SEREA2017
Seminario Iberoamericano de Redes de Agua

XV Seminario Iberoamericano de Redes de Agua y Drenaje, SEREA2017

GEOSTATISTICAL MODELS FOR THE PREDICTION OF WATER SUPPLY NETWORK FAILURES IN BOGOTÁ, INTEGRATING MACHINE LEARNING ALGORITHMS

MODELOS GEOESTADÍSTICOS PARA LA PREDICCIÓN DE FALLOS DE UNA ZONA DE LA RED DE ABASTECIMIENTO DE AGUA DE BOGOTÁ, INTEGRANDO ALGORITMOS DE MACHINE LEARNING

C. Navarrete-López[11], D.S. Calderón Rivera[2], J.L. Díaz-Arévalo[3], M. Herrera[4], J. Izquierdo[5]

[1,2]Universidad Santo Tomas/Facultad de Ingeniería Ambiental, Bogotá, Colombia
[3]Universidad Santo Tomas/Facultad de Ingeniería Civil, Bogotá, Colombia
[4]University of Bath, Bath (UK)/ EDEn - Dept. of Architecture and Civil Eng, Bath, UK
[5]Universitat Politècnica de València/Fluing - Instituto de Matemática Multidisciplinar, Valencia, España

Abstract

Currently new strategies of spacial referencing, data analisis, and machine learning methods are integrated with Geographical Information Systems (GISs) to understand specific characteristics and water supply dinamics. This work explores the variables that can cause spacial failures and potential risk areas with application to a zone in the Bogotá water supply network. Machine learning algorithms are proposed to generate prediction models and potential failure maps. A sensitivity analysis was held to identify the model with the best fit for the estimation. This study will allow water supply decisions makers to focalize their efforts in the field.

Keywords: Water supply network failures; Machine Learning; GIS

Resumen

Actualmente se buscan nuevas estrategias y/o metodologías basadas en la integración de los Sistemas de Información Geográfica (SIGs) como forma de georeferenciacion espacial y visualización de las variables analizadas, junto con métodos de aprendizaje automático (Machine Learning) que permitan entender características puntuales, variables influyentes y dinámicas de los sistemas de abastecimiento de agua potable.En este trabajo se hace la identificación espacial de los fallos y zonas potenciales de riesgo que se presentan en una zona de la red de abastecimiento de Bogotá, explorando las variables que puedan tener mayor incidencia en los mismos. Se propone el uso de algoritmos de aprendizaje automático para la generación de modelos de predicción y la elaboración de mapas de fallos potenciales, identificando, a través de un análisis de sensibilidad, cuál de estos modelos presenta un mejor ajuste en la estimación. Este estudio permite a los gestores del abastecimiento una localización precisa y eficiente de los fallos en la red, apoyando el proceso de toma de decisiones.

Palabras clave: Fallos en redes de distribución; Machine Learning; SIG

1 Autor de correspondencia: Tel.: +57-1-5878797 Ext. 1571
Correo electrónico: claudianavarrete@usantotomas.edu.co

1. Introducción

Es necesario identificar y reconocer nuevas herramientas, que las técnicas de Machine Learning ofrecen, que sirvan como apoyo al estudio de las complejas interacciones entre factores relacionados con las prácticas en la operación y gestión de redes de abastecimiento urbano y las estructuras sociales, procesos físicos, sistemas de ingeniería y factores ambientales. Es importante profundizar en la identificación y el desarrollo de estas técnicas brindando alternativas novedosas a las técnicas actualmente existentes y otras que han sido recientemente exploradas por otras áreas de conocimiento, tales como la epidemiología [1] [2] y más recientemente la ingeniería.

En este último caso, a partir de los estudios de investigación y empíricos en epidemiología humana, se introduce un contexto para analizar las variaciones en el tiempo de rotura de tuberías, que pueden ayudar a las empresas de gestión del agua urbana, a entender mejor los fallos en las redes de tuberías e instaurar medidas para minimizar las interrupciones causadas por ellos. Se postula que en cualquier momento, una cohorte en el caso de tuberías metálicas que comprende el sistema de distribución de agua estará en un estado debilitado debido a la fatiga del metal y la corrosión. Esta cohorte frágil se vuelve vulnerable durante el curso de las operaciones normales y, en última instancia, se rompe debido al rápido aumento de longitudes de grieta inducidos por factores de estrés anormales [3].

Comprender los modelos para los sistemas de distribución de agua, beneficiaría a sus gestores y podría ayudar a prevenir y mitigar los brotes de roturas de tuberías de una manera comparable a las medidas de salud pública que han mitigado con éxito los efectos de las olas de calor en la población más vulnerable [3]

En los últimos años, con el rápido desarrollo de las tecnologías de la información y de bases de datos, los algoritmos de minería de datos se han utilizado para aplicaciones que van más allá de la tecnología de la información [4].

La minería de datos es un proceso por el cual se extrae información y conocimiento potencialmente útil a partir de bases de datos grandes, incompletas, aleatorias y difusas, cuya utilidad no se puede conocer de antemano [4].

A nivel de redes de distribución de agua se han obtenido modelos híbridos para la toma de decisiones sobre la renovación de tuberías, analizando los principales factores de influencia en el deterioro de las mismas, proponiendo un nuevo método de apoyo a la toma de decisiones, y priorizando las necesidades de renovación en dichas redes. Para el desarrollo de este trabajo se aplicaron técnicas de sistemas de ayuda a la decisión multicriterio, así como algoritmos genéticos, lógica difusa, y el análisis del riesgo de la probabilidad de fallo y sus consecuencias [5].

Este trabajo compara varios modelos predictivos para la estimación de demanda de agua, presenta el uso de redes neuronales artificiales (ANN, *Artificial Neural Networks*), proyección para la búsqueda de regresiones (PPR, *Projection Pursuit Regression*), splines de regresión adaptativa multivariante (MARS, *Multivariate Adaptive Regression Splines*), regresión vectorial de soporte (SVR, *Support Vector Regression*), bosques aleatorios (RFs, *Random Forests*) y un modelo ponderado para la predicción de la demanda de agua en series altamente no lineales. Los resultados de esta comparación identificaron a los SVRs como los modelos más precisos, seguidos de cerca por MARS, PPR y RFs [6].

Actualmente se ha dado gran relevancia a la tecnología de Sistemas de Información Geográfica (SIGs) proporcionando alternativas adecuadas para una gestión eficaz de las grandes y complejas bases de datos geoespaciales [7]. Estudios recientes han mostrado altas fiabilidades localizando fugas a través de la simulación hidráulica con EPANET y aprendizaje automático usando algoritmos de clustering que tienen en cuenta la estructura topológica de un grafo[8]. Usando diferentes enfoques de *Machine Learning* se predicen anomalías en sistemas de redes de agua, mostrando que dichas técnicas pueden tener gran potencial de estimación [9].

En este trabajo, el uso de los algoritmos RFs, MARS [10] y máxima entropia (EM) [11], se integra con los SIGs, para estimar la cartografía potencial de aguas subterránea en la región de Mehran. Los resultados indican que los modelos MARS, RF y EM son buenos estimadores del potencial de fuentes de agua subterránea en el área de estudio.

Algunos métodos estadísticos bivariados y multivariados tienen inconvenientes para medir la relación entre los factores condicionantes y la ocurrencia de agua subterránea, debido a la definición de suposiciones estadísticas antes del estudio [11]; en contraste con los enfoques mencionados, técnica como los RFs y EM, que puede manejar datos de varias escalas de medición y no precisan de ninguna suposición estadística, resultaron útiles para modelar el potencial del agua subterránea [11].

Este estudio evalúa los algoritmos RF y las máquinas de vectores soporte (SVMs, support vector machines) integrados en los SIGs para la localización cartográfica de potenciales fallos y zonas de riesgo para una zona de la red de abastecimiento de agua de Bogotá.

2. Materiales y métodos

2.1 Descripción del área de estudio

La ciudad de Bogotá está situada entre los 04°36′35″ de latitud norte y los 74°04′54″ de longitud oeste, a 2625 m.s.n.m. Es la capital de Colombia. Posee aproximadamente 8.081.000 millones de habitantes. Está ubicada en la altiplanicie sobre la cordillera oriental de los Andes Sudamericanos y tiene una extensión de 1775 km2. La Figura 1 (a) muestra la localización de los fallos en la red de abastecimiento de agua para la zona de estudio.

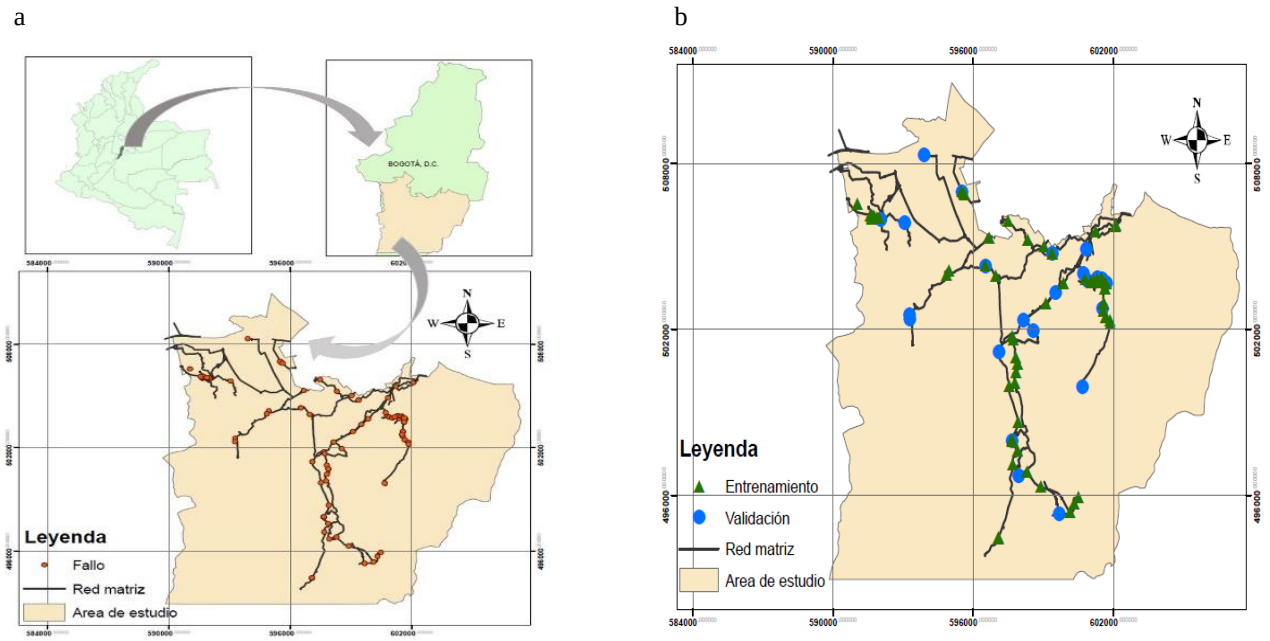
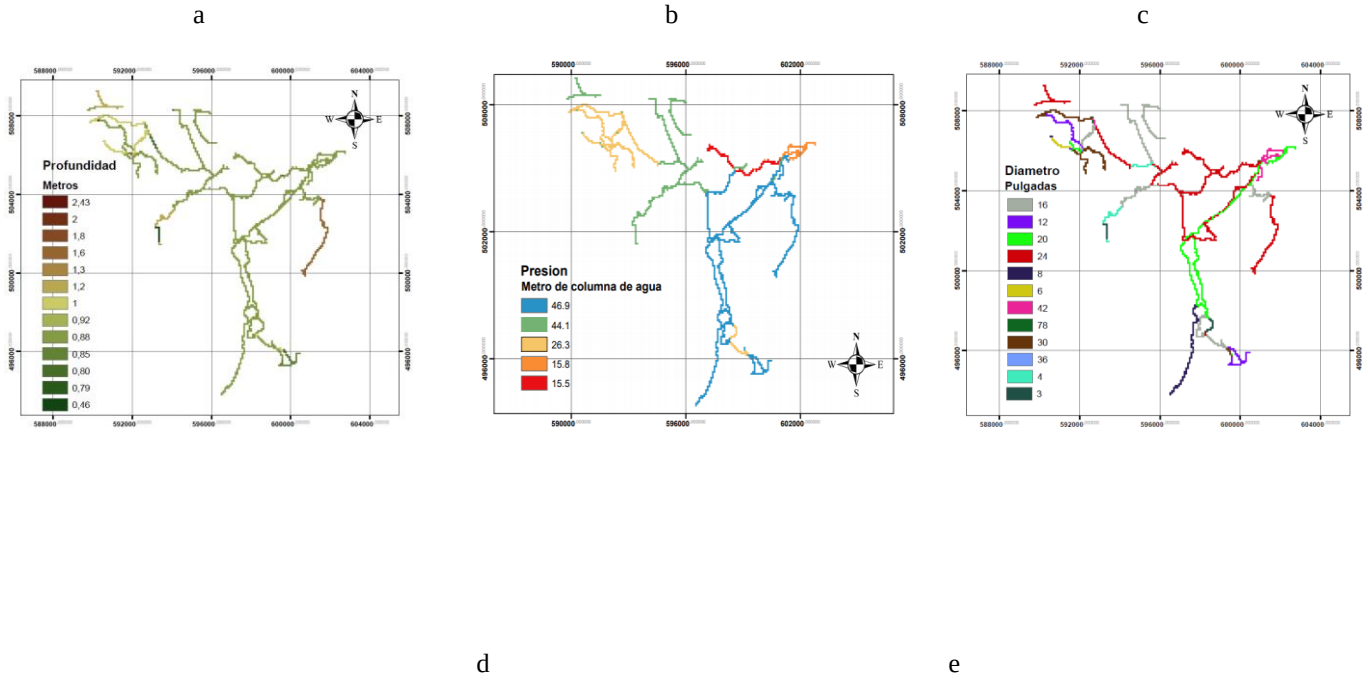


Figura. 1. (a) Localización del área de estudio, con los fallos; (b) Datos para entrenamiento y para validación

2.2 Conjunto de datos utilizado

Se analizaron los fallos presentados en la zona de estudio para el periodo de tiempo comprendido entre 2009 a 2014. En total 100 fallos fueron detectados para este periodo. Estos datos de los fallos fueron divididos utilizando un algoritmo de partición aleatoria para organizar una base de entrenamiento (70% del conjunto de datos) y una base de validación (30%). La Figura 1 (b) ilustra tanto el conjunto de datos de entrenamiento como el de validación.

Este estudio consideró 5 variables explicativas. Estas fueron: profundidad, presión, diámetro, material y pendiente. La Figura 2, presenta los mapas de las variables explicativas en el área de estudio.



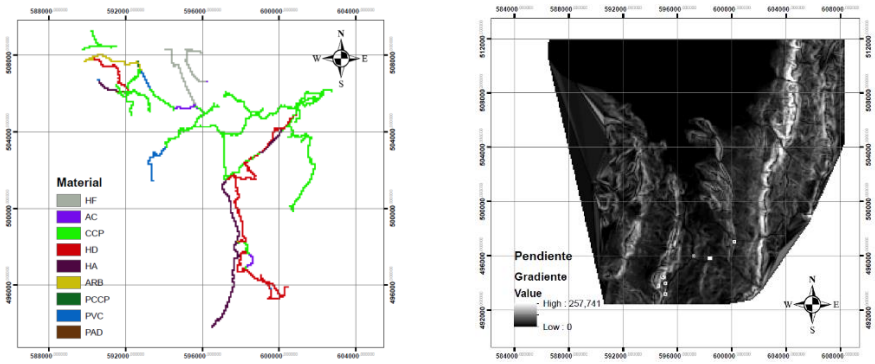


Figura. 2. Mapas de las variables explicativas; (**a)** profundidad (m); (**b**) presión (mca); (**c)** diámetro (pulgadas); (**d)** material[2]; (**e)** pendiente (gradiente)

2.3 Descripción de los modelos

2.3.1 Modelo Random Forest (RF)

Los RFs son una técnica no paramétrica [12] que se desarrolló como una extensión de los árboles de clasificación y regresión (CART) y la cual genera muchos árboles de clasificación [13] para mejorar la predicción y el rendimiento del modelo. Los RFs son una combinación de árboles predictores, de tal manera que cada árbol depende de los valores de un vector aleatorio muestreado de manera independiente y con la misma distribución para todos los árboles en el bosque. Después de que se genere un gran número de árboles, se vota por la clase más popular [12]. La Figura 3 muestra el principio de clasificación de los bosques aleatorios [14].

El algoritmo RF maneja árboles binarios aleatorios que utilizan un subconjunto de las observaciones a través de técnicas de *bootstrapping*: desde el conjunto de datos original se toma una selección aleatoria de los datos de entrenamiento y se utiliza para construir el modelo; los datos no incluidos se denominan (OOB, *out-of-bag*) [15].

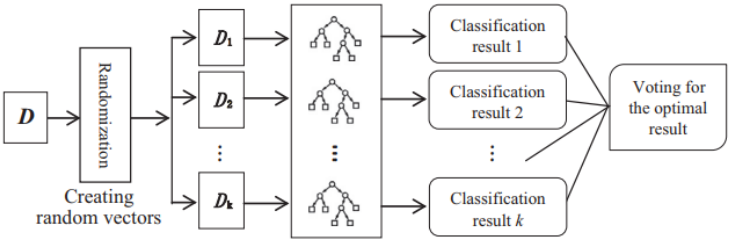


Figura. 3. Principio de clasificación de los bosques aleatorios

2.3.1.1 Definición

Un RF es un clasificador que consiste en una colección de árboles clasificadores $\{h(x, \Theta_k),\ k = 1, \ldots\}$ donde $\{\Theta_k\}$ son vectores aleatorios independientes e idénticamente distribuidos y cada árbol arroja un voto unitario para la clase más popular en la entrada x.

2.3.1.2 Características del algoritmo RF [12]

- Buena precisión
- Es relativamente robusto con los valores atípicos y el ruido.
- Es más rápido que ensayar o aumentar.
- Da estimaciones internas útiles de error, fuerza, correlación e importancia de variables.
- Es simple y fácilmente paralelizable.

[2] Material: HF: hierro fundido, AC: asbesto-cemento, CCP: acero revestido con concreto, HD: hierro dúctil, HA: hierro acerado, ARB: acero inoxidable, PCCP: tubos de presión de hormigón, PVC: cloruro de polivinilo, PAD: polietileno de alta densidad

2.3.2 Máquinas de Vectores de Soporte (SVM)

Los SVMs son una técnica de aprendizaje de máquina avanzado, basado en la minimización del riesgo estructural (SRM, *structure risk minimization*), que minimiza el error esperado de un modelo de aprendizaje y reduce el problema del sobreajuste [16].

Existen dos categorías principales de SVMs utilizadas para: la clasificación (SVM) y la regresión (SVR) mediante vectores soporte. Los procedimientos estadísticos de regresión se expresan a menudo como los procesos que derivan de una función f(x) que tiene la menor desviación entre las respuestas previstas y las observadas experimentalmente para todos los ejemplos de entrenamiento. Una de las principales características de la regresión mediante vectores soporte es que, en lugar de minimizar el error de la información observada, los SVRs intentan minimizar el límite generalizado del error para alcanzar una mejor eficiencia en el algoritmo de regresión [17].

2.3.2.1 Definición

Dado un conjunto de ejemplos de entrenamiento $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$, en el que se asume que los valores y_i de todos los ejemplos de S se pueden ajustar (o cuasi-ajustar) mediante una función lineal, el objetivo de la tarea de regresión es encontrar el valor de los parámetros $w = (w_1, \dots, w_d)$ y b que permitan definir dicha función lineal [18]

$$f(x) = (w_1 x_1 + \cdots + w_d x_d) + b = < w, x > + b \tag{1}$$

Dado que, en la práctica, es muy difícil que los ejemplos de entrenamiento se ajusten al modelo lineal con un error de predicción igual a cero, se recurre al concepto de margen blando. De esta forma se permite cierto ruido en los ejemplos de entrenamiento y, por tanto, se puede relajar la condición del error existente entre el valor predicho por la función y el valor real. Para ello se utiliza la denominada *función de pérdida e-insensible*, L_e, caracterizada por ser una función con una zona insensible, de anchura $2e$, en la que el error es nulo, y definida por [18]

$$L_e\big(y, f(x)\big) = \begin{cases} 0 & si\ |y - f(x)| \le e \\ |y - f(x)| - e & en\ otro\ caso \end{cases}. \tag{2}$$

La principal razón para elegir esta función es la de permitir cierta dispersión en el regresor linear, de tal forma que todos los ejemplos que quedan confinados en la region tubular definida por $\pm e$ no serán considerados vectores soporte [18].

2.3.2.2 Características de las máquinas de vectores soporte (SVM)[19]

- Permiten obtener una solución óptima global.
- El resultado es una solución general que evita el sobreentrenamiento.
- SVM gestiona eficientemente la información contenida en los datos y solo necesita usar un conjunto limitado de puntos de entrenamiento para alcanzar una solución.
- Las soluciones no lineales se pueden calcular eficientemente.

2.4 Importancia de las variables

Las variables de entrada en el modelo optimizado pueden ser clasificadas por importancia relativa sobre la base de la disminución media en el coeficiente de Gini. Esta medidas de importancia se pueden utilizar para clasificar variables y para la selección de variables [20].

2.5 Validación de los mapas potenciales de fallo

El paso de validación es el proceso más importante de modelado. La curva ROC se ha aplicado ampliamente en varias investigaciones para evaluar cuantitativamente la eficacia de la cartografía potencial [21]. En [22] se demuestra que el área bajo la curva ROC (AUC: *area under curve*) es útil para cuantificar la incertidumbre en predicciones de modelos, así como también, puede explicar los sesgos de detección asociados con la estimación. En el análisis, la relación cualitativa-cuantitativa entre el AUC y la precisión de predicción se clasifica de la siguiente manera: 50-60% (pobres), 60-70% (promedio), 70-80% (buena), 80-90% (muy buena) y 90-100% (excelente) [23].

3. Resultados y Análisis

3.1 Modelo Random Forest

La Figura 4 muestra las cinco variables ordenadas por la disminución media Gini. Los valores más altos indican que la variable es relativamente más importante [24]; la variable *pendiente* presentó una disminución media Gini mayor, las variables diámetro, material, profundidad y presión presentaron una menor importancia dentro del modelo.

La Figura 5 presenta la curva ROC y el AUC, el cual es de 66%; de acuerdo con lo considerado en la escala de clasificación, se podría suponer que se encuentra en un rango de predicción promedio.

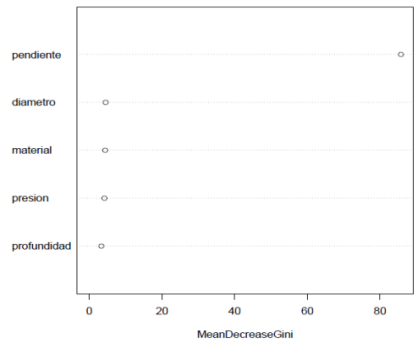


Figura. 4. Importancia de variables derivada del modelo RF

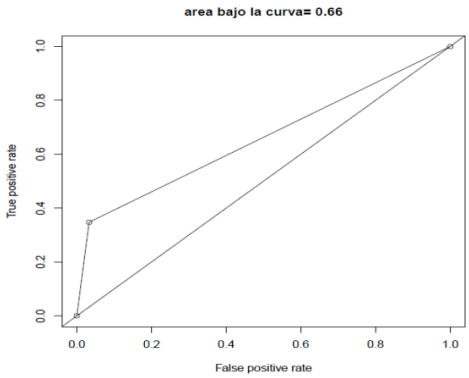


Figura. 5. Curva ROC y área bajo la curva para RF

La Figura 6 presenta el grado de importancia dado a las variables usando el SVR. De igual manera, la *pendiente* resultó ser la variable de más importancia dentro del modelo. Nuevamente, las variables diámetro, material, presión y profundidad aportan una menor variabilidad a la variable dependiente.

En la Figura 7 se presentan los resultados de la curva ROC y el área bajo la curva (AUC=0.62) para el modelo mediante la estimación SVR, presentando una categoría de predicción dentro del rango promedio.

3.2 Regresión de Soporte Vectorial

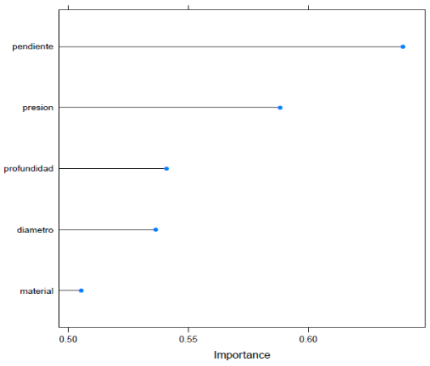


Figura. 6. Importancia de variables derivada del modelo SVR

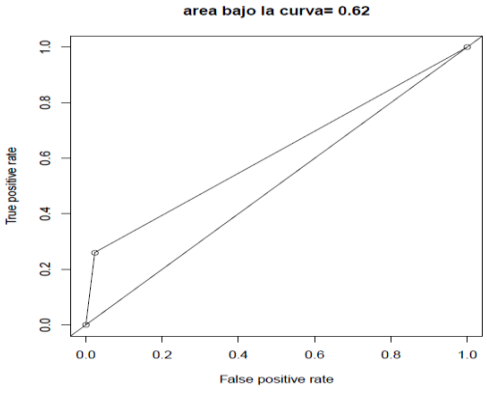


Figura. 7. Curva ROC y área bajo la curva para SVR

Finalmente, en la Figura 8 se presenta la cartografía con los posibles puntos de fallos, estimados mediante los dos algoritmos. Los puntos presentados por ambos métodos son similares en ubicación, aunque el algoritmo RF estimó una mayor cantidad de puntos de fallo, en relación a lo estimado por el SVR.

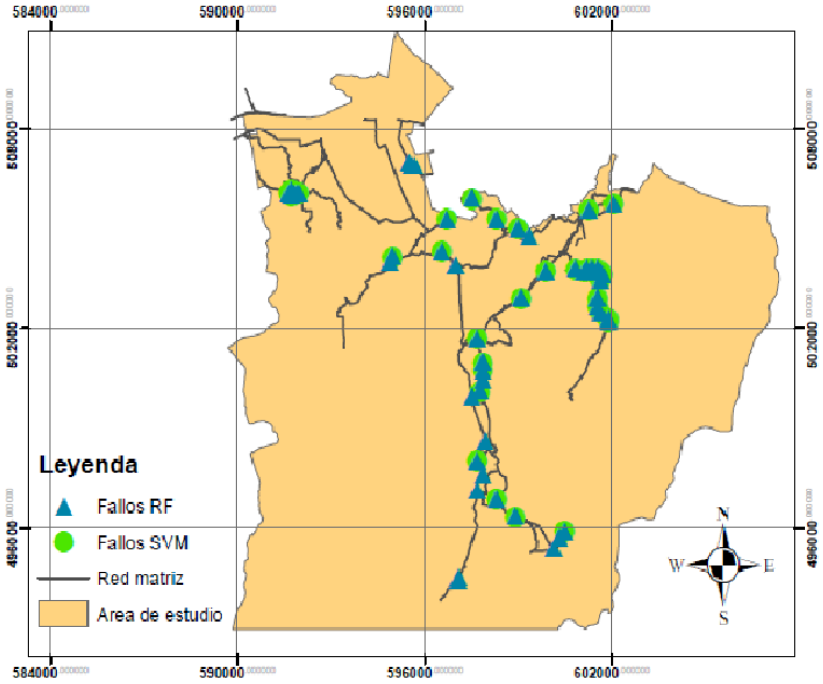


Figura. 8. Estimación de posibles fallos por los algoritmos RF y SVR

4. Discusiones y conclusiones

Este estudio identificó potenciales puntos de fallo dentro de un sector de la red de abastecimiento de aguas de Bogotá, utilizando modelos RF y SVR, esta predicción basada en variables explicativas mapeables.

La validación de los resultados indicó que el modelo RF presenta una mejor predicción (66%) comparado con el modelo de SVR (62%). El modelo RF tiene varias ventajas, el crecimiento de un gran número de árboles no genera un sobreajuste de los datos y la selección aleatoria mantiene el sesgo bajo, proporcionando mejores modelos para la predicción [25].

Los dos modelos, RF y SVR, coincidieron en considerar a la variable pendiente como la de más importancia; esto es consistente con lo que se evidencia en el mapa de posibles fallos (Figura 8) en el que éstos tienden a concentrarse mayormente hacia donde se tiene un mayor gradiente de pendiente.

Los valores promedio obtenidos en las estimaciones de predicción de los modelos pueden estar afectados por los pocos valores de fallos trabajados para el entrenamiento y validación de los algoritmos. Se sugiere tener un mayor número de datos para entrenar los modelos, ya que estos algoritmos han mostrado mejores ajustes con mayor cantidad de datos de entrada.

Otras técnicas de modelización enmarcadas dentro del *Machine Learning*, pueden ser usadas y comparadas con lo obtenido en este trabajo, revisando su nivel de adecuación y brindando nuevas alternativas en el estudio de la variable fallos dentro de las redes de abastecimiento de agua.

Los mapas de potenciales fallos, así como la importancia de las variables, pueden ser de gran utilidad a las empresas, gestores y tomadores de decisiones en lo que respecta a un manejo preventivo, brindando y asegurando una mejor continuidad en el servicio y evitando sobrecostos por atención de fallos no previstos.

Para futuros estudios, se recomienda analizar los algoritmos con mayor cantidad de variables y de puntos de muestreo, ya que, la precisión de los resultados podrían mejorarse si se aumenta la calidad de los datos; Así como analizar los algoritmos en otras zonas mapeadas por la empresa de Acueducto y Alcantarillado de Bogotá, para tener un juicio confiable sobre la eficiencia de los modelos de RF y SVR testados en éste estudio.

Agradecimientos

A la Empresa de Acueducto y Alcantarillado de Bogotá (EAAB), por la disposición en colaborar con esta clase de estudios y facilitar la información. Nuestro agradecimiento cordial, también, a los profesores Ronal Sierra y Rafael Barragán por sus sugerencias y comentarios pertinentes, de gran ayuda para el buen desarrollo de este trabajo.

Referencias

[1] C. Navarrete, B. M. Brentan, M. Herrera, J. Izquierdo, E. Luvizotto Jr, and R. PerezGarcia, "Epidemiological approach to forecasting water demand consumption through SAX."

[2] C. Navarrete-López, M. Herrera, B. M. Brentan, E. Luvizotto, and J. Izquierdo, "Network analysis for inferring spatio-temporal predictive models in water demand consumption."

[3] J. P. Bardet and R. Little, "Epidemiology of urban water distribution systems," *Water Resources Research*, vol. 50, no. 8, pp. 6447-6465, 2014.

[4] D. Yao, J. Yang, and X. Zhan, "A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines," *J Compu*, vol. 8, no. 1, pp. 170-7, 2013.

[5] C. D. Guzmán, "Programación óptima de la renovación de tuberías en un sistema de abastecimiento urbano: Análisis de los factores de influencia," 2012.

[6] M. Herrera, L. Torgo, J. Izquierdo, and R. Pérez-García, "Predictive models for forecasting hourly urban water demand," *Journal of hydrology*, vol. 387, no. 1, pp. 141-150, 2010.

[7] M. Waikar and A. P. Nilawar, "Identification of groundwater potential zone using remote sensing and GIS technique," *Int J Innov Res Sci Eng Technol*, vol. 3, no. 5, pp. 1264-1274, 2014.

[8] A. Candelieri, D. Soldi, D. Conti, and F. Archetti, "Analytical leakages localization in water distribution networks through spectral clustering and support vector machines. the icewater approach," *Procedia Engineering*, vol. 89, pp. 1080-1088, 2014.

[9] D. Vries, B. van den Akker, E. Vonk, W. de Jong, and J. van Summeren, "Application of machine learning techniques to predict anomalies in water supply networks," *Water Science and Technology: Water Supply*, vol. 16, no. 6, pp. 1528-1535, 2016.

[10] M. Zabihi, H. R. Pourghasemi, Z. S. Pourtaghi, and M. Behzadfar, "GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran," *Environmental Earth Sciences*, vol. 75, no. 8, p. 665, 2016.

[11] O. Rahmati, H. R. Pourghasemi, and A. M. Melesse, "Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran," *Catena*, vol. 137, pp. 360-372, 2016.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[13] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[14] Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, "Flood hazard risk assessment model based on random forest," *Journal of Hydrology*, vol. 527, pp. 1130-1141, 2015.

[15] F. Catani, D. Lagomarsino, S. Segoni, and V. Tofani, "Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues," *Natural Hazards and Earth System Sciences*, vol. 13, no. 11, pp. 2815-2831, 2013.

[16] P.-S. Yu, S.-T. Chen, and I.-F. Chang, "Support vector regression for real-time flood stage forecasting," *Journal of Hydrology*, vol. 328, no. 3, pp. 704-716, 2006.

[17] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203-224, 2007.

[18] E. J. C. Suárez, "Tutorial sobre máquinas de vectores soporte (sVM)," *Tutorial sobre Máquinas de Vectores Soporte (SVM)*, 2014.

[19] U. Thissen, R. Van Brakel, A. De Weijer, W. Melssen, and L. Buydens, "Using support vector machines for time series prediction," *Chemometrics and intelligent laboratory systems*, vol. 69, no. 1, pp. 35-49, 2003.

[20] M. L. Calle and V. Urrea, "Letter to the editor: stability of random forest importance measures," *Briefings in bioinformatics*, vol. 12, no. 1, pp. 86-89, 2010.

[21] H. Nampak, B. Pradhan, and M. A. Manap, "Application of GIS based data driven evidential belief function model to predict groundwater potential zonation," *Journal of Hydrology*, vol. 513, pp. 283-300, 2014.

[22] E. F. Zipkin, E. H. C. Grant, and W. F. Fagan, "Evaluating the predictive abilities of community occupancy models using AUC while accounting for imperfect detection," *Ecological Applications*, vol. 22, no. 7, pp. 1962-1972, 2012.

[23] E. K. Yesilnacar, *The application of computational intelligence to landslide susceptibility mapping in Turkey*. University of Melbourne, Department, 200., 2005.

[24] G. Williams, *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media, 2011.

[25] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181-199, 2006.