ISBN 978 -987-633-574-4

Identificación de Variedad Contextual en Modelado de Sistemas Big Data

Líam Osycka[1] *, Agustina Buccella[1], and Alejandra Cechich[1]

GIISCO Research Group
Departamento de Ingeniería de Sistemas - Facultad de Informática
Universidad Nacional del Comahue
Neuquen, Argentina
liam.osycka, agustina.buccella, alejandra.cechich@fi.uncoma.edu.ar

Resumen La propiedad de los sistemas Big Data con respecto a diversidad de los datos se denomina Variedad y su análisis permite identificar distintos tipos; por ejemplo, la variedad estructural denota la variedad en formatos y tipos de datos, clasificándolos como estructurados, semi-estructurados y no estructurados. En particular, el agregado de información de contexto (o dominio) permite análisis más complejos en la variedad, llevando a una nueva fase de investigación en su modelado que incluye la posibilidad de reuso. En este artículo, presentamos una propuesta para modelar sistemas Big Data para/con reuso teniendo en cuenta variaciones en el contexto que surgen del análisis de datos existentes para un problema dado. La propuesta incluye un caso de estudio a modo de prueba de conceptos.

Keywords: Modelado de Sistemas Big Data, Reusabilidad, Variedad, Líneas de Productos Software

1. Introducción

La propiedad de los sistemas Big Data (SBD) [2] con respecto a diversidad de los datos se denomina Variedad; y en [1] se clasifica en una taxonomía que divide el análisis de variedad en cuatro casos de diversidad: estructural, de las fuentes, de contenido y de procesamiento. Por ejemplo, la *diversidad estructural* denota la variedad en formatos y tipos de datos, clasificándolos como estructurados, semi-estructurados y no estructurados; la *diversidad de las fuentes* se clasifica en tres grupos - datos generados por humanos, generados por máquinas o mediados por procesos; la *diversidad de contenido* aborda diferentes tipos de soporte; y la *diversidad de procesamiento* enfoca en las distintas necesidades de procesamiento algorítmico.

La variedad en los datos también ha sido considerada desde el punto de vista de incorporación de semántica al proceso de modelado de arquitecturas en SBDs,

* Este trabajo esta parcialmente soportado por el Proyecto Desarrollo de Software basado en Reuso Parte II

ISBN 978 -987-633-574-4

e incluso ha sido relacionada con diversas propiedades como interoperabilidad, seguridad, reusabilidad, etc. En SBDs, la reusabilidad ha sido abordada también desde diversos ángulos. Por ejemplo, en [9] se discuten conceptos de reusabilidad en el contexto de analítica de datos distinguiendo entre uso y reuso del dato. En otro sentido, incorporando la detección de aspectos comunes y variables a modo de familia de sistemas, en [7] se propone una arquitectura de referencia acotada por medio de casos de uso. De esos casos, se identifican requerimientos relevantes al SBD, incluyendo categorías, como tipos de datos, transformaciones, visualizaciones, etc. Luego, la arquitectura se organiza como una colección de módulos que descomponen la solución en funciones o capacidades para un conjunto de aspectos. En este contexto, y respondiendo a la pregunta de investigación:

RQ: *¿Cómo puede identificarse la variedad de la información de dominio de manera de incorporar reusabilidad en el desarrollo de SBDs?*

este artículo extiende la arquitectura para la construcción de SBDs presentada en [5], incorporando variedad a modo de líneas de productos [10]. A diferencia de la propuesta en [7], que descompone la arquitectura en módulos asociados a intereses guiados por soluciones existentes, nuestra propuesta toma como partida una estructura de etapas asociadas al desarrollo de SBDs, instanciada en artefactos software producidos durante esas etapas, e incorpora el modelado de variedad de contexto de manera similar a líneas de productos. La propuesta se ejemplifica mediante un caso de estudio en el dominio hidrológico a modo de prueba de conceptos.

El artículo se organiza de la siguiente manera. En la sección 2 se introduce nuestro enfoque en el sentido bottom-up y luego, la sección 3 presenta el caso de estudio. Finalmente, se abordan conclusiones y trabajos futuros.

2. Enfoque bottom-up para identificar características de contexto variantes

A partir de la pregunta de investigación que hemos definido en la introducción (RQ), en la Figura 1 mostramos la visión global del enfoque bottom-up de nuestra propuesta, es decir, la identificación de variedad a partir de los datos.

En principio, al centrarnos en SBDs, el primer elemento a considerar es el proceso de desarrollo, donde las etapas básicas pueden resumirse en [6] (círculos centrales de la Figura): (1) *Adquisición de datos*, que consiste en extraer los datos desde las fuentes, agregando un proceso de carga y filtrado para que los datos sean adecuados a su posterior procesamiento; (2) *Transformación y Mejora*, que consiste en estructurar el formato de los datos, realizar la limpieza de los mismos y eventualmente, también su integración; (3) *Análisis*, que contiene las funcionalidades que permiten derivar conocimiento a partir de los datos, enfocando en análisis descriptivo, predictivo y/o prescriptivo; y (4) *Visualización*, que es el punto de acceso a los resultados del proceso.

A su vez, en la Figura podemos observar que nuestro enfoque parte de un proceso bottom-up. Esto significa que las características variantes serán identificadas a partir de un proceso de análisis de datos, donde las variedades serán

ISBN 978 -987-633-574-4

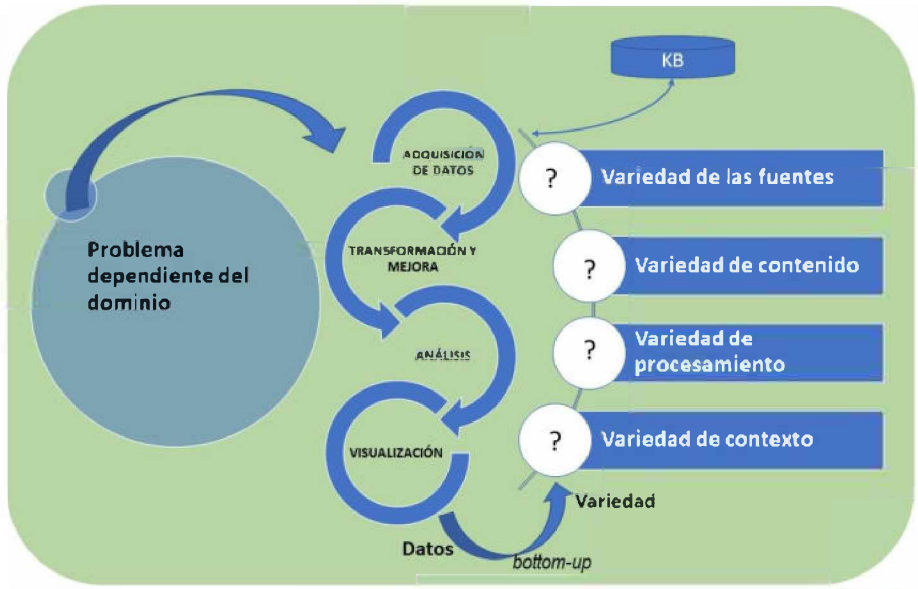

Figura 1: Visión global del enfoque bottom-up: desde los datos al dominio

inferencias a ser corroboradas por expertos de dominio. En contraposición, un enfoque top-down iniciaría con análisis de información de dominio, a ser corroborada por evidencia en las bases de datos. En principio, nuestra propuesta engloba ambos enfoques; sin embargo, en este artículo sólo presentaremos el sentido bottom-up para una mejor comprensión.

Así, partimos desde la *definición de un problema dependiente del dominio* e intentamos detectar características variantes dentro de cada una de las etapas del proceso de análisis de datos. Por ejemplo, la *variedad de las fuentes*, dentro de la etapa de *adquisición de datos* intentará detectar diferencias en las fuentes en cuanto a los cambios de su estructura, datos, formas de adquirirlas, etc. La *variedad de contenido* detectará cambios en las formas en que los datos deben ser transformados y procesados de acuerdo a los objetivos planteados, sobre todo considerando cambios o evoluciones de las fuentes. La *variedad de procesamiento* permitirá detectar variantes en cuanto a las técnicas de análisis posibles de utilizar y por último la *variedad de contexto* permitirá detectar variaciones del dominio que condicionen o cambien los resultados de los análisis.

Nuestro enfoque propone documentar las variedades encontradas en un dominio determinado de forma tal de almacenarlas en una base de conocimiento (KB) para que puedan ser reusadas en las mismas situaciones pero en contextos (dominios o casos) diferentes [5].

En trabajos previos, hemos presentado una propuesta de diseño de Líneas de Productos Software (LPSs) dirigida por funcionalidades, donde cada fun-

ISBN 978 -987-633-574-4

cionalidad se documenta a través de una hoja de datos funcional (*datasheet*), representando el conjunto de servicios comunes[1] y variantes [3,4].

Para el caso de reusabilidad en SBDs, la Figura 2 muestra la hoja de datos funcional definida para reusar modelos de análisis y detectar *variedades de contexto*. En la primera funcionalidad, *Buscar modelo a utilizar*, vemos las acciones necesarias para definir el objetivo del proceso de análisis y recuperar la técnica que se desea aplicar. Allí podemos observar el modelo de variabilidad asociado que posee el punto de variación *técnicas de análisis* con una variabilidad alternativa, es decir, se puede instanciar sólo una de las variantes definidas, ya sea una *red neuronal*, un análisis usando *k-means*, etc. Una vez seleccionada la técnica, en la siguiente datasheet podemos ver la funcionalidad que se despliega por haber elegido el modelo de red neuronal. Aquí debemos buscar los modelos existentes en la KB y determinar si podemos reusarlos, o si existen variedades de contexto que requieren la creación de nuevos modelos. Así, en la Figura vemos asociados dos modelos de variabilidad con puntos de variación opcionales (para el caso de encontrar similitudes entre el modelo a aplicar y los existentes) y puntos de variación alternativos para almacenar un nuevo modelo (cuando el mismo debe ser creado[2]) o reusar alguno existente.

3. Aplicación del enfoque bottom-up al caso de estudio

La calidad del agua es medida por los cambios en los parámetros químicos, ecológicos y espaciales, de los cuales además de estudiar sus valores, hay que ver sus interdependencias. Entre esos parámetros se encuentran la *concentración de pH* (una medida usada para testear acidez), el *Oxigeno Disuelto*, la *Temperatura del Agua*, etc. [8]. La Figura 3 muestra el enfoque bottom-up definido en la sección anterior, pero instanciado a nuestro caso de estudio. A la izquierda de la Figura puede verse un problema de dominio dado, en el cual se debe detectar variedad contextual (*Causas de variación de la temperatura en dos localizaciones de un curso de agua*). En nuestro caso de estudio y a modo de prueba de conceptos, estableceremos estabilidad (no variación) en las fuentes, contenido y procesamiento, intentando identificar variaciones contextuales en el dominio de estudio. La variedad de contexto a identificar consiste en relacionar las inferencias realizadas a partir de los datos con información del dominio (ubicaciones geográficas del curso de agua en **L1** y **L2** que sean caracterizadas en términos de variables comunes y variantes).

Adquisición de los datos

Para el caso de estudio, seleccionamos un dataset que contiene muestras de cuerpos de agua en King County, Washington, Estados Unidos. La cantidad de

[1] Los servicios comunes son aquellos que son parte de todos los productos derivados de la LPS

[2] La restricción «require» entre las variantes determina que si se debe crear un nuevo modelo, debe a su vez guardarse en la KB con la documentación correspondiente

ISBN 978 -987-633-574-4

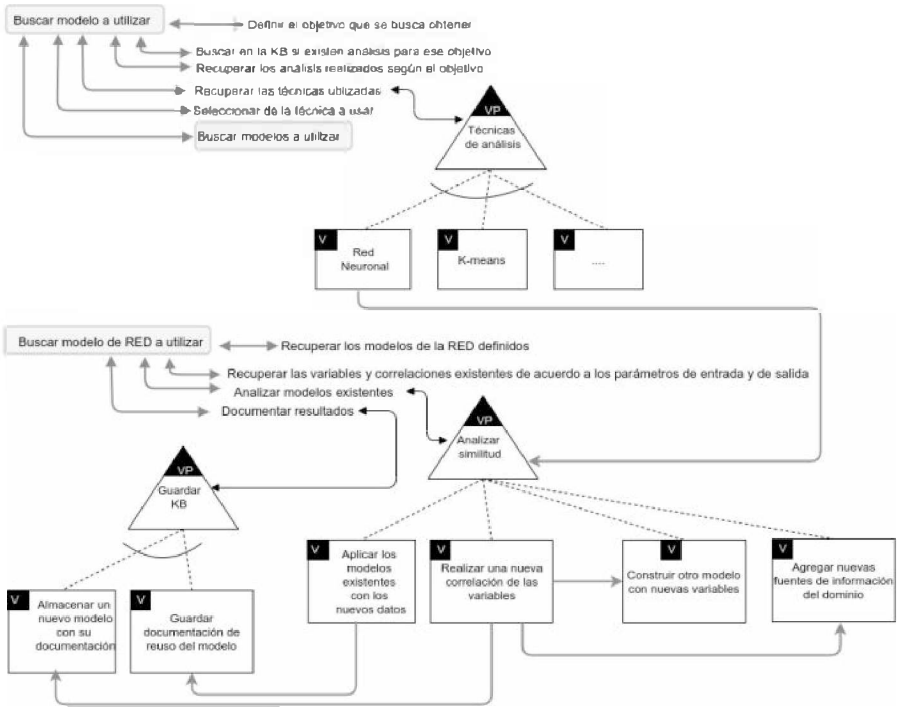

Figura 2: Datasheets para reusar modelos de análisis y detectar *variedades de contexto*

tuplas (1.589.362) y 25 columnas de variables lo hicieron adecuado para nuestra prueba de conceptos. Entre esas variables, se registran la identificación de la muestra, su fecha de recolección, tipo de sitio donde se recolectó (ej. ríos, lagos), punto donde se extraen las muestras (*locator*), el tipo de parámetro correspondiente a la muestra, etc. En particular, esta última variable identifica si el parámetro corresponde a pH, temperatura, turbidez, total de fósforo, coliformes fecales, oxígeno disuelto, conductividad, amoníaco de nitrógeno, densidad o clorofila.

Este dataset es entrada para el análisis de dos localizaciones - **L1** y **L2** (Figura 3), con lo que se mantiene la igualdad de la fuentes y por lo tanto, no hay variedad que se incorpore en esta etapa.

Transformación y mejora

Después de analizar los tipos de datos suministrados y de seleccionar los relevantes al problema abordado, procedimos a la transformación en columnas, agregando aquellas correspondientes a información geográfica de cada muestra en cada localización y seleccionando dos de ellos (**L1** y **L2**) para nuestro análisis. Para identificar relaciones entre los parámetros y la variable *temperatura*, foco

ISBN 978 -987-633-574-4

Figura 3: Enfoque bottom-up aplicado al caso de estudio

del problema, realizamos un análisis de correlación de Pearson determinando el grado de intensidad y dirección de las relaciones lineales entre cada par de variables. Este análisis se aplicó primero en **L1**, donde pudimos observar que las variables más relacionadas con el objetivo eran *'Nitrógeno Total'*, *'Alcalinidad Total'*, *'Oxígeno Disuelto del Suelo'*, *'Oxígeno Disuelto'* y *'Conductividad del Suelo'*. En la Figura 4, del lado izquierdo, podemos observar gráficamente estas correlaciones. Al tomar sólo variables significativas para la temperatura en **L1** y graficándo nuevamente las correlaciones, pero ahora con los datos en **L2**, observamos que existen diferencias en la intensidad de las relaciones. Esto llevó a que repliquemos el análisis para obtener las variables más significativas en esta segunda localización.

En la Tabla 1 se observan las relaciones de cada variable con temperatura para **L1** y **L2**. Puede verse que algunas variables, como el *'pH del suelo'*, tienen una fuerte relación con temperatura en **L2** mientras que en **L1** no sucede lo mismo (*'Oxígeno Disuelto'* muestra mayor impacto). Resumiendo en las dos columnas intermedias de la Tabla 1, podemos ver la diferencia que hay entre los valores de ambas localizaciones para cada parámetro y el orden de los mismos de acuerdo a su incidencia.

Análisis

En la Figura 5 mostramos la instanciación del datasheet para este caso. De acuerdo al problema enunciado en la Figura 3 y considerando inexistencia de antecedentes en la base de conocimientos (KB), para este caso se decidió la utili-

ISBN 978 -987-633-574-4

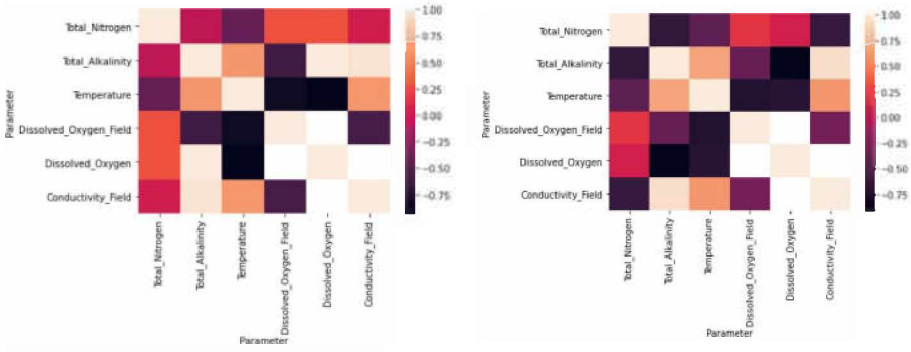

Figura 4: Correlación gráfica de variables seleccionadas en L1 (izquierda) y L2 (derecha)

Parámetro	L1	L2	Parámetro
Oxígeno Disuelto	-0.928462	-0.703626	pH, del Suelo
Oxígeno Disuelto, del Suelo	-0.857255	-0.74100	Silica
Alcalinidad Total	0.591276	0.654430	Nitrito + Nitrato Nitrógeno
Conductividad, del Suelo	0.582982	0.594806	pH
Nitrógeno Total	-0.420579	-0.439275	Conductividad
Ortofosfato de Fósforo	0.320313	0.110690	Nitrógeno Amoniacal
Nitrógeno Amoniacal	0.257200	0.019802	Fósforo Total
Nitrito + Nitrato Nitrógeno	-0.246479	-0.622934	Oxígeno Disuelto
Fósforo Total	0.242684	-0.012616	Ortofosfato de Fósforo
Conductividad	0.208746	0.509993	Sólidos Suspendidos Totales
Sólidos Suspendidos Totales	-0.202858	-0.038395	Coliformes Fecales
Coliformes Fecales	0.047256	0.205362	Turbidez
Turbidez	-0.171750	-0.029406	E. coli
E. coli	0.111787	0.253594	Oxígeno Disuelto, del Suelo
Silica	0.073081	0.527815	Alcalinidad Total
pH	-0.050635	0.424880	Enterococo
Enterococo	0.042639	0.093202	Nitrógeno Total
pH, del Suelo	0.011134	0.557562	Conductividad, del Suelo

Cuadro 1: Correlación de parámetros con temperatura en L1 y L2

ISBN 978 -987-633-574-4

zación de redes neuronales con el objetivo de estimar el valor de la temperatura, en base a otros parámetros relacionados, para **L1** y **L2**.

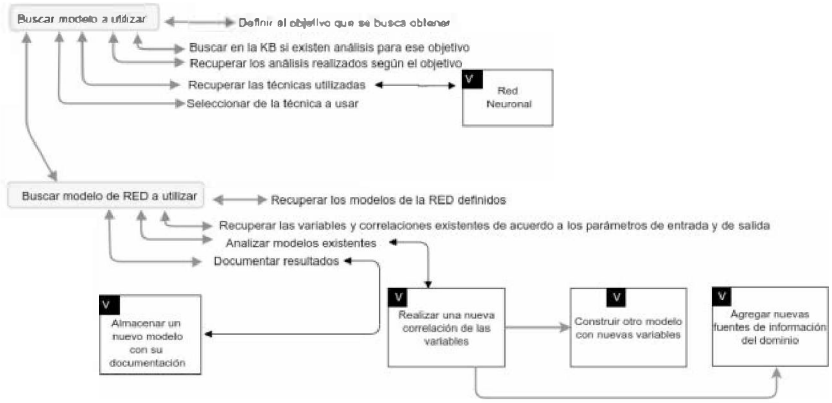

Figura 5: Modelo de Variabilidad instanciado para reflejar nuestro caso de estudio

Así, el primer modelo M_1 fue entrenado en **L1** y se procedió a comprobar la variación entre la temperatura real y estimada, calculando la diferencia entre las mismas. Para M_1 en L1 el valor promedio de las diferencias, calculadas para una serie de variables aleatorias, fue de **2.56841430**. Luego, el modelo M_1 se reutilizó en **L2** sin modificación en la configuración y sin realizar ningún ajuste de contexto (Figura 5 "Recuperar los modelos de la RED definidos"). El resultado de esta reutilización arrojó una diferencia promedio de **3.43957511** (Figura 6 (a)).

Como pudimos observar en la Tabla 1, los valores de las correlaciones de la temperatura en **L1** y **L2** muestran variaciones entre ellos (Figura 5 "Recuperar las variables y correlaciones ..."). Teniendo en cuenta esta variedad, se definió un nuevo modelo, M_2, con la misma arquitectura de M_1 (sin variedad de procesamiento) pero cambiando las variables de entrada. Mientras que en M_1 se utilizaron las variables *'Nitrógeno Total'*, *'Alcalinidad Total'*, *'Oxígeno Disuelto, del Suelo'*, *'Oxígeno Disuelto' y 'Conductividad del Suelo'*; para M_2 se utilizaron *'Alcalinidad Total'*, *'Oxígeno Disuelto, del Suelo'*, *'Oxígeno Disuelto'*, *'Nitrito + Nitrato de Nitrógeno'*, *'Conductividad del Suelo'*, *'pH del Suelo'* y *'Silica'*. Este nuevo modelo se ajusta a las características de contexto de **L2**, por lo que debería tener mejor desempeño en la predicción con respecto al valor obtenido al reusar M_1 (Figura 5 "Analizar los modelos existentes"). Efectivamente, en la Figura 6 (b) puede observarse la ejecución de M_2, mostrando una diferencia en promedio **2.303304169** que es menor que la obtenida con la ejecución anterior.

Ahora, el nuevo modelo documentado y almacenado en la KB podrá ser en un futuro elegido en otro contexto con características similares (Figura 5 "Documentar resultados").

ISBN 978 -987-633-574-4

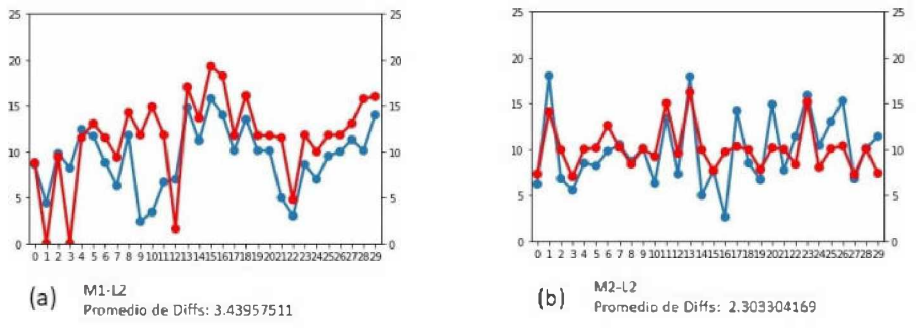

Figura 6: Diferencias entre valores reales y estimados de M_1 en L2 (a) y M_2 en L2 (b)

Visualización

La Figura 6 permite visualizar las ejecuciones realizadas durante el análisis. Sin embargo, existen posibilidades adicionales en el ejemplo presentado. La Figura 7 muestra las localizaciones de **L1** y **L2** en el espacio geográfico. Como parte del análisis, es interesante notar que, aunque ambas pertenecen al mismo curso de agua, su entorno es bastante diferente. Mientras que **L1** se encuentra en una zona relativamente urbanizada, **L2** es un área boscosa con poca intervención humana. Esas características, que se obtienen a simple vista, podrían complementar las condiciones de contexto que den explicación a las diferentes variaciones. Por ejemplo, en **L2** el *'pH del Suelo'* es mucho mayor que en **L1** - probablemente debido a la zona forestada y al tipo de vegetación. Este tipo de inferencias, debidamente contrastadas por expertos de dominio, podrían enriquecer la caracterización de cada localización y almacenarse para futuras identificaciones de áreas geográficas donde el mismo problema sea relevante.

4. Conclusiones y Trabajo Futuro

En este artículo, hemos presentado una propuesta para incorporar reusabilidad en el modelado de sistemas big data, identificando la manera en que la variedad de contexto puede impactar en actividades típicas como la transformación y el análisis de los datos.

Como hemos visto, la participación activa de expertos de dominio es fundamental para la definición del problema y para el análisis de los resultados. En ese sentido, actualmente, estamos definiendo casos de estudio con el acompañamiento de expertos del INTA (Instituto Nacional de Tecnología Agropecuaria) para extender la propuesta con su enfoque top-down y validar los resultados de ambos enfoques en SBDs para el análisis de la napa freática, en función de la variedad de fuentes acuíferas de diversas zonas geográficas.

ISBN 978 -987-633-574-4

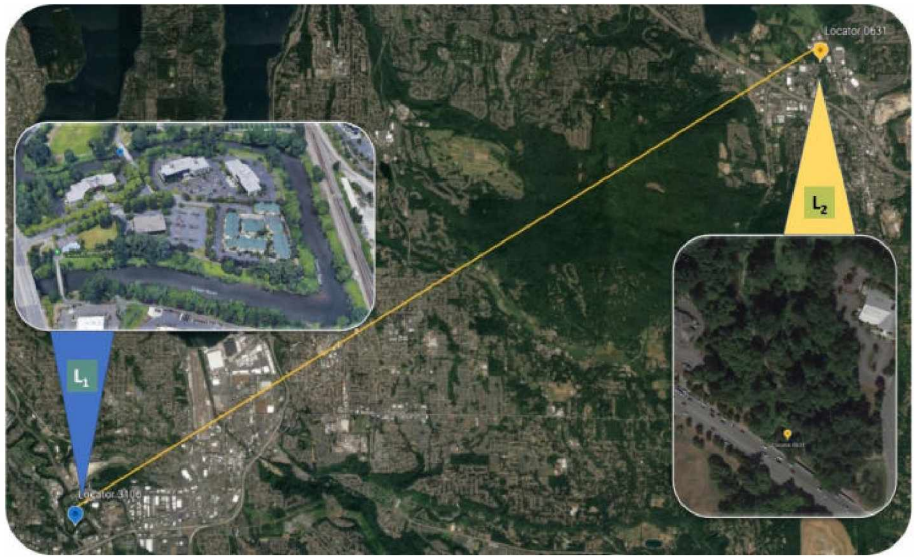

Figura 7: Localizaciones de L1 y L2 en el espacio geográfico

Referencias

1. Abawajy, J.: Comprehensive analysis of big data variety landscape. International Journal of Parallel, Emergent and Distributed Systems 30(1), 5–14 (2015)
2. Bahga, A., Madisetti, V.: Big Data Science & Analytics: A Hands-On Approach. VPT, 1st edn. (2016)
3. Buccella, A., Cechich, A., Arias, M., Pol'la, M., Doldan, S., Morsan, E.: Towards systematic software reuse of gis: Insights from a case study. Computers & Geosciences 54(0), 9 – 20 (2013)
4. Buccella, A., Cechich, A., Pol'la, M., Arias, M., Doldan, S., Morsan, E.: Marine ecology service reuse through taxonomy-oriented SPL development. Computers & Geosciences 73(0), 108 – 121 (2014)
5. Buccella, A., Luzuriaga, J., Cechich, A., Osycka, L., Paterno, F., Pol'la, M., Cruz, M., Martinez, R., Mazalu, R., Moyano, M.: Reusabilidad en el contexto de desarrollo de sistemas para big data. In: Actas del XXIII Workshop de Investigadores en Ciencias de la Computación, Chilecito, La Rioja. pp. 525–529 (2021)
6. Davoudian, A., L., M.: Big data systems: A software engineering perspective. ACM Computing Surveys 53(5) (2020)
7. Klein, J.: Reference architectures for big data systems, carnegie mellon university's software engineering institute blog. http://insights.sei.cmu.edu/blog/reference-architectures-for-big-data-systems/ (Accessed June 9, 2021) (2017)
8. Loucks, D.P., van Beek, E.: Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications. Springer (2017)
9. Pasquetto, I., Randles, B., Borgman, C.: On the reuse of scientific data. Data Science Journal 16(8) (201720)
10. Pohl, K., Böckle, G., Linden, F.J.v.d.: Software Product Line Engineering: Foundations, Principles and Techniques. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2005)