientes. El mismo resultado puede obtenerse si se combinan N observaciones un único sensor.
- Si se utilizan parámetros conocidos que regulan la relación entre los sensores, se agrega un conocimiento adicional a la información. Por ejemplo si se utiliza la posición relativa de múltiples sensores, el proceso puede generar información más enriquecida (por ejemplo dos sensores que miden la velocidad angular a un objetos, pueden ser combinados por tiangulación y estimar la posición del mismo).
- La combinación de sensores con habilidades complementarias mejora la observabilidad. Ampliar la línea base de observaciones mejora el resultado general. La mayoría de las aplicaciones comerciales de percepción proximal para agricultura de precisión en la actualidad utilizan en mayor o menor medida un enfoque de fusión de sensores,

5 Intersección entre los tópicos

Desde hace ya varias décadas, las maquinarias agrícolas comenzaron a incorporar sensores y sistemas de bitácora que permiten recolectar grandes cantidades de información durante sus actividades rutinarias. Con la llegada de IoT y la computación en la nube, este proceso creció exponencialmente tanto en cantidad, como en calidad. Luego, los conceptos involucrados en Agricultura de Precisión y Smart Farm despegaron a estas maquinarias de la idea de que son simples recolectores de datos, generando un sistema capaz de controlar y mejorar las actividades agrícolas. A medida que se incremente la adopción de sensores y robots agrícolas, se incrementará la cantidad y alcance de los datos, permitiendo que los procesos agrícolas sean guiados y habilitados por ellos (data driven and data enabled). Es en este contexto que Grandes Datos cobra un rol fundamental, dada las necesidades de gestión de las grandes cantidades de datos recolectadas. Además, las herramientas de agricultura de precisión, junto con técnicas de extracción del conocimiento e inteligencia artificial que permiten generar modelos predictivos, generan la posibilidad de proveer recomendaciones personalizadas, en una escala espacial apropiada, que permiten mejorar notoriamente la productividad agrícola y el desempeño del medioambiente. Morimoto (2018)

Tomando Agroinformática como punto de pivote, en los siguientes párrafos, exploraremos la intersección entre ésta y los otros temas de interés: Internet de las Cosas, Grandes Datos, y Extracción del Conocimiento. Se han identificado 10 espacios de intersección, cada uno de los cuales será analizado en una sección dedicada: Variabilidad temporal y espacial (Sección 5.1), Zonas de Manejo (Sección 5.2), Monitoreo en tiempo real y soporte de decisiones (Sección 5.3), Percepción Remota (Sección 5.4), Percepción Proximal de Suelos y Plantas (Sección 5.5), Sistemas de dosificación variable (Sección 5.6), Fenotipado de alto rendimiento (Sección 5.7), Detección y manejo de pestes (Sección 5.8), Fertilización y Control de Pestes (Sección 5.9) y Agricultura inteligente (Sección 5.10)

Descubriremos así que las intersecciones entre cada línea guía y cada una de las temáticas nombradas no son vacías, así mismo, en general la línea atraviesa a todas y cada una de los dominios nombrado. Además, en una buena parte del material con el que trabajamos se observó la confluencia simultanea de más de una línea guía.

5.1 Variabilidad temporal y espacial

Desde hace ya varias décadas las maquinarias agrícolas permiten recoger información de rendimiento y otros datos operativos durante la ejecución de sus actividades rutinarias (lo que las transforma en ejemplos de IoT). Esta información, permite mapear la variabilidad temporal y espacial del campo, entre otras aplicaciones y, en muchos casos, no ha sido aprovechada. En Leroux, Jones, Taylor, Clenet, & Tisseyre (2018) los autores proponen un enfoque novedoso para generar zonas de administración, basado en estos datos, que saca provecho de la amplitud temporal registrada y se diferencia, de la más difundida, aplicación de técnicas de clustering con ese fin.

En Maestrini & Basso (2018) los autores, utilizan datos de rendimiento (ver sección "análisis de datos de rendimiento") de cultivos como soja, maíz, trigo y algodón en distintos campos del área central de Estados Unidos, junto con datos topográficos, de lluvias e información suelo de esas locaciones, para comprobar mediante técnicas estadísticas y de grandes datos, la relación del rendimiento con dichas variables. De esta forma comprueban que el rendimiento relativo en áreas con un alto índice de humedad depende en gran medida de los patrones de lluvia, porque tiende a rendir menos por anegarse en las épocas húmedas, y rendir más en las de sequía.

5.2 Zonas de Manejo

La delimitación de Zonas de Manejo (mencionadas en el capítulo anterior) tiene un papel importante en las actividades de Agricultura de Precisión, ya que permite la gestión del cultivo de acuerdo a las necesidades particulares de cada zona del campo (en lugar de tomarlo como un homogéneo). La técnica más usada para la delimitación de zonas de manejo es el análisis de los mapas de rendimiento (generados mediante el análisis de datos de rendimiento georeferenciados recolectados por sensores emplazados en las maquinarias utilizadas en la cosecha - existen desde hace varias décadas múltiples alternativas comerciales que cumplen este fin). En Damian et al. (2020) los autores exploran, un mecanismo alternativo al mencionado: la delimitación de zonas de manejo mediante información proveniente de Percepción Remota, en este caso imágenes satelitales LANDSAT en las que se calcula el Índice de Vegetación NDVI (índice de vegetación de diferencia normalizada). En el trabajo se comparan ambas técnicas (delimitación mediante información de rendimiento y NDVI) aplicando el algoritmo Fuzzy C-Means para delimitar las zonas, y luego se comparan los resultados mediante la matriz de correlación de Spearmans, llegando a la conclusión que ambas delimitaciones tienen una correlación de $0.48<r<0.61$, pudiendo usarse este método en forma alternativa o complementaria.

En el trabajo mencionado anteriormente, se contrastan dos metodologías de demarcación de zonas de manejo, por un lado la utilización de imágenes de Percepción Remota (que suscribe al enfoque basado en la utilización de datos sobre las propiedades del suelo y/o

del terreno) y por otro lado mapas de rendimiento (que constituye el segundo enfoque más utilizado), por su parte Miao, Mulla, & Robert (2018) propone un enfoque integrador (utiliza datos del suelo o el terreno, junto con información de rendimiento) llamado ROSE-YSTTS que utiliza: datos topológicos (elevación y pendiente) del terreno, materia orgánica, conductividad eléctrica del suelo, mapas de tendencias de rendimiento espacial, y mapa de estabilidad de rendimiento temporal. El delineamiento de zonas resultante se compara con el obtenido con otros dos métodos también integradores, llamados ROSE y CMYYM. En el proceso, luego de la normalización y limpieza de los datos, se aplica el algoritmo de fuzzy clustering provisto por el software Management Zone Analyst. Finalmente los clusters hallados son tratados según procedimientos prescriptos por cada uno de los tres métodos y los resultados comparados.

5.3 Monitoreo en tiempo cercano al real y sistemas de soporte de decisiones

En el ámbito de los sistemas de monitoreo para el agro, unos de los temas en juego es la interconexión de múltiples sensores de diferentes tipos (bajo las restricciones impuestas por el contexto en el que operan) y el análisis de los datos recolectados bajo los requerimientos de las tres V de big data (volumen, variedad y velocidad). En Xian (2017) se mencionan estas problemáticas y se analiza un modelo propuesto basado en tecnologías como ZigBee para la operación en malla (mesh network) de la red de sensores (que gestiona la comunicación entre ellos bajo un esquema de corto alcance, auto-organización y bajo consumo), los datos recolectados son enviados a la nube mediante IoT gateways basados en Raspberry Pi operando bajo el Sistema Operativo Raspbian (basado en Linux Debian). Dada la cantidad masiva de datos no es posible consultar directamente la información, para el análisis de grandes datos en este caso se utiliza un servicio provisto por Amazon llamado EMR (es una plataforma de grandes datos que utiliza software de código abierto como Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi y Presto), se utilizan servicios temporizados para analizar periódicamente los datos y almacenar la información resultante en una combinación de DynamoDB y Oracle Amazon RDS proveyendo al usuario final capacidades de consulta y visualización en tiempo cercano al real necesarias para el monitoreo de datos del agro.

Siguiendo una línea similar al trabajo anteriormente nombrado, los autores de Vuran, Salam, Wong, & Irmak (2018) describen la aplicación en Agricultura de Precisión de lo que se denomina IoUT por las siglas en inglés de Internet de las cosas subterráneas (Internet of Underground Things), o sea, la aplicación de redes de sensores emplazados de forma subterránea dedicadas a la percepción proximal con el propósito de monitoreo en tiempo cercano al real y asistencia en las tomas de decisiones en ambientes del agro. En general se comparten las restricciones del trabajo anterior en cuanto a disponibilidad de recursos de interconexión y disponibilidad de energía eléctrica, con algunas otras

propias del contexto (mantenimiento de los sensores subterráneos, atenuación de las señales electromagnéticas, etc.). Se describe, en el trabajo, una arquitectura completa de IoUT y se listan algunos desafíos en su aplicación (bajo costo, baja complejidad para no aumentar el consumo de energía, integración de los sensores - múltiples formatos, sensores de propiedades del suelo de bajo costo, resistencia a los factores climáticos, adaptación dinámica a los cambios en el medio de comunicación el suelo)

5.4 Percepción Remota

En cuanto a la temática de predicción de rendimiento, en la revisión de literatura Liakos, Busato, Moshou, Pearson, & Bochtis (2018), los autores mencionan algunos trabajos pertenecientes a la intersección entre percepción remota y extracción del conocimiento. Entre ellos, la aplicación de redes neuronales artificiales (ANN) en imágenes satelitales multitemporales de las bandas espectrales del rojo y NIR con el propósito de aplicarlos a la estimación de biomasa (estiman kg seco por Ha por día). Mencionan, además, otro trabajo en el que se fusionan imágenes satelitales con predicciones sobre las características del suelo donde se aplica también ANN con el fin de predecir el rendimiento del trigo dentro de la variabilidad del campo. En cuanto a la aplicación de imágenes RGB de alta resolución tomadas por un vehículo no tripulado (UAV) y se aplica k-means clustering, expectation maximization (EM) y self-organizing map (SOM) con el propósito de detectar los tomates de un cultivo.

El trabajo Maxwell, Warner, & Fang (2018) es una amplia revisión sobre el uso de machine learning en clasificación de datos provenientes de sensores de Percepción Remota. En este trabajo se mencionan como los seis métodos más maduros en este área: máquinas de soporte vectorial (SVM), arboles de decisión simples (DT), arboles de decisión potenciados, random forest (RF), redes neuronales artificiales (ANN) y k-nearest neighbor (k-NN). Mencionan algunos métodos más modernos, como extreme learning machines (ELM) y redes neuronales de convulsión profunda (Deep Convolution Neural Networks CNN), pero aclaran que toman los métodos de mayor madurez. Además de las tareas de clasificación mencionadas como tema principal de la revisión, también se resalta el uso de algunos de los métodos listados en funciones de regresión como estimación de área de cobertura del dosel, predicción de calidad de agua, estimación de contenido de clorofila, entre otros son nombrados.

5.5 Percepción Proximal de Suelos y Plantas

En Anastasiou, Castrignanò, Arvanitis, & Fountas (2019) los autores aplican geoestadística multivariada en un esquema de Fusión de Sensores con el fin de evaluar de forma muy precisa la variabilidad espacial y temporal, permitiendo de esta forma delimitar zonas de manejo homogéneas y aplicar a ellas procedimientos de agricultura

de precisión (esto permite cuantificar el impacto de la aplicación de nutrientes, agua, efectos de enfermedades y otros modificadores del crecimiento de las plantas). En este trabajo se fusionan datos provenientes de dos tipos de sensores de percepción proximal (tecnologías detalladas previamente en la sección Percepción Proximal de este trabajo): un sensor comercial de follaje (que permite tanto, la obtención de índices de vegetación clásicos, como de información básica de reflectancia del dosel y el suelo) y un sensor de inducción electromagnética (EMI). En el trabajo se muestra que el uso integrado de sensores de varios tipos, junto con la aplicación de técnicas de Fusión de Sensores, genera un ambiente propicio para la administración zonificada en cualquier tipo de escala espacial. Además, se muestra que las técnicas de geoestadística multivariada ponderan de forma satisfactoria los datos provenientes de dichos sensores, incluso, siendo el caso de que tienen soportes de medida dispares (dado que refieren a volúmenes de espacio y/o intervalos de tiempo disímiles).

El área de visión artificial a registrado grandes avances en la última década. Algunas de las técnicas mencionadas en la sección extracción del conocimiento, en particular Redes Neuronales Artificiales y Aprendizaje Profundo se utilizan exitosamente en problemas de visión artificial aplicados al Agro. En el trabajo Too, Yujian, Njuki, & Yingchun (2019) los autores exploran distintas técnicas de clasificación de imágenes mediante redes neuronales de convulsión profunda (Deep Convolution Neural Networks CNN) aplicándolas a la detección y clasificación de enfermedades en plantas. En la sección de trabajos relacionados, los autores, listan algunas aplicaciones de visión artificial tales como: diagnostico de enfermedades mediante smart phone, clasificación de especies de plantas que permite distinguir plantas de malezas (con aplicación potencial en la erradicación automática de las mismas), detección de frutos (con el objetivo de ser utilizado en estimación de rendimiento y cosecha automática), y algunos otros cuyo objetivo es la detección de ciertas enfermedades en particular. Uno de los grandes desafíos en este tipo de aplicaciones es la variabilidad en las imágenes con las que debe operar el modelo entrenado, muchos de los modelos que tienen éxito en las pruebas de laboratorio se desempeñan pobremente en el campo de aplicación cuando son utilizados con imágenes reales, fotogramas con iluminación variada, con ruido, diferentes inclinaciones y características propias de la cámara, existen distintos esfuerzos para superar estos desafíos, entre ellos el entrenamiento de modelos con imágenes sintéticas.

5.6 Sistemas de dosificación variable

Mantener apropiadamente la humedad del suelo mediante irrigación óptima de acuerdo al cultivo, las características de la superficie cultivada y las condiciones ambientales, permite utilizar recursos de forma óptima, maximizando resultados, bajando costos e impacto ambiental. En Muangprathub et al. (2019) se describe la utilización de sensores de humedad interconectados mediante una Red Inalámbricas de Sensores (WSN por las

siglas de Wireless Sensor Network) para recoger los datos en el campo que son analizados mediante minería de datos (mediante reglas de asociación) de forma de predecir las condiciones de crecimiento óptimo del cultivo y generar información de control que permita aplicar de forma automática o supervisada irrigación (mediante la participación del técnico que puede acceder a esta información en su teléfono inteligente y aplicar las decisiones necesarias desde el mismo).

La cuestión de la detección y erradicación de malezas, desde la intersección con IoT y Extracción del Conocimiento (en particular visión por computador), es tratada en Dankhara, Patel, & Doshi (2019) donde se describen tres enfoques para la detección de malezas en un robot equipado con visión artificial: detección mediante redes neuronales de convulsión (CNN), generación automática de un conjunto de datos para la detección precisa (SegNet), detección mediante la utilización de un conjunto de datos resumido (cNET). El propósito perseguido es la creación de un robot de erradicación automática de malezas mediante un robot de dosificación de herbicida que utilice la cantidad óptima del mismo, manteniendo a la vez al agricultor lejos del mismo con el fin de evitar sus efectos nocivos, tenga bajo impacto sobre el medioambiente y la salud de los consumidores, entre otros beneficios.

5.7 Fenotipado de alto rendimiento

La relación entre el fenotipo (organismo funcional de la planta), su genotipo (estructura genética) y el entorno en que se desarrolla, están íntimamente ligados con el rendimiento y la productividad vegetal. El fenotipado permite evaluar los rasgos genéticos en forma cuantitativa, en especial la relación entre ellos y el rendimiento o la tolerancia al estrés. El fenotipado es una herramienta que, mediante la evaluación de las características y propiedades fenotípicas de las plantas, permite predecir sus interacciones con el medioambiente, como ser el estrés biótico y abiótico en ellas.

Según Araus & Cairns (2014) la actividad de buscar mejores variaciones se realiza desde antes, incluso, que hayamos elaborado el conocimiento sobre el ADN. Aunque a partir de la evolución del fitomejoramiento se observa que la obtención de mejores variaciones depende de la capacidad de fenotipar grandes cantidades de combinaciones genéticas de forma de identificar la que mejor progenie. Esto da nacimiento al fenotipado de alto rendimiento (HTPPs). En el trabajo citado los autores mencionan el gran avance que han tenido los sensores, análisis de imágenes, modelado, robótica, control remoto y minería de datos contribuyendo a la actividad (intersección con internet de las cosas). Pero mencionan como los dos factores limitantes: el manejo de la gran cantidad de datos y el uso de bioinformática para analizar los datos de fenotipado (intersecciones con Grandes Datos y Extracción del Conocimiento).

En el tema de fenotipado, siguiendo una línea más general que en el trabajo mencionado en el párrafo anterior, tenemos el estudio Yandun Narvaez, Reina, Torres-Torriti,

Kantor, & Cheein (2017) donde se aborda el fenotipado desde los puntos de vista: la evaluación morfológica (caracterización estructural y detección de plantas/frutos) y la evaluación fisiología. En lo que respecta a sensores para evaluación morfológica los autores mencionan dos categorías de sensores: sensores de recorrido (ultrasonido, tiempo de vuelo ToF y LiDAR) y sensores de visión artificial (cámaras de luz estructurada, cámaras color y visión estéreo). Mientras que para la evaluación fisiológica, los sensores utilizados son: cámaras térmicas y cámaras multiespectrales e hiperespectrales. En cuanto al procesamiento de los datos con el propósito de extraer información, los autores mencionan estrategias variadas: clustering y algoritmos de matching se suele utilizar en los datos de sensores de recorrido (para separar los objetos de interés del resto de la escena), en las imagines provenientes de cámaras color se utilizan técnicas de visión artificial, en cuanto a aprendizaje no supervisado un algoritmo muy utilizado es K-Means, también redes neuronales, y en aprendizaje supervisado mencionan algunos métodos basados en instancias (descriptos en la sección 4.3.2 del presente trabajo) como k-nearest neighbor y support vector machines. En el mencionado sondeo se nombra como desafíos a los que intentan contribuir esta técnica: administración de fertilizantes y pesticida (mediante la caracterización de la superficie cultivada), dispositivos de poda (decidiendo inteligentemente las partes a podar, gracias a la posibilidad de evaluar la salud de ellas), monitoreo de cultivos e mejoramiento genético (mediante fenotipado y genotipado se puede identificar las mejores combinaciones de recursos genéticos).

5.8 Detección y manejo de pestes

La aparición de enfermedades y las deficiencias en la nutrición de los cultivos son una de las causas de perdidas económicas más frecuentes. La detección y diagnostico efectivo de las mismas son uno de los temas en que la aplicación de IoT y Extracción del Conocimiento cobran cada vez mayor relevancia. Dado que este tipo de anomalías suele reflejarse en la forma y coloración del follaje de las plantas, puede realizarse evaluación mediante imágenes provenientes de sensores de percepción proximal, en general fusionados, mediante técnicas de Fusión de Sensores, con datos provenientes de sensores que reflejan el contexto (de las imágenes).

Este es el caso presentado en Kale & Sonavane (2019) donde se recolectan imágenes junto con datos de temperatura y humedad, se las normaliza y aplica un proceso de FSS (feature subset selection) basado en algoritmos genéricos y un proceso de clasificación basado en el algoritmo de red neuronal artificial (NN) de propagación hacia adelante del tipo Extreme Learning Machine (ELM), aunque referencia la aplicación de otros algoritmos como redes neuronales artificiales ANN, Support Vector Machine (SVM) y redes neuronales Fuzzy NN (todos mencionados en la sección de Extracción del Conocimiento del presente trabajo). En el trabajo mencionado, dicho esquema, se muestra aplicado a un sistema de soporte de decisión para la clasificación

de enfermedades y deficiencias de nutrición en viñedos, que puede operar en tiempo cercano al real.

5.9 Fertilización y Control de Pestes

La fertilización mediante dosificación variable comienza, en general, con la confección del mapa de aplicación (también llamado mapa de prescripción) que indica en donde y en que cantidades se realizarán las aplicaciones, y para la confección de este el primer paso es la delimitación de Zonas de Manejo. El tema se mencionó unos párrafos más arriba, pero se retoma aquí desde la perspectiva de la aplicación de fertilizantes. En Nawar, Corstanje, Halcro, Mulla, & Mouazen (2017) se realiza una revisión completa del tema, de la que tomamos algunos puntos que resumimos a continuación. En cuanto al muestreo, existen numerosas técnicas de interpolación espacial que permiten predecir los valores de los atributos en puntos que no han sido medidos, permitiendo obtener una suerte de mapa continuo. Los métodos de interpolación más comunes en los trabajos incluidos en la revisión son: interpolación con la distancia inversa ponderada (IDW) y el enfoque geoestadístico Kriging (regresión en procesos Gaussianos), el último es más utilizado a medida que aumenta la densidad del muestreo. En cuanto al proceso de delineación las técnicas de clasificación más utilizadas son: Fuzzy c-means no supervisado y k-means clustering. Para la confección del mapa de prescripción de nitrógeno y fósforo no existe un enfoque predominante (en los trabajos incluidos en la revisión que estamos comentando), suele utilizarse una predicción del rendimiento potencial (basada en la cantidad de agua disponible, las lluvias esperadas y la cantidad de agua necesaria para el cultivo especifico), rendimiento buscado (de acuerdo a las zonas de manejo delimitadas), y en general se contrasta la dosificación uniforme (aquella prescripta para el campo de forma homogénea, como si no se usara Dosificación Variable) usual con la dosificación variable (algunos trabajos fijan un porcentaje máximo de fluctuación sobre la dosificación uniforme).

5.10 Agricultura inteligente (Smart Farm)

En el caso de smart farming el alcance de las actividades relacionadas con la información se extienden más allá de la agricultura en sí hasta incluir todos los componentes de la cadena de valor. En ese escenario la creciente aplicación de Internet of Things hace que volúmenes muy grandes de información heterogénea sean producidos y estén disponibles en tiempo real (en cada punto de la cadena), es aquí donde encontramos la intersección entre Smart Farming, IoT y Grandes Datos. El trabajo Wolfert, Ge, Verdouw, & Bogaardt (2017) explora esta coyuntura y nombra los siguientes desafíos a abordar: 1) propiedad de los datos, seguridad y privacidad (abordar estos temas puede enlentecer la innovación, es importante encontrar mecanismos que permitan un equilibrio), 2) calidad

de la información, 3) procesamiento inteligente y análisis de datos (teniendo en cuenta las fuentes desestructuradas y heterogéneas que intervienen dada las características de la actividad), 4) generación de modelos de negocios atractivos para los proveedores de soluciones, pero que contribuyan en un reparto equitativo entre los participantes, 5) generar plataformas abiertas que aceleren la innovación y posicionen a los agricultores dentro de la cadena de valor.

6 Conclusiones

Alcanzar los objetivos de producción agrícola en términos de productividad, impacto ambiental, seguridad alimentaria y sostenibilidad, requieren de un mejor entendimiento de las complejas relaciones entre las variables con las que se trabaja. En las prácticas agrícolas, como vimos, podemos llegar a este entendimiento mediante la aplicación conjunta de las tecnologías de la información y comunicaciones (TICs en la agricultura, agroinformática), Internet de las cosas (IoT), grandes datos y extracción del conocimiento.

En este sentido, observamos que para encarar las temáticas del Agro de manera innovadora la aplicación de Tecnologías de la Información y la Comunicación actuales es fundamental, apoyadas con la aplicación de Internet de las Cosas que permite la recolección de datos y la actuación a una velocidad y un volumen nunca antes alcanzados, siendo requeridas aquí herramientas modernas de Grandes Datos para poder manejar estás características de forma satisfactoria y generar información valiosa para la toma de decisiones mediante la utilización de técnicas de Extracción del Conocimiento. Reflejando de esta forma la intersección de los cuatro temas vistos en este trabajo complementándose en un circuito virtuoso que permite un abordaje moderno y eficiente a las problemáticas del Agro.

Las combinaciones de las disciplinas mencionadas con cada una de las temáticas listadas en Agroinformática no han podido ser abarcadas en su totalidad en el presente trabajo debido su cantidad y variedad. Hemos seleccionado, sin embargo, las más representativas a nuestro humilde entender. Dejando abiertas para su posterior profundización algunas de las líneas planteadas.

Líneas futuras de investigación (algunas):

- En la confluencia entre Agro y IoT sería interesante profundizar en los aspectos relacionados con la seguridad: autenticación, confidencialidad y control de accesos. De manera similar, en la confluencia entre Agro y Grandes Datos tenemos las temáticas relativas a propiedad de los datos y temas relativos a privacidad y seguridad de los mismos. Para abordarlos, es importante encontrar un punto equilibrado, ya que el hacerlo de forma demasiado estricta puede enlentecer la velocidad de adopción de estas tecnologías.
- Interoperabilidad (sintáctica y semántica). Los datos recolectados se han ido complicando, comenzaron siendo mediciones de valores numéricos simples, a estar geolocalizados, espacio-temporalmente demarcados, incluyendo múltiples valores

de mediciones complejas, imágenes multiespectrales, video e, incluso, en ciertas aplicaciones datos totalmente desestructurados. La extracción del conocimiento y aplicación de grandes datos es desafiante en este contexto, aun cuando se han realizado ya grandes avances en la última década.

- Dados las características del ambiente de aplicación de IoT, donde se carece de conectividad (sobretodo en países en vías de desarrollo donde el acceso a infraestructura es dificultoso) o es muy limitada, es importante abordar el problema aplicando tecnologías que lo mitiguen (redes de sensores, auto organización y auto descubrimiento). Tecnologías como Edge Computing y Fog Computing pueden ser aplicadas de forma de mejorar la situación.
- Los proveedores de insumos y tecnología para el Agro, hace tiempo, vieron las posibilidades que las temáticas aquí tratadas tienen para potenciar sus negocios. Las soluciones que ellos ofrecen al productor agropecuario, son en general cerradas y poco compatibles entre distintos proveedores. Sin embargo, en los últimos tiempos, se ha dado cierta toma de conciencia por parte de los productores, que son conscientes del valor de sus datos y de las posibilidades que traen las plataformas abiertas, o simplemente la cooperación con sus pares (facilitada si las herramientas utilizadas pueden integrarse). Además, el sector público está ofreciendo plataformas de Datos abiertos para el agro cuya explotación e integración a las privadas es más que interesante.
- Las técnicas tradicionales de fusión de datos multi-origen son adecuadas cuando se trata de datos estructurados, no siendo este el caso del dominio del Agro. Entonces, la aplicación de Fusión de Sensores en este contexto es un desafío aún abierto. La utilización de aprendizaje profundo ha mostrado buenos resultados en cuanto a integración de este tipo de datos y manejo de problemas de representación semántica.
- Se observa una desproporción, entre la aplicación de sistemas de sensores y monitoreo (muchas veces por debido a una cuestión comercial, dado que muchas de las máquinas agrarias hoy en día ya vienen provistas con sensores y sistemas de almacenamiento de información de fábrica), versus sistemas con actuación y control a distancia.
- Para poder afrontar los desafíos del Agro, grandes cantidades de datos provenientes de múltiples y variados sensores deben ser ingeridos, entregados y procesado como un continuo en línea de manera de poder generar abstracciones y extraer información accionable (actionable information). Esto plantea la necesidad de generar soluciones de procesamiento de streams de datos específicas que cumplan, al menos, con los siguientes atributos: deben poder manejar la heterogeneidad de los datos de entrada, responder en forma dinámica y poder operar a la velocidad del flujo de datos de entrada. Para alcanzar este objetivo se aplica analítica de secuencias de datos en tiempo real (real-time streaming data analytic), dado que

se trata de tecnologías que están en constante movimiento aún, se genera aquí un interesante foco de trabajo.

- Los avances en Grandes Datos han permitido afrontar con éxito las características que los datos del agro proseen en cuanto a Volumen y Velocidad, aunque sigue siendo un desafío las cuestiones relativas a Variedad (3Vs de Laney). En este sentido sigue siendo requerido esfuerzo en temas de interoperatividad semántica, esta cuestión debe ser encarada más que desde lo tecnológico, desde un enfoque multidisciplinario que ayude a interconectar fuentes de datos heterogéneas pertenecientes a distintas disciplinas y propósitos diversos.
- Las máquinas agrícolas de operación autónoma (robots) están revolucionando el agro y son prometedoras en áreas de aplicación tales como cosecha autónoma, robots que eliminan malezas, identificación y erradicación automática de pestes, entre otras. En esta temática los avances en Extracción del Conocimiento, que siguen creciendo a gran velocidad, están aportando nuevos y novedosos modos de encarar estos desafíos.
- Los invernaderos, en tanto ambientes controlados, permiten un amplio uso de las temáticas analizadas en el presente documento. Temáticas relativas a IoT son muy aplicadas al monitoreo de invernaderos, Grandes Datos y Extracción del Conocimiento son invaluables en análisis, soporte de decisiones y automatización. Los ambientes controlados en que se ejecutan cultivos de Hidroponía, Acuaponía y cultivos verticales tipo AeroFarm permiten la aplicación de tecnologías de gestión inteligente de recursos y mejora de productividad pertenecientes a la confluencia entre las temáticas presentadas.
- Los sistemas de manejo de riesgos utilizan técnicas de análisis de datos para predecir posibles fallas en los cultivos, reduciendo la incertidumbre y permitiendo la intervención temprana. Estos sistemas gestionan los riesgos particulares para el cultivo especifico, en la zona de ejecución y según los parámetros particulares recogidos en ella. Temas como adaptación al cambio climático caen dentro de las cuestiones que estos sistemas intentan abordar. Se trata de una cuestión de gran trascendencia, que ha registrado grandes avance en los últimos años y posee aún varias líneas de investigación para abordar.

Hemos intentado, realizar una exploración de la intersección entre Agroinformática, Extracción del Conocimiento, Grandes Datos e Internet de las Cosas. Somos conscientes que estos temas son extensos y pese a los esfuerzos, mucho ha quedado afuera. Las décadas pasadas han demostrado grandes avances para el Agro, es nuestro deber continuar de forma de poder alcanzar los objetivos que la actividad tiene por delante.

7 Bibliografía

.

Abdmeziem, M. R., Tandjaoui, D., & Romdhani, I. (2016). Architecting the Internet of Things: State of the Art. En A. Koubaa & E. Shakshuki (Eds.), *Robots and Sensor Clouds* (Vol. 36, pp. 55-75). https://doi.org/10.1007/978-3-319-22168-7_3

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, *44*(1), 71-91. https://doi.org/10.1016/j.compag.2004.03.002

Adamchuk, V. I., & Tremblay, N. (2017). *New developments in proximal soil sensing*. 4.

Adamchuk, V. I., & Viscarra Rossel, R. A. (2011). Precision Agriculture: Proximal Soil Sensing. En J. Gliński, J. Horabik, & J. Lipiec (Eds.), *Encyclopedia of Agrophysics* (pp. 650-656). https://doi.org/10.1007/978-90-481-3585-1_126

Adamo, F., Andria, G., Attivissimo, F., & Giaquinto, N. (2004). An Acoustic Method for Soil Moisture Measurement. *IEEE Transactions on Instrumentation and Measurement*, *53*(4), 891-898. https://doi.org/10.1109/TIM.2004.831126

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., … others. (2012). Challenges and Opportunities with Big Data—A Community White Paper Developed by Leading Researchers across the United States. 2012. *Google Scholar*.

Anastasiou, E., Castrignanò, A., Arvanitis, K., & Fountas, S. (2019). A multi-source data fusion approach to assess spatial-temporal variability and delineate homogeneous zones: A use case in a table grape vineyard in Greece. *Science of the Total Environment*, *684*, 155-163. https://doi.org/10.1016/j.scitotenv.2019.05.324

Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C., & Yang, G.-Z. (2015). Big Data for Health. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1193-1208. https://doi.org/10.1109/JBHI.2015.2450362

Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, *19*(1), 52-61. https://doi.org/10.1016/j.tplants.2013.09.008

Arun, K., & Jabasheela, D. L. (2014). Big Data: Review, Classification and Analysis Survey. *International Journal of Innovative Research in Information Security*, *1*(3), 7.

Baesens, B. (2014). *Analytics in a Big Data World*. Wiley.

Blum, A., Flammer, I., Friedli, T., & Germann, P. (2004). Acoustic Tomography Applied to Water Flow in Unsaturated Soils. *Vadose Zone Journal*, *3*(1), 288. https://doi.org/10.2136/vzj2004.2880

Bongiovanni, R. (2006). *Agricultura de precisión integrando conocimientos para una agricultura moderna y sustentable.* Montevideo: Procisur/IICA.

Bradley, B. A. (2014). Remote detection of invasive plants: A review of spectral, textural and phenological approaches. *Biological Invasions*, *16*(7), 1411-1425. https://doi.org/10.1007/s10530-013-0578-9

Brase, T., Shannon, D. K., Clay, D. E., & Kitchen, N. R. (2018). Basics of a Geographic Information System. En *ACSESS Publications*. https://doi.org/10.2134/precisionagbasics.2016.0119

Clay, D. E., Hatfield, G., & Clay, S. A. (2017a). An Introduction to Experimental Design and Models. En *Practical Mathematics for Precision Farming* (pp. 52-64). https://doi.org/10.2134/practicalmath2017.0104

Clay, D. E., Kitchen, N. R., Byamukama, E., & Bruggeman, S. (2017b). Calculations Supporting Management Zones. En *Practical Mathematics for Precision Farming* (pp. 122-135). https://doi.org/10.2134/practicalmath2017.0024

Corwin, D. L., & Lesch, S. M. (2003). Application of Soil Electrical Conductivity to Precision Agriculture: Theory, Principles, and Guidelines. *AGRONOMY JOURNAL*, *95*, 17.

Damian, J. M., Pias, O. H. de C., Cherubin, M. R., Fonseca, A. Z. da, Fornari, E. Z., & Santi, A. L. (2020). Applying the NDVI from satellite images in delimiting management zones for annual crops. *Scientia Agricola*, *77*(1). https://doi.org/10.1590/1678-992x-2018-0055

Dankhara, F., Patel, K., & Doshi, N. (2019). Analysis of robust weed detection techniques based on the Internet of Things (IoT). *Procedia Computer Science*, *160*, 696-701. https://doi.org/10.1016/j.procs.2019.11.025

Data Never Sleeps 6 | Domo. (2018). Recuperado 10 de noviembre de 2018, de https://www.domo.com/learn/data-never-sleeps-6

Davis, J. L., & Annan, A. P. (1989). GROUND-PENETRATING RADAR FOR HIGH-RESOLUTION MAPPING OF SOIL AND ROCK STRATIGRAPHY1. *Geophysical Prospecting*, *37*(5), 531-551. https://doi.org/10.1111/j.1365-2478.1989.tb02221.x

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122-135. https://doi.org/10.1108/LR-06-2015-0061

Deng, L. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, *7*(3-4), 197-387. https://doi.org/10.1561/2000000039

Doerge, T. A. (1999). Management Zone Concepts (SSMG-2). *SSMG-2. Site*, 4.

Ferguson, R., Rundquist, D., Shannon, D. K., Clay, D. E., & Kitchen, N. R. (2018). Remote Sensing for Site-Specific Crop Management. En *ACSESS Publications*. https://doi.org/10.2134/precisionagbasics.2016.0092

Fulton, J., Hawkins, E., Taylor, R., Franzen, A., Shannon, D. K., Clay, D. E., & Kitchen, N. R. (2018). Yield Monitoring and Mapping. En *ACSESS Publications*. https://doi.org/10.2134/precisionagbasics.2016.0089

Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos. *IDC Iview*, *1142*(2011), 12.

Gartner Says 8.4 Billion Connected. (2017). Recuperado 12 de julio de 2018, de https://www.gartner.com/newsroom/id/3598917

Google Flu Trends. (2014). Recuperado 22 de noviembre de 2018, de https://www.google.org/flutrends/about/

Group, S. M. A., Engineering (NITIE), N. I. of I., Lake, V., Mumbai, Group, I. R. A., Engineering (NITIE), N. I. of I., … India. (2015). Internet of Things (IoT): A Literature Review. *Journal of Computer and Communications*, *03*, 164. https://doi.org/10.4236/jcc.2015.35021

Guadalupe Ramos Hernández, J., Gracia-Sánchez, J., Patricia Rodríguez-Martínez, T., & Adalberto Zuñiga-Morales, J. (2019). Correlation between TDR and FDR Soil Moisture Measurements at Different Scales to Establish Water Availability at the South of the Yucatan Peninsula. En G. Civeira (Ed.), *Soil Moisture*. https://doi.org/10.5772/intechopen.81477

Hardy, Q. (2012, marzo 15). Better Economic Forecasts, From the Cloud. Recuperado 25 de noviembre de 2018, de Bits Blog website: https://bits.blogs.nytimes.com/2012/03/15/better-forecasts-from-the-cloud/

Helbing, D., & Balietti, S. (2011). From Social Data Mining to Forecasting Socio-Economic Crises. *The European Physical Journal Special Topics*, *195*(1), 3.

Hemmat, A., & Adamchuk, V. I. (2008). Sensor systems for measuring soil compaction: Review and analysis. *Computers and Electronics in Agriculture*, *63*(2), 89-103. https://doi.org/10.1016/j.compag.2008.03.001

Hiremath, S., Yang, G., & Mankodiya, K. (2014). Wearable Internet of Things: Concept, Architectural Components and Promises for Person-Centered Healthcare. *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies"*. Presentado en 4th International Conference on Wireless Mobile Communication and

Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies". https://doi.org/10.4108/icst.mobihealth.2014.257440

Holland, K. H., Lamb, D. W., & Schepers, J. S. (2012). Radiometry of Proximal Active Optical Sensors (AOS) for Agricultural Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(6), 1793-1802. https://doi.org/10.1109/JSTARS.2012.2198049

Holzinger, A., Kieseberg, P., Weippl, E., & Tjoa, A. M. (2018). Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. En A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (Vol. 11015, pp. 1-8). https://doi.org/10.1007/978-3-319-99740-7_1

Huang, Y.-b., Lan, Y.-b., Hoffmann, W. C., & Lacey, R. E. (2007). Multisensor data fusion for high quality data analysis and processing in measurement and instrumentation. *Journal of Bionic Engineering*, *4*(1), 53-62. https://doi.org/10.1016/S1672-6529(07)60013-4

Jayasekara, P. K., & Abu, K. S. (2018). Top Fifty Highly Cited Publications on the Internet of Things. *Journal of the University Librarians Association of Sri Lanka*, (2), 17.

Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., & Wang, J. (2018). A review of data assimilation of remote sensing and crop models. *European Journal of Agronomy*, *92*, 141-152. https://doi.org/10.1016/j.eja.2017.11.002

Kale, A. P., & Sonavane, S. P. (2019). IoT based Smart Farming : Feature subset selection for optimized high-dimensional data using improved GA based approach for ELM. *Computers and Electronics in Agriculture*, *161*, 225-232. https://doi.org/10.1016/j.compag.2018.04.027

Kwak, H., Blackburn, J., & Han, S. (2015). Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 3739-3748. https://doi.org/10.1145/2702123.2702529

Laney, D. (2001). *3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Report.*

Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74-81. https://doi.org/10.1016/j.bdr.2015.01.003

Leroux, C., Jones, H., Taylor, J., Clenet, A., & Tisseyre, B. (2018). A zone-based approach for processing and interpreting variability in multi-temporal yield data sets. *Computers and Electronics in Agriculture*, *148*, 299-308. https://doi.org/10.1016/j.compag.2018.03.029

Liakos, K., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Liggins, M. E., Hall, D. L., & Llinas, J. (Eds.). (2009). *Handbook of multisensor data fusion: Theory and practice* (2. ed). Boca Raton, Fla.: CRC Press.

Lin, J., Wang, M., Zhang, M., Zhang, Y., & Chen, L. (2008). Electrochemical Sensors for Soil Nutrient Detection: Opportunity and Challenge. En D. Li (Ed.), *Computer And Computing Technologies In Agriculture, Volume II* (Vol. 259, pp. 1349-1353). https://doi.org/10.1007/978-0-387-77253-0_77

Lu, Z., & Sabatier, J. M. (2009). Effects of Soil Water Potential and Moisture Content on Sound Speed. *Soil Science Society of America Journal*, *73*(5), 1614. https://doi.org/10.2136/sssaj2008.0073

Maestrini, B., & Basso, B. (2018). Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-32779-3

Maglogiannis, I., Karpouzis, K., & Wallace, M. (2007). *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Recuperado de http://public.eblib.com/choice/publicfullrecord.aspx?p=329926

Mahmood, H., Hoogmoed, W., & van Henten, E. (2013). Proximal Gamma-Ray Spectroscopy to Predict Soil Properties Using Windows and Full-Spectrum Analysis Methods. *Sensors*, *13*(12), 16263-16280. https://doi.org/10.3390/s131216263

MARTELLOTTO, E., MENDEZ, A. A., VON MARTINI, A., & BIANCHINI, A. (2007). *Percepción Remota. Proyecto Agricultura de Precisión, INTA Manfredi*. Recuperado de https://inta.gob.ar/documentos/percepcion-remota

Martinelli, F., Scalenghe, R., Davino, S., Panno, S., Scuderi, G., Ruisi, P., … Dandekar, A. M. (2015). Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development*, *35*(1), 1-25. https://doi.org/10.1007/s13593-014-0246-1

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784-2817. https://doi.org/10.1080/01431161.2018.1433343

Miao, Y., Mulla, D. J., & Robert, P. C. (2018). An integrated approach to site-specific management zone delineation. *Frontiers of Agricultural Science and Engineering*, *0*(0), 0. https://doi.org/10.15302/J-FASE-2018230

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Morimoto, E. (2018). What is cyber-physical system driven agriculture? - Redesign of big data for outstanding farmer management -. *2018 Detroit, Michigan July 29 - August 1, 2018*. Presentado en 2018 Detroit, Michigan July 29 - August 1, 2018. https://doi.org/10.13031/aim.201800486

Muangprathub, J., Boonnam, N., Kajornkasirat, S., Lekbangpong, N., Wanichsombat, A., & Nillaor, P. (2019). IoT and agriculture data analysis for smart farm. *Computers and Electronics in Agriculture*, *156*, 467-474. https://doi.org/10.1016/j.compag.2018.12.011

Nawar, S., Corstanje, R., Halcro, G., Mulla, D., & Mouazen, A. M. (2017). Delineation of Soil Management Zones for Variable-Rate Fertilization. En *Advances in Agronomy* (Vol. 143, pp. 175-245). https://doi.org/10.1016/bs.agron.2017.01.003

Noborio, K. (2001). Measurement of soil water content and electrical conductivity by time domain reflectometry: A review. *Computers and Electronics in Agriculture*, *31*(3), 213-237. https://doi.org/10.1016/S0168-1699(00)00184-8

Olagunju, A. O., & Khan, F. (2016). Challenges of Interdisciplinary IoT Curriculum. *SIGITE 2016 - Proceedings of the 17th Annual Conference on Information Technology Education*, 110. https://doi.org/10.1145/2978192.2978200

Patel, K. K., Patel, S. M., & Professor, P. S. A. (2016). Internet of Things-IOT: Definition, Characteristics, Architecture, Enabling Technologies, Application & Future Challenges. *Int. J. Eng. Sci. Comput*, *6*(5).

Pepó, P. (2013). *Alternative Crop Production Strategies*. Recuperado de https://www.tankonyvtar.hu/en/tartalom/tamop412A/2011_0009_Pepo_Peter-Alternative_Crop_Production_Strategies/adatok.html

Pundir, Y., Sharma, N., & Singh, D. Y. (2016). *Internet of Things (IoT) : Challenges and Future Directions*. *5*(3), 5.

Samouëlian, A., Cousin, I., Tabbagh, A., Bruand, A., & Richard, G. (2005). Electrical resistivity survey in soil science: A review. *Soil and Tillage Research*, *83*(2), 173-193. https://doi.org/10.1016/j.still.2004.10.004

Schowengerdt, R. A. (2007). *Remote sensing, models, and methods for image processing* (3rd ed). Burlington, MA: Academic Press.

Shanmugapriya, P., Rathika, S., Ramesh, T., & Janaki, P. (2019). Applications of Remote Sensing in Agriculture - A Review. *International Journal of Current Microbiology and Applied Sciences*, *8*(01), 2270-2283. https://doi.org/10.20546/ijcmas.2019.801.238

Shearer, S. A., Fulton, J. P., McNeill, S. G., Higgins, S. F., & Mueller, T. G. (1999). *Pa-1: Elements of Precision Agriculture: Basics of Yield Monitor Installation and Operation*. 10.

Sheets, K. R., & Hendrickx, J. M. H. (1995). Noninvasive Soil Water Content Measurement Using Electromagnetic Induction. *Water Resources Research*, *31*(10), 2401-2409. https://doi.org/10.1029/95WR01949

Shimotashiro, T., Inanaga, S., Sugimoto, Y., Matsuura, A., & Ashimori, M. (1998). Non-destructive Method for Root Elongation Measurement in Soil Using Acoustic Emission Sensors. *Plant Production Science*, *1*(1), 25-29. https://doi.org/10.1626/pps.1.25

Sudduth, K. A., Drummond, S. T., & Kitchen, N. R. (2001). Accuracy issues in electromagnetic induction sensing of soil electrical conductivity for precision agriculture. *Computers and Electronics in Agriculture*, *31*(3), 239-264. https://doi.org/10.1016/S0168-1699(00)00185-X

Talavera, J. M., Tobón, L. E., Gómez, J. A., Culman, M. A., Aranda, J. M., Parra, D. T., … Garreta, L. E. (2017). Review of IoT Applications in Agro-Industrial and Environmental Fields. *Computers and Electronics in Agriculture*, *142*, 283-297. https://doi.org/10.1016/j.compag.2017.09.015

Team, O. R. (2011). Big Data Now: Current Perspectives from OReilly Radar. *OReilly Media*.

T. E. Grift, M. Z. Tekeste, & R. L. Raper. (2005). ACOUSTIC COMPACTION LAYER DETECTION. *Transactions of the ASAE*, *48*(5), 1723-1730. https://doi.org/10.13031/2013.20006

That 'Internet of Things' Thing - 2009-06-22 - Page 1 - RFID Journal. (2009). Recuperado 11 de julio de 2018, de http://www.rfidjournal.com/articles/view?4986

Theodoridis, E., Mylonas, G., & Chatzigiannakis, I. (2013). Developing an IoT Smart City framework. *IISA 2013*, 1-6. https://doi.org/10.1109/IISA.2013.6623710

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, *161*, 272-279. https://doi.org/10.1016/j.compag.2018.03.032

Tr, T. (2009). *Dimensionality Reduction: A Comparative Review*. 36.

Vermesan, O., & Friess, P. (2014). *Internet of Things-from Research and Innovation to Market Deployment* (Vol. 29). River publishers Aalborg.

Viacheslav I. Adamchuk, Raphael A. Viscarra Rossel, Kenneth A. Sudduth, & Peter Schulze Lammers. (2011). *Sensor Fusion for Precision Agriculture.* Recuperado de http://www.intechopen.com/articles/show/title/sensor-fusion-for-precision-agriculture

Viscarra Rossel, R. A., Adamchuk, V. I., Sudduth, K. A., McKenzie, N. J., & Lobsey, C. (2011). Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. En *Advances in Agronomy* (Vol. 113, pp. 243-291). https://doi.org/10.1016/B978-0-12-386473-4.00005-1

Vuran, M. C., Salam, A., Wong, R., & Irmak, S. (2018). Internet of underground things in precision agriculture: Architecture and technology aspects. *Ad Hoc Networks*, *81*, 160-173. https://doi.org/10.1016/j.adhoc.2018.07.017

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, *153*, 69-80. https://doi.org/10.1016/j.agsy.2017.01.023

Xian, K. (2017). Internet of Things Online Monitoring System Based on Cloud Computing. *International Journal of Online Engineering (iJOE)*, *13*(09), 123. https://doi.org/10.3991/ijoe.v13i09.7591

Xue, J., & Su, B. (2017). Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors*, *2017*, 1-17. https://doi.org/10.1155/2017/1353691

Yandun Narvaez, F., Reina, G., Torres-Torriti, M., Kantor, G., & Cheein, F. A. (2017). A Survey of Ranging and Imaging Techniques for Precision Agriculture Phenotyping. *IEEE/ASME Transactions on Mechatronics*, *22*(6), 2428-2439. https://doi.org/10.1109/TMECH.2017.2760866