TESIS DE MAESTRÍA

Determinación del estado de redes de alcantarillado y su necesidad o no de ser sometidas a renovación/rehabilitación teniendo en cuenta Minería de Datos

Alejandra Posada Obando

Asesor: Juan G. Saldarriaga Valderrama

UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA CIVIL Y AMBIENTAL
MAESTRÍA EN INGENIERÍA CIVIL
BOGOTÁ D.C.
2019

AGRADECIMIENTOS

A mi familia, por apoyarme incondicionalmente.

A mi asesor de tesis, por confiar en mis capacidades y siempre brindarme oportunidades de crecimiento personal y profesional.

A mis amigos y compañeros de la maestría, por hacer el camino más grato al recorrerlo conmigo.

A Daniela, cuyo apoyo y conocimiento hicieron este logro posible.

A Daniel Rodriguez y la Empresa de Acueducto de Bogotá, por su apoyo y colaboración.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

1 INTRODUCCIÓN

Los Sistemas de Drenaje Urbano (SDU) son considerados una parte esencial de la infraestructura de las ciudades debido a la función que cumplen, de recolectar, transportar y disponer apropiadamente los caudales de aguas residuales y aguas lluvias producidas al interior de estos territorios. Además, estos sistemas generalmente están constituidos por las instalaciones de tratamiento y miles de kilómetros de tuberías y otras estructuras de recolección y transporte de aguas, con lo cual se pueden considerar como una de las infraestructuras que requieren inversiones intensivas de capital (Wirahadikusumah, Abraham, & Iseley, 2001). Dado lo anterior, cada vez es más esperado que estos sistemas cumplan ciertos requerimientos de desempeño que incluyen el funcionamiento de las tuberías sin obstrucciones, bajas tasas de casos de inundación, la protección de la infraestructura adyacente y el resguardo de la salud pública, con el propósito de garantizar la prestación de un servicio continuo y de calidad (Davies, Clarke, Whiter, & Cunningham, 2001).

Sin embargo, el deterioro de los componentes de estos sistemas es inevitable a lo largo de su vida útil para la mayoría de las ciudades en el mundo ya que sus sistemas fueron construidos en su mayor parte hace décadas y ha pasado el pico de inversión de capital (Rokstad & Ugarelli, 2015), por lo cual, muchas ciudades en la actualidad reciben la prestación de servicios de acueducto y alcantarillado por sistemas cuya infraestructura es vulnerable a la falla. Comúnmente, estas fallas se clasifican considerando el tipo de afectación que se produce en el sistema, ya sea el colapso o ruptura de las tuberías (fallas estructurales) o la disminución de su capacidad debido a diversos factores (operacionales). Entonces, cuando estos eventos ocurren se requieren inversiones de capital para llevar a cabo la reparación o reemplazo de las tuberías que pueden alcanzar los millones de euros (Caradot, Sonnenberg, et al., 2017). Adicionalmente a estos costos, se deben considerar costos indirectos adicionales y costos sociales que ocurren debido a la afectación de la sociedad. Estos costos corresponden, entre otros, a retrasos en el tráfico, inundación de propiedades, incomodidades generadas por el ruido, la contaminación y el olor, consecuencias ambientales y consecuencias para la salud pública (Davies et al., 2001).

Debido a lo anterior, resulta indispensable que la infraestructura y los componentes de los SDU sean administrados de tal manera que se minimice la probabilidad de ocurrencia de falla del sistema y se haga uso óptimo de los recursos disponibles. Esta administración de los componentes de SDU generalmente es conocida como Gestión de Activos bajo tierra, y consiste de varias etapas que incluyen actividades de inspección, categorización, rehabilitación y reemplazo de estos componentes, entre otras.

La gestión de activos bajo tierra en los sistemas de alcantarillado se realiza con el propósito de mantener, durante el mayor tiempo posible, la funcionalidad del sistema o para restaurar/mejorar

el estado de alguno de los componentes del mismo, y así garantizar el desempeño óptimo del diseño inicial a lo largo de su vida útil. Esta gestión se puede realizar mediante dos enfoques, los cuales se basan en el mantenimiento mediante actividades proactivas (antes de la falla) o reactivas (después de la falla). Con base en lo anterior, se puede realizar una clasificación de tres tipos de mantenimiento comúnmente implementados: correctivo, preventivo y predictivo (New England Interstate Water Pollution Control Commission, 2003).

En general, en los sistemas reales de alcantarillado la gestión de activos se lleva a cabo mediante una combinación de estos tres tipos de mantenimiento, siendo el objetivo la reducción de actividades correctivas y la priorización de actividades preventivas y predictivas. Sin embargo, la capacidad de alcanzar este objetivo se ve limitada en muchos casos debido al mantenimiento histórico que se ha dado a las redes de drenaje, con lo cual, sistemas en los cuales se ha priorizado un mantenimiento correctivo tendrán mayor dificultad para implementar un enfoque proactivo debido a que sus recursos se dirigen hacia el enfoque reactivo (New England Interstate Water Pollution Control Commission, 2003). De acuerdo con la normativa de inspección y rehabilitación de redes de alcantarillado de la ciudad de Bogotá (NS-058 y NS-061), lo anterior corresponde al caso que se presenta en el sistema de esta ciudad, en donde los esfuerzos de mantenimiento son principalmente reactivos debido a la distribución de recursos, a pesar de contar con algunas herramientas y procedimientos de mantenimiento proactivo.

Realizar esta transición hacia medidas de mantenimiento proactivo se vuelve una necesidad para las empresas prestadoras del servicio de alcantarillado pues a medida que los sistemas envejecen las estrategias reactivas se vuelven menos viables y se incrementan los costos de financiación de estas (Rokstad & Ugarelli, 2015). Así mismo, conocer el estado de los componentes de los sistemas es necesario para identificar posibles fallas futuras y garantizar la prestación de un buen servicio; sin embargo, realizar la inspección de las redes en su totalidad requiere de una gran cantidad de tiempo y dinero, lo cual resulta generalmente en bajas tasas de inspección de las redes debido a restricciones de presupuesto (Harvey & McBean, 2014). Por lo tanto, como respuesta a estas necesidades, se han desarrollado diversos modelos de deterioro con el propósito de determinar y predecir el estado de los componentes del sistema a lo largo de su vida útil; estos modelos se pueden clasificar en modelos determinísticos, estadísticos y de aprendizaje automático (machine learning) (Caradot, Sonnenberg, et al., 2017). Los modelos determinísticos buscan entender los diferentes mecanismos físicos que generan el deterioro de las tuberías, pero generalmente incluso los modelos más complejos son demasiado simplificados para modelar el complejo proceso de deterioro. Los modelos estadísticos usan relaciones matemáticas para relacionar la caracterización histórica del estado de las tuberías con factores de deterioro de las tuberías y sus resultados se expresan como valores de probabilidad. Finalmente, los modelos de aprendizaje automático permiten identificar relaciones complejas no lineales entre las entradas y las salidas, sin limitarse a

una expresión predefinida vinculando las variables de entrada y los resultados obtenidos (Caradot et al., 2017).

Algunos de los modelos estadísticos y de aprendizaje automático también se pueden clasificar como modelos de Minería de datos y han sido ampliamente investigados para el caso de SDU en los últimos años debido a que el estudio de la caracterización del estado de las tuberías se puede entender como un problema de clasificación supervisada en términos de Minería de Datos y es una de las áreas más estudiadas en materia de análisis de datos (Wright, Heany, & Dent, 2006). De igual manera, se ha observado que estos permiten hacer frente al reto que representa lograr una gestión eficiente, sostenible y racional de la infraestructura de las redes de alcantarillado, considerando las limitaciones de información y capital y las regulaciones ambientales cada vez más exigentes (Baik, Jeong, & Abraham, 2006), al mismo tiempo que permiten el entendimiento en más detalle del comportamiento de los sistemas de drenaje urbano y la caracterización del proceso de falla de tuberías en términos de variables registradas en los procesos de inspección de las redes.

Estas últimas razones resultan particularmente relevantes en el caso de países en desarrollo ya que, en primer lugar, se tienen grandes limitaciones de recursos (financieros y humanos) disponibles para los procesos de mantenimiento, por lo cual no es posible inspeccionar el 100% de los sistemas y, en segundo lugar, como consecuencia de lo anterior, el proceso de toma de decisiones para la priorización del mantenimiento de las redes requiere ser apoyado por información técnica y relaciones justificadas entre las variables de deterioro y las fallas en el sistema. Entonces, el soporte que dan estos modelos respecto a los recursos limitados se percibe cuando la información que se registra en la inspección de una porción de las redes de alcantarillado permite obtener un material del cual es posible explotar conocimiento y patrones para predecir el comportamiento de la totalidad del sistema, al igual que garantizar la inversión de recursos de manera más eficiente; y también, cuando se da un soporte técnico a un proceso de toma de decisiones inmerso en un contexto sociotécnico complejo (van Riel et al., 2014a), debido a la consideración de otros factores como: la inversión de capital en mejorar infraestructura visible, la rehabilitación simultanea de otro tipo de infraestructura como vías y carreteras, e incluso la influencia de la opinión y el instinto de los administradores del sistema encargados de estas actividades en las empresas prestadoras del servicio (New England Interstate Water Pollution Control Commission, 2003; Van Riel, Langeveld, Herder, & Clemens, 2014b; Van Riel et al., 2014a).

Así, este trabajo busca identificar, mediante la literatura técnica disponible, el proceso sistemático requerido para llevar a cabo el mantenimiento proactivo de los sistemas de drenaje urbano, haciendo principal énfasis en el modelamiento predictivo de la condición de las redes. Lo anterior, teniendo en cuenta que muchos autores resaltan tres actividades fundamentales para llevar a cabo una gestión proactiva de activos en las empresas de alcantarillado: primero, evaluar la condición, desempeño y capacidad de cada componente de la red; segundo, realizar predicciones del estado futuro de sus componentes durante su ciclo de vida; y finalmente, determinar la prioridad de las

intervenciones a realizar antes de los eventos de falla (EPA, 2009). Así mismo, se busca analizar la viabilidad de implementar métodos de modelación predictiva en ciudades con bajas tasas de inspección y limitaciones de presupuesto y recursos. Entonces, el capítulo 3 de este documento se enfoca en identificar y describir los diversos factores que se han considerado influyentes en el deterioro de las tuberías, al igual que los procesos actuales para la clasificación del estado de las tuberías y la incertidumbre asociada a estos procesos. Luego, el capítulo 4 describe los diferentes tipos mantenimiento realizados actualmente por las empresas prestadoras del servicio de alcantarillado incluyendo: mantenimiento correctivo, preventivo y predictivo. A continuación, el capítulo 5 describe las diferentes metodologías de Minería de Datos aplicadas al problema de predicción del estado de redes de drenaje urbano, las medidas de desempeño utilizadas en estos modelos y los casos de estudio en los cuales han sido aplicadas estas metodologías a la predicción de fallas en redes de alcantarillado, considerando la cantidad y calidad de la información disponible. Seguido, en el capítulo 6 se realiza un diagnóstico de los métodos de gestión actuales de la EAB, la normativa con la cual se realiza la estimación del estado operacional y estructural de sus activos y la frecuencia con que se realizan las inspecciones en las redes. El capítulo 7 presenta la aplicación de diferentes modelos de minería de datos a un caso de estudio sintético generado a partir del comportamiento reportado para un conjunto de tuberías inspeccionadas en la ciudad y, a partir de este, la capacidad de predicción generalizada de cada modelo considerando el efecto de la información disponible para su calibración. Finalmente, en el capítulo 8 se analizan las ventajas y retos de la aplicación de técnicas de minería de datos como herramientas para la gestión proactiva de redes de alcantarillado y en el capítulo 9 se presentan las conclusiones de la investigación.

2 OBJETIVOS

2.1.1 Objetivo General

Analizar la viabilidad de la aplicación de metodologías de Minería de datos para determinar la necesidad de rehabilitación de tuberías en redes de drenaje. Así, implementar y evaluar el desempeño de los modelos más apropiados para la predicción del deterioro estructural de sistemas de alcantarillado en un caso de estudio de la ciudad de Bogotá.

2.1.2 Objetivos Específicos

- Realizar una revisión bibliográfica crítica de modelos de deterioro utilizados para la rehabilitación de redes de alcantarillado, incluyendo métodos de Minería de datos y otros métodos utilizados actualmente para la determinación de fallas en tuberías.
- Identificar las medidas de desempeño de los modelos de deterioro y las metodologías de pre y post procesamiento de datos utilizadas en la rehabilitación de redes de alcantarillado.
- Determinar las ventajas, las limitaciones y retos de los modelos de deterioro utilizados para la rehabilitación de redes de alcantarillado reportados en la literatura.
- Estudiar los tipos de mantenimiento implementados en redes de alcantarillado y su relación con el proceso de deterioro de las mismas.
- Establecer las principales causas de deterioro de tuberías en las redes de alcantarillado y su relación con las variables topológicas de las redes.
- Realizar un análisis de sensibilidad de las variables involucradas en el proceso de deterioro de tuberías y su capacidad explicativa en la predicción de fallas en redes de alcantarillado.
- Establecer los lineamientos generales del modelamiento predictivo en el ámbito de redes de alcantarillado, explicando el buen procesamiento y uso de datos, la complejidad del problema y la evaluación de modelos adecuados.
- Realizar un diagnóstico de la gestión actual de los activos de las redes de alcantarillado de la ciudad de Bogotá
- Analizar la viabilidad de implementar los modelos de minería de datos más apropiados en un caso de estudio de la ciudad de Bogotá, considerando el registro limitado de información y las variables inspeccionadas por las empresas prestadoras del servicio de agua potable y alcantarillado.

3 PROBLEMAS COMUNES EN SISTEMAS DE ALCANTARILLADO

3.1 Generalidades y Descripción del problema

Los sistemas de recolección y transporte de alcantarillado corresponden a una parte valiosa, extensa y compleja de la infraestructura de los países. Estos sistemas están compuestos por tuberías, conductos, estaciones de bombeo, pozos de inspección y otras estructuras que sirven para la recolección de aguas lluvias y aguas residuales, y el transporte de las mismas a las instalaciones que proporcionan un adecuado tratamiento y disposición a los cuerpos receptores (New England Interstate Water Pollution Control Commission, 2003).

En países en desarrollo, se conoce que existe un reto frente a la cobertura de los servicios de acueducto y alcantarillado, pues generalmente las tasas de cobertura no alcanzan un 100%, a pesar del incremento en las inversiones de los últimos años. En particular, en el caso de Colombia, la cobertura de Acueducto es del 92,3% y la de Alcantarillado corresponde a 88,2% (El Espectador, 2018). Sin embargo, también es necesario considerar que estos países, al igual que los países desarrollados, deben lidiar con el envejecimiento de sus sistemas y las diferentes consecuencias que se generan cuando no se realiza una adecuada evaluación y administración del comportamiento de sus componentes.

En general, es considerado que la ausencia de los sistemas de alcantarillado está directamente vinculada a la generación y propagación de problemas de salud pública. Asimismo, también es posible inferir que los problemas anteriores se pueden generar debido al malfuncionamiento de estos sistemas y que las consecuencias de los problemas de envejecimiento de las redes se vuelven más graves cuanto más tiempo pasen desatendidos, generando así riesgos inaceptables para la salud humana, el medio ambiente y la infraestructura próxima, al igual que afectando la economía de las ciudades (EPA, 2017). En particular, la Agencia de Protección Ambiental de los Estados Unidos (US EPA) ha manifestado una alta preocupación por el control y la eliminación de los desbordamientos en los sistemas de alcantarillado (SSO's y CSO's), los cuales ocurren en un alto porcentaje como resultado del deterioro de los sistemas de drenaje o ausencia de mantenimiento de los mismos. De hecho, en Estados Unidos se pueden asociar cerca del 70% de eventos de desbordamiento del sistema con colapsos o bloqueos de tuberías principales o secundarias de las redes de alcantarillado; causando así un vínculo directo entre las fallas en sistemas de alcantarillado y las enfermedades causadas debido a la exposición a patógenos presentes en aguas residuales (EPA, 2017).

Luego, es esperado que debido a la gravedad de las consecuencias de los eventos de falla en las redes alcantarillado y a la necesidad de gestionar de forma eficiente los recursos económicos, las empresas se vean en la necesidad gestionar eficientemente su infraestructura y más aún de realizar la transición de un enfoque de mantenimiento reactivo a uno proactivo, implementando

gradualmente procedimientos preventivos y predictivos para la operación de sus redes. Varios autores han demostrado que para efectuar un mantenimiento proactivo se requiere un enfoque sistemático en que primero se evalúe la condición, desempeño y capacidad de cada componente de la red; en segundo lugar se realicen predicciones del estado futuro de estos componentes durante su ciclo de vida; y por último, se determine la prioridad de las intervenciones a realizar antes de los eventos de falla (EPA, 2009).

Entonces, para llevar a cabo el primero paso de este enfoque sistemático, que corresponde a la evaluación de la condición de los componentes de las redes, es necesario determinar y clasificar las fallas típicas que ocurren en las redes de alcantarillado, evaluar los factores que son relevantes para los diversos mecanismos de falla y establecer los lineamientos con los cuales se clasifica la condición (estructural o de servicio) de las tuberías. En los siguientes dos capítulos de este documento (3.2, 3.3) se desarrollan los temas mencionados anteriormente.

3.2 Fallas típicas en Sistemas de Alcantarillado

Estudiar los tipos de fallas que pueden ocurrir en los sistemas de alcantarillado es de interés en el proceso de rehabilitación pues mediante la caracterización de estas fallas es posible identificar cuando surge la necesidad de llevar a cabo la rehabilitación de la red, al igual que la tecnología y las actividades requeridas restaurar el desempeño (hidráulico o estructural) de un componente en la red (Empresa de Acueducto de Bogotá E.S.P., 2001). Estas fallas generalmente son clasificadas en dos categorías, dependiendo de si el defecto encontrado afecta el desempeño hidráulico (fallas operacionales) o el desempeño estructural de las redes (fallas estructurales) (Mcdonald & Zhao, 2001). En general, se puede considerar la ocurrencia de fallas en todos los componentes que constituyen los sistemas de drenaje, abarcando las tuberías de recolección y disposición, los pozos de inspección, las instalaciones de bombeo y demás estructuras mencionadas anteriormente; sin embargo, las investigaciones generalmente se encuentran enfocadas en el estudio del comportamiento de las tuberías en las redes y los diferentes mecanismos de falla que se presentan en estas dadas sus características. Lo anterior puede deberse al alto porcentaje que estos elementos representan de la infraestructura total de la red, por lo cual muchas de las investigaciones se enfocan en el estudio de las fallas específicamente en estos componentes. Por esta razón, las fallas estudiadas en este capítulo estarán principalmente enfocadas en aquellas que ocurren en las tuberías de las redes.

Estas fallas en las tuberías pueden ocurrir por diversos factores externos o internos en el sistema, que someten a las tuberías a condiciones en las cuales se presentará algún mecanismo de falla, dependiendo de las características propias de las tuberías y/o las condiciones en las cuales se encuentren instaladas (Davies et al., 2001). Al igual que en los demás componentes de las redes, los defectos encontrados en las tuberías se agrupan en problemas estructurales (degradación, riesgo de colapso, etc.) o problemas operacionales (obstrucciones, infiltraciones, etc.) y estos pueden ser

analizados como fallas localizadas (que afectan uno o varios puntos de una tubería) o fallas generalizadas que requieren la intervención de un área aferente (Empresa de Acueducto de Bogotá E.S.P., 2001; Mcdonald & Zhao, 2001).

A continuación se presenta una descripción de los defectos más comunes encontrados en estos sistemas categorizados en las fallas operacionales y estructurales.

Figura 3-1. Clasificación de los tipos de fallas en sistemas de alcantarillado.

3.2.1 Fallas operacionales

Estas fallas son aquellas que se encuentran relacionadas con la perdida de la capacidad en la conducción de los flujos establecida en el diseño de las tuberías, ya sea debido a reducción de la sección transversal o por el aumento no esperado de los caudales. Así, estos defectos también se conocen como fallas de servicio pues reducen la capacidad hidráulica del sistema y el nivel de servicio entregado. Entre estos defectos se encuentran: infiltraciones y exfiltraciones generalizadas, intrusiones de raíz, incrustaciones y acumulación de residuos.

Así, las principales fallas operacionales que son consideradas en los sistemas de alcantarillado son (Damvergis, 2014; Mcdonald & Zhao, 2001):

- Infiltraciones y exfiltraciones generalizadas:

Las entradas de agua en las tuberías pueden ocurrir directamente en las tuberías (a través de los pozos o las conexiones de servicio) o por la entrada lateral de flujo subsuperficial. Las infiltraciones se entienden como el proceso lento mediante el cual se incrementan los caudales de flujo debido a altos niveles freáticos que ingresan en el sistema de drenaje, y las exfiltraciones corresponden a la salida de agua de las tuberías de alcantarillado (Ver **Figura 3-2**) . Estas fallas ocurren comúnmente debido a defectos ya existentes en las tuberías (Damvergis, 2014).

Estas fallas se pueden producir por deficiencias en el diseño o construcción de los sistemas, altos niveles freáticos y alta permeabilidad del suelo, conexiones ilegales al sistema y envejecimiento de los materiales.

Figura 3-2. Infiltración de agua en sistemas de alcantarillado. Tomado de Stein & Stein, 2004.

- Obstrucciones/bloqueos de flujo:

Las obstrucciones de flujo son objetos o materiales que se encuentran en la sección transversal de la tubería y que representan un obstáculo para el flujo normal de las aguas residuales, implicando la reducción del área de la sección transversal requerida para el flujo habitual del agua. Los obstáculos típicamente encontrados son: depósitos de residuos sólidos (colmatación), acumulación de sedimentos minerales (incrustaciones) y la intrusión de raíces en las tuberías.

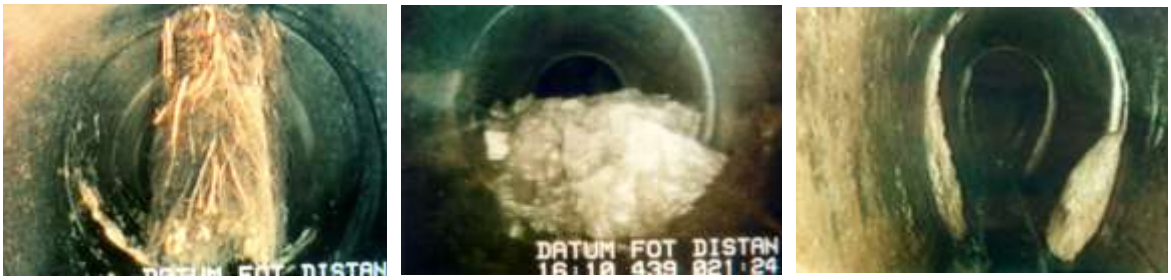


Figura 3-3. Obstrucciones de flujo en tuberías. (a) Intrusión de raíces, (b) depósitos de residuos, (c) incrustación. Tomado de Stein & Stein, 2004.

3.2.2 Fallas estructurales

Estas fallas corresponden a los defectos en las tuberías que generan una disminución parcial o completa de la capacidad estructural del sistema, y por lo tanto, se puede inferir que están relacionadas típicamente con las cargas verticales a las que son sometidas, la capacidad portante del suelo y el material de las tuberías. Estos defectos eventualmente conducen al colapso de las tuberías, mediante un proceso que se divide comúnmente en tres etapas (Davies et al., 2001; WRc, 2001):

1. Defectos iniciales – el colapso generalmente se origina con un defecto menor que permite el desarrollo de un mayor deterioro. Estos pueden generarse debido a cargas verticales excesivas, suelos de pobres características o malas prácticas de construcción e instalación de las tuberías.

2. Deterioro – Esta etapa corresponde comúnmente a los defectos que se generan debido a la perdida de la capacidad portante del suelo en que se encuentra instalada la tubería o a causa del deterioro propio del material ya sea por la interacción entre el sistema suelo-tubería o el contacto de químicos en el agua con las paredes de la tubería.

3. Colapso – estas fallas ocurren generalmente debido a un evento especifico, después de que se ha producido el deterioro suficiente en la tubería para que el colapso sea probable. Debido a lo anterior, no es posible predecir cuándo se producirá el colapso de una tubería y se considera más viable establecer el grado de deterioro a partir del cual este evento tiene una alta probabilidad.

Así, las principales fallas estructurales que son consideradas en los sistemas de alcantarillado son (Mcdonald & Zhao, 2001; Stein & Stein, 2004):

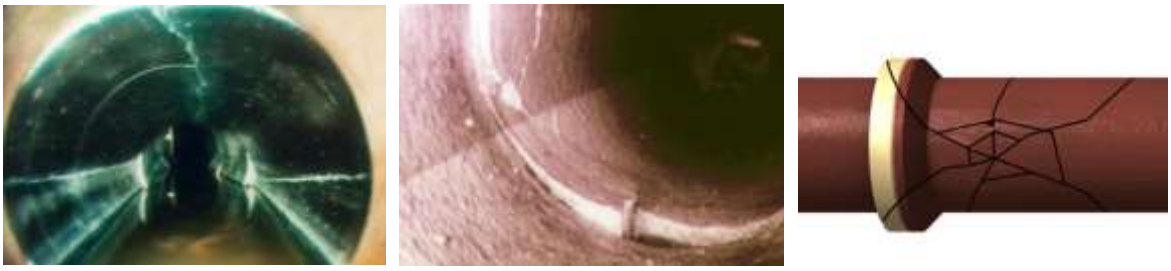
- Daños superficiales debido a Corrosión - Abrasión:

La abrasión de las paredes de las tuberías ocurre debido al desgaste producido por la alta velocidad a la cual pueden llegar a moverse la cantidad considerable de solidos inorgánicos (arena o arenilla) en suspensión que contiene el agua transportada; mientras que la corrosión de los componentes del sistema ocurre debido a que las redes de alcantarillado transportan aguas residuales que pueden liberar gases como H_2S, que al entrar en contacto con la humedad se convierten en ácido sulfúrico (Zaher, s. f.) (Ver **Figura 3-4**). La ocurrencia de estos dos eventos puede ser causada por deficiencias en los diseños, aguas residuales con componentes especiales debido a descargas industriales no controladas o la selección inapropiada de materiales en las tuberías, resultando en el problema estructural del adelgazamiento de la pared de la tubería y/o la creación de agujeros.

Figura 3-4. Corrosión de la superficie interior de una tubería. Tomado de Stein & Stein, 2004.

- Fisuras - grietas laterales y longitudinales:

Este tipo de defecto ocurre principalmente en tuberías rígidas y comúnmente se realiza una distinción de tres tipos de grietas que pueden llevar a la rotura o colapso de las tuberías: fisuras longitudinales, laterales y puntuales (Ver **Figura 3-5**). Las causas de este tipo de defecto están estrechamente relacionadas con el tipo de grieta, por lo cual su forma, dimensión y curso representan información importante para la determinación de la fuente de la falla; estos defectos se pueden originar por deficiencias en los diseños o la construcción de las tuberías, daños ocasionados durante su transporte e instalación o debido al proceso de deterioro.



a. **b.** **c.**

Figura 3-5. Fisuras en las tuberías. (a) Longitudinal, (b) Lateral, (c) Puntual. Tomado de Stein & Stein, 2004.

- Rotura o colapso de las tuberías:

Se entiende por rotura de una tubería la falta de piezas de la pared de la tubería de un tamaño significativo. Por otro lado, el colapso se entiende como la pérdida total de capacidad de carga debido a la destrucción de la tubería (Ver **Figura 3-6**). Estas dos fallas generalmente ocurren como consecuencia de otros defectos menores que progresan debido al proceso de deterioro de las tuberías.

a. **b.**

Figura 3-6. (a) Rotura de una tubería, (b) Colapso de una tubería. Tomado de Stein & Stein, 2004.

- Deformaciones en las tuberías:

Las deformaciones en las tuberías comúnmente se estudian considerando la rigidez que presenta el sistema tubería – suelo, analizando la forma en que se realiza la transferencia de cargas. Estos defectos pueden generarse por deficiencias de diseño y construcción, materiales deficientes en las estructuras de soporte de las tuberías o suelos de pobres características, y efectos de la temperatura (Ver **Figura 3-7**). Como consecuencia de las deformaciones se pueden presentar fugas, fracturas y colapso de las tuberías.

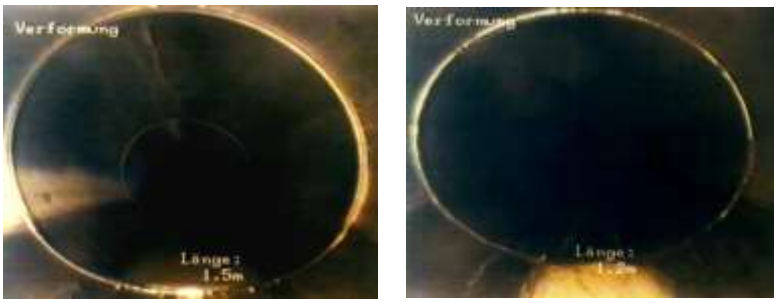


Figura 3-7. Deformación de una tubería de polietileno de alta densidad (HDPE). Tomado de Stein & Stein, 2004.

- Desviación de la posición:

Este defecto se entiende como la desviación no planificada de las tuberías y otras estructuras de su posición nominal establecida en la planeación y construcción del sistema. En el caso de las tuberías, esta desviación se puede distinguir de acuerdo a la dirección en que ocurre el desplazamiento: vertical, horizontal o longitudinal (Ver **Figura 3-8**). Estos movimientos pueden ocasionarse debido a asentamientos del suelo, eventos extremos como terremotos, cambios hidrogeológicos, cambios en las cargas de las tuberías o como consecuencia de fugas; y comúnmente tienen consecuencias problemas de alineación de las tuberías, fugas, ruptura y colapso de las tuberías.

a. b. c.

Figura 3-8. Desviación de posición en las tuberías. (a) Vertical, (b) Longitudinal, (c) Horizontal. Tomado de Stein & Stein, 2004.

3.3 Factores que influencian el deterioro y colapso de tuberías

Se considera que la ocurrencia y propagación de los defectos en las tuberías, al igual que la tasa de deterioro estructural se ve afectada por un gran número de factores. Entre otras, se ha estudiado en diferentes investigaciones, la influencia de variables físicas de las tuberías como el diámetro, la longitud, la profundidad, el material, el tipo de junta y la pendiente; al igual que la influencia de factores externos como las características del suelo, el uso del suelo y variables ambientales; asimismo, se han investigado otros factores entre los cuales: la edad de las tuberías, el tipo de sistema y el tipo de tubería (Davies et al., 2001; López Kleine, Hernandez, & Torres, 2016). Por otro lado, recientemente se han considerado influyentes nuevos factores como el cambio climático, el cambio del suelo y el crecimiento demográfico (Kleidorfer et al., 2013). Los trabajos de Harvey & McBean (2014) y Bailey et al. (2015) son ejemplos de investigaciones en las cuales se ha buscado entender la influencia de variables adicionales como la densidad de propiedades circundantes y la velocidad de flujo.

En general, la consideración de estos factores que influencian el proceso de deterioro en las tuberías puede basarse en el conocimiento que se tiene de los procesos físicos y químicos que originan los diferentes mecanismos de falla, como algunos de los que se mencionan en las secciones 3.2.1 y 3.2.2 de este documento. Sin embargo, como se ha mencionado anteriormente, encontrar modelos determinísticos o relaciones explicitas entre estas variables y las fallas con base en el proceso de deterioro puede constituir un problema muy complejo; por lo cual, el impacto de estas variables en los procesos de deterioro ha sido comúnmente evaluado a partir de la aplicación de modelos estadísticos y modelos de aprendizaje automático.

Estos factores generalmente se clasifican en cuatro categorías que tienen en cuenta las características internas del sistema, las condiciones externas y los aspectos operacionales de las redes. La primera categoría (factores físicos) corresponde a los atributos físicos del sistema y/o las tuberías; la segunda (factores ambientales) hace referencia a las características ambientales del entorno circundante; la tercera (factores operacionales) contiene las variables implicadas con el

funcionamiento u operación de las tuberías; y la cuarta (factores de construcción) incluye las variables de los métodos de instalación y construcción empleados (Ver **Figura 3-9**).

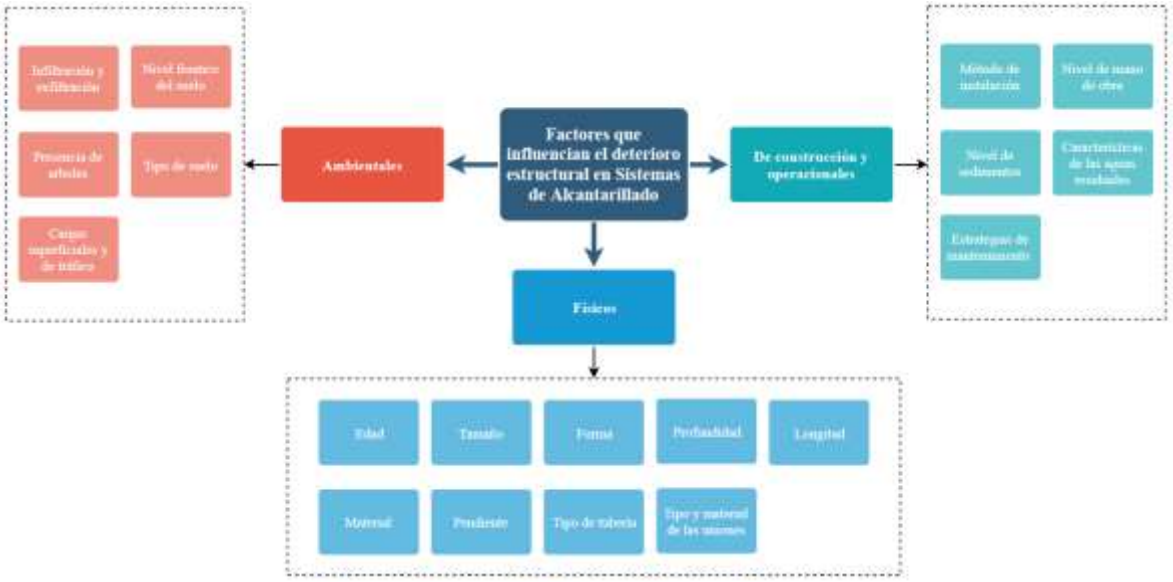


Figura 3-9. Factores que influencian el deterioro estructural en sistemas de alcantarillado. Adaptado de Ana & Bauwens, 2010.

A continuación, se presenta una descripción de los factores comúnmente encontrados como relevantes para estimar y predecir la generación y potencial evolución de las fallas en redes de alcantarillado:

- Edad de las tuberías:

La edad de las tuberías es uno de los factores que presenta gran variabilidad al estudiar componentes en redes de alcantarillado, lo cual se debe en general debido al desarrollo histórico y expansión de los sistemas de drenaje que se ha llevado a cabo durante largos periodos de tiempo.

Múltiples investigaciones, entre las cuales: Ariaratnam, El-Assaly, & Yang (2001); Savic, Giustolisi, & Laucelli (2009); Mashford, Marlow, Tran, & May (2011); Harvey & McBean (2014); Harvey, Wheeler, & Mcbean (2015); Salman & Salem (2012); (Caradot, Kley, Kropp, & Schmidt, 2013); Nicolas Caradot et al. (2017) y Berardi, Giustolisi, Savic, & Kapelan (2009) encuentran la edad de las tuberías o su año de instalación como un factor relevante para la predicción de su condición estructural. En general, se espera que el deterioro de estos componentes sea mayor a medida que se incrementa su edad; sin embargo, las tasas de deterioro de las tuberías son muy variables dependiendo de la construcción de las tuberías y factores operacionales. Así, no es posible generalizar la predicción de una peor condición estructural en tuberías de mayor edad, ni lo contrario en tuberías recientemente instaladas (Caradot et al., 2014).

Al igual que se ha encontrado la significancia de este factor al describir el proceso de deterioro de las tuberías, se han considerado las posibles limitaciones que implica el uso exclusivo de esta variable para estimar la condición. La edad de las tuberías o su año de instalación puede resultar en un indicador erróneo de las fallas en tuberías en los casos en los cuales no se considera que las tuberías han sido rehabilitadas a lo largo de su vida útil y las diferentes condiciones de construcción y de operación bajo las cuales se han manejado. Lo anterior afecta la relación entre la tasa de deterioro y la edad de estos componentes encontrada para estados estructurales buenos e intermedios, pues el principal caso en el cual este factor no resulta en un predictor apropiado corresponde al caso de las tuberías en pobres estados estructurales, cuya rehabilitación ha ocurrido poco tiempo antes de realizar las inspecciones utilizadas para el desarrollo de cualquier modelo de deterioro. Mas aún, en el estudio realizado por Syachrani, Seok, & Chung (2013) se sugiere el uso de la "edad real" de las tuberías para la predicción del deterioro de las tuberías, correspondiendo esta nueva variable a la edad ajustada dada la ubicación y las condiciones operacionales de estos componentes.

Así, contar con información adicional a la edad de las tuberías como el año de instalación, el año de inspección y sus condiciones de operación al momento de explicar el proceso de deterioro puede incrementar la asertividad de los modelos utilizados.

- Tamaño de las tuberías:

El efecto del tamaño de las tuberías en la estabilidad estructural de las mismas ha sido estudiado por un gran número de autores. Muchos autores (Angarita, Vargas, & Torres, 2017; Baik et al., 2006; Bailey et al., 2015; Harvey & McBean, 2014; Harvey et al., 2015; Jung, Garrett Jr., Soibelman, & Lipkin, 2012; Laakso, Kokkonen, Mellin, & Vahala, 2018; López Kleine et al., 2016; Mashford et al., 2011; Savic, Giustolisi, & Shepherd, 2006; Wright et al., 2006) han considerado la importancia de esta variable en la implementación de diferentes modelos estadísticos y de aprendizaje automático, encontrando en la mayoría de los casos que este factor es significativo para explicar las fallas en tuberías en redes de alcantarillado. Sin embargo, los resultados de estas investigaciones no siempre son consistentes respecto a si existe un aumento en la probabilidad de falla en las tuberías de mayor diámetro o si ocurre lo contrario.

Micevski, Kuczera, & Coombes (2002) encontraron mediante la aplicación de un modelo de Cadenas de Markov que el deterioro en tuberías de aguas lluvias con menores diámetros es mayor que en tuberías más grandes, atribuyendo esto a la subestimación de las cargas de tráfico o de la profundidad mínima requerida para su cobertura en tuberías pequeñas. Otros autores como Baik et al. (2006); Khan, Zayed, & Moselhi (2010) encuentran que la tasa de falla se incrementa en tuberías de mayores diámetros; mientras que los estudios realizados por Berardi, Giustolisi, Savic, & Kapelan (2009); N. Caradot et al. (2018); Harvey & McBean (2014); Harvey et al. (2015); Laakso,

Kokkonen, et al. (2018); Savic et al. (2006) y Wright et al. (2006) indican que tuberías de menor tamaño tiene una mayor probabilidad de falla que tuberías más grandes.

Entre las explicaciones consideradas por los autores para una tasa de fallas mayor en tuberías de menor tamaño se encuentran: una supervisión menos cuidadosa del proceso de instalación en tuberías pequeñas, la subestimación de los esfuerzos a los cuales se verán sometidas, el uso de códigos más estrictos en la ubicación histórica de tuberías más grandes debido a su utilización en proyectos de mayor relevancia y el hecho de que tuberías de menor tamaño son más probables a desarrollar bloqueos que influyan en el desempeño estructural.

Por otro lado, un argumento por el cual se explican valores de las tasas de fallas mayores en tuberías más grandes corresponde al incremento del área expuesta a las aguas residuales y el suelo circundante en tuberías de mayor diámetro. Otras consideraciones como la importancia de la criticalidad de tuberías más grandes también ha sido relevante en el momento de estudiar este factor, incluyendo así conceptos de riesgo en sistemas de infraestructura y las consecuencias de la falla de componentes para la identificación y priorización de la rehabilitación de tuberías (Caradot, Sonnenberg, et al., 2017; Laakso, Ahopelto, Lampola, Kokkonen, & Vahala, 2018; Savic et al., 2006).

- Profundidad:

Este factor corresponde a la profundidad a la cual las tuberías de los sistemas de drenaje son instaladas, medida como la profundidad vertical desde la parte superior del tubo y la superficie; y en general, se sabe que los valores mínimos y máximos que puede tomar esta variable están reglamentados por la normativa local de cada país o ciudad. Davies et al. (2001) investiga el efecto de la profundidad en la condición estructural de las tuberías, encontrando mediante la recopilación de investigaciones, que existe una disminución de la probabilidad de ocurrencia de fallas en la medida en que se incrementa la profundidad de instalación; sin embargo, igualmente se sugiere que profundidades mayores a un valor límite resultan por lo contrario en el incremento de esta probabilidad. Berardi et al. (2009) encuentra mediante la aplicación de un modelo de Regresión polinómica evolutiva (EPR) que el incremento de la profundidad de cobertura de las tuberías resulta en un menor efecto de la transmisión de cargas directa desde la superficie. Igualmente, Laakso, Kokkonen, et al. (2018) concluyen que las profundidades entre 2 y 3 metros son las que se encuentran menos asociadas con malas condiciones de las tuberías, considerando la aplicación de dos modelos (regresión logística y arboles aleatorio) en la red analizada en su caso de estudio.

Por otro lado, se encontró que la profundidad de instalación es insignificante en los estudios realizados por Ana et al. (2009), Ariaratnam, El-Assaly, & Yang (2001) y Ahmadi, Cherqui, De Massiac, & Le Gauffre (2015) en los cuales se desarrollaron modelos de regresión logística; al igual que en la investigación llevada a cabo por Salman & Salem (2012), en la cual examinan la significancia de ocho

variables independientes en la probabilidad de falla y encuentran que la profundidad es la única variable no significativa a un nivel de 0,05, aunque si lo es a un nivel de 0,1.

Dados los estudios anteriores, se podría inferir que el análisis de la influencia de esta variable en la determinación de la condición estructural de las tuberías requiere estudiar la distribución de los valores que toma esta variable en un caso de estudio particular; ya que, es posible que al considerar un conjunto de datos que presente poca variabilidad en la profundidad de instalación de las tuberías, se encuentren resultados sesgados a condiciones óptimas de la profundidad. Así, acarreando lo anterior a la conclusión de la insignificancia de la profundidad en el proceso de deterioro de las tuberías.

- Material:

El material de las tuberías es un factor que se ha considerado como influyente en el proceso de deterioro debido a los diferentes mecanismos de falla que se han observado pueden ocurrir debido a la selección inapropiada de los materiales en tuberías de alcantarillado sometidas a ciertas condiciones específicas. Entre las fallas que se pueden presentar se encuentran la deformación de las tuberías debido al uso de tuberías rígidas o flexibles, la abrasión o corrosión de las superficies internas debido a la incompatibilidad de materiales y la presencia de fugas en tuberías debido al uso de materiales que incumplen los estándares requeridos (Stein & Stein, 2004).

Entre los materiales comúnmente encontrados para la construcción de tuberías en redes de alcantarillado son: concreto, plástico, fibrocemento, mampostería, hierro, arcilla vitrificada y gres. Las redes pueden estar constituidas por tuberías de dos o más tipos de materiales diferentes, pero generalmente existe la predominancia de un material en un municipio mientras que los otros tipos de materiales se usan en menor proporción (Stein & Stein, 2004).

En el trabajo realizado por Salman & Salem (2012) encuentran que todos sus cuatro tipos de materiales considerados (mampostería, arcilla, concreto y concreto reforzado) son significativos para la estimación de la probabilidad de falla. Así mismo, en la investigación de Caradot et al. (2017) se utiliza el material como un factor de influencia para la estimación de la condición de las tuberías obteniendo resultados muy acertados a los valores de referencia observados. Sin embargo, Ahmadi et al. (2015) al estudiar los efectos de la construcción de modelos de regresión logística usando únicamente información básica de las tuberías (tamaño, tipo, gradiente, longitud y profundidad) encuentra que a pesar de que el material si es significante en la estimación de la condición estructural, un modelo menos complejo en el cual no se incluye esta variable no difiere significativamente de un modelo en que si se incorpore. En línea con lo anterior, las investigaciones de Ariaratnam et al. (2001) y Baik et al. (2006) mostraron que el tipo de material en la red no constituía una variable significativa en el deterioro de tuberías.

- Longitud:

La longitud de las tuberías corresponde a una de las variables que comúnmente son conocidas o han sido ampliamente registradas en las redes de alcantarillado, por lo cual evaluar el proceso de deterioro considerando el impacto de esta variable permite facilitar el modelamiento predictivo a partir de variables inspeccionadas y con poca incertidumbre. Salman & Salem (2012) encuentran que la longitud de las tuberías resulta significativa para la predicción de fallas al implementar cuatro modelos de regresión logística, encontrando que la alta importancia de esta variable es constante tanto en modelos en los que se utiliza información básica así como modelos que implementen más factores explicativos. Mas específicamente, Laakso, Kokkonen, et al. (2018) obtienen mediante la aplicación de dos modelos de minería de datos que tuberías con una longitud inferior a 40 metros tienden a presentar mejor condición estructural, mientras que es más común que tuberías con una longitud mayor a 60 metros se encuentren en peor condición. Lo anterior es consistente con los resultados de Berardi et al. (2009) según los cuales la tasa de bloqueos en tuberías de alcantarillado es directamente proporcional a la longitud de las mismas.

El anterior comportamiento del deterioro de tuberías relacionado con su longitud es comúnmente explicado por el potencial que tienen tuberías de mayores longitudes a presentar una mayor cantidad de defectos al igual que la presencia de un número mayor de conexiones de servicio que pueden acarrear la generación y propagación de defectos. Así, tuberías más largas tienden a ser clasificadas en una peor condición estructural.

- Material de lecho:

Las tuberías de los sistemas de acueducto y alcantarillado requieren tener un material de lecho apropiado que soporte estructural adecuado y garantizar el desempeño estructural a largo plazo. Este factor es de relevancia al considerar el desempeño estructural pues la efectividad con la cual se transmiten las cargas y presiones que actúan sobre las tuberías es dependiente del tipo de material de lecho sobre el cual se ubiquen estos componentes y el relleno alrededor de las tuberías; siendo una condición ideal que los esfuerzos se distribuyan de manera uniforme alrededor de la tubería (Davies et al., 2001). Sin embargo, a pesar de lo indicado por Davies et al. (2001) en su trabajo respecto a las múltiples investigaciones acerca de la clasificación del material de lecho en factores que permitan cuantificar su efectividad (especialmente en el Reino Unido), pocos estudios en los cuales se estudie la influencia de este factor como una variable predictora de la condición estructural de las tuberías han sido llevados a cabo. Ugarelli, Selseth, Le Gat, Rostum, & Krogh (2013) incluyen esta variable en su modelo de Cadenas de Markov aplicado a un caso de estudio en Oslo, encontrando que la tasa de deterioro incrementa con los tipos de suelo marino y roca de los 4 tipos analizados (depósitos marinos, detritus, relleno y roca).

- Tipo de vía/Tráfico:

Se espera que la ubicación de una tubería afecte la magnitud y el tipo de carga a la que se verá sometida a lo largo de su vida útil. Más aún, en el caso de tuberías que se encuentran instaladas bajo vías, se espera que las principales cargas correspondan a las generadas debido al tráfico en esa vía en particular (Davies et al., 2001).

Este factor fue considerado insignificante en el trabajo de Caradot et al. (2017), similar al resultado encontrado por Laakso, Kokkonen, et al. (2018), en el cual encuentran una conexión débil entre la clase de carretera y la condición de las tuberías. Contrario a esto, diversos autores (Ahmadi et al., 2015; Mashford et al., 2011; Salman & Salem, 2012; Ugarelli et al., 2013) han considerado el tipo de vía para la predicción del estado estructural de las tuberías, considerándola como una variable categórica dependiendo del tipo de vía bajo el cual se encuentran las tuberías o considerando el número de vehículos por día que transitan por una vía. Más específicamente, Mashford et al. (2011) considera el hecho de que una tubería se encuentre o no bajo una vía como una variable predictora binaria para la aplicación de un modelo de minería de datos en el cual se obtiene una precisión del 91% en la predicción de la condición de las tuberías.

La significancia de esta variable en el proceso de deterioro es esperada pues es posible que actúe como un indicador o permita cuantificar las cargas superficiales a las cuales se encuentran sometidas las tuberías a lo largo de su vida útil. Teniendo en cuenta que este resultado no es consistente en todos los casos de estudio, es posible que se deba analizar la significancia de esta variable en los modelos bajo las mismas condiciones de una red particular con el propósito de entender en más detalle el efecto de este parámetro en la tasa de deterioro.

Otros factores externos locales como el tipo de suelo, la ubicación de la tubería, el uso del suelo y la localización o cercanía de árboles se consideran relevantes para la descripción del proceso de deterioro; además de factores de construcción como los métodos de instalación y el nivel de mano de obra (Davies et al., 2001); sin embargo, estos factores son pocas veces incluidos en las investigaciones realizadas hasta el momento ya que la disponibilidad de información para estos factores generalmente se encuentra muy limitada. Igualmente, en investigaciones más recientes, se han considerado nuevos factores como la velocidad o flujo en las tuberías, el número de propiedades conectadas por unidad de longitud y el número de fallas reportadas en tuberías circundantes (Bailey et al., 2015; Berardi et al., 2009; Harvey & McBean, 2014; Jung et al., 2012).

Una de las conclusiones más importantes a partir de la influencia observada de factores en diferentes estudios corresponde a que la influencia de las variables predictoras para el deterioro no se puede generalizar para todos los casos de estudio, pues se observa que las condiciones externas (ambientales, operacionales, etc.) pueden llegar a ser influyentes al relacionar la falla de las tuberías con el incremento o disminución de una variable (en el caso de variables continuas, por ejemplo el diámetro) o con una categoría especifica (en el caso de variables discretas, como por ejemplo el material). Lo anterior puede ser indicador en algunos casos de la redundancia de información que

existe entre las diferentes variables en un caso de estudio en particular, al igual que indicativo de la significancia de diferentes variables en la predicción de diversos mecanismos de falla. Lo anterior indica la necesidad de llevar a cabo futuras investigaciones en las cuales se evalúen los mecanismos de falla de las tuberías al clasificar su estado de deterioro.

4 METODOS DE GESTIÓN DE ACTIVOS EN SISTEMAS DE ALCANTARILLADOS

Teniendo en cuenta la identificación y clasificación de los problemas típicos en sistemas de alcantarillado mencionados en el capítulo anterior (3), es esperado que para evitar las severas consecuencias ocasionadas por los problemas anteriores, las empresas prestadoras del servicio realicen una gestión eficiente de su infraestructura (Carvalho, 2015). La Gestión de Activos de Infraestructura, más conocida como Infrastructure Asset Management (IAM) es un componente crucial de los procesos que deben llevar a cabo las empresas de prestación de servicios para administrar sus activos físicos, su desempeño, riesgos y costos asociados a su ciclo de vida, de manera que estos cumplan sus funciones y garantizar un buen nivel de servicio de forma rentable; para lo cual, se llevan a cabo un conjunto de actividades de administración, financieras, económicas y de ingeniería que se realizan de forma coordinada y sistemática (Rokstad & Ugarelli, 2015). Luego, para efectuar esta gestión, es necesario conocer el estado actual de los sistemas, al igual que el estado que se desea alcanzar, y así evaluar la necesidad o no de llevar a cabo un proceso de rehabilitación, que puede consistir de actividades reactivas (antes de la falla) o proactivas (después de la falla).

Por otro lado, debido a la creciente necesidad de manejar de gestionar de forma eficiente y racional los recursos económicos, las empresas se han visto en la necesidad de realizar la transición de un enfoque de mantenimiento reactivo a uno proactivo. No obstante, cualquiera que sea el tipo de mantenimiento implementado por las empresas prestadoras del servicio, para que este sea efectivo se requiere tener conocimiento de todos los componentes del sistema, su ubicación y su condición. Se han desarrollado diversos procesos en los cuales se sugiere como llevar a cabo este mantenimiento, los cuales pueden ser simples o complejos. En general, estos procesos consisten en: establecer el estado objetivo que se desea alcanzar, identificar los activos y datos disponibles, inspeccionar los activos, analizar los datos recolectados e implementar el proceso de toma de decisiones (Mcdonald & Zhao, 2001).

Una de las aproximaciones para llevar a cabo el mantenimiento de estos sistema se ve sintetizado en la **Figura 4-1**, en la cual se presentan los pasos necesarios recomendados por el Consejo Nacional de Investigación de Canadá (NRC por sus siglas en inglés) para mantener el desempeño de los sistemas de alcantarillado, y es aplicable para el mantenimiento de tuberías y pozos de inspección (Mcdonald & Zhao, 2001).

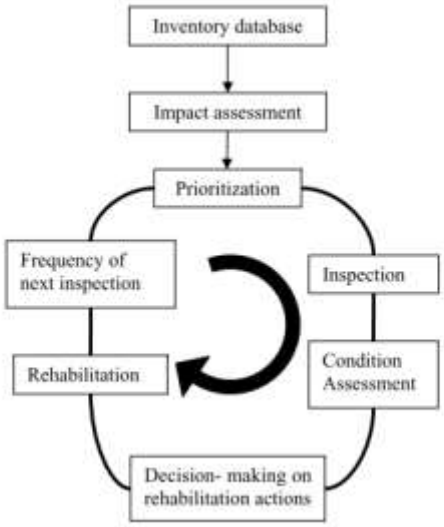


Figura 4-1. Pasos para la gestión de activos de alcantarillado. Tomado de (EPA, 2009).

En esta figura se observa que el enfoque recomendado por la NRC realiza la priorización de los componentes a rehabilitar teniendo en cuenta no solo los factores físicos de la tubería recolectados en el primer paso (Inventory database), sino también la evaluación del impacto que tendría una falla en el componente estudiado (Impact assessment), correspondiendo así a un enfoque para la priorización de activos basado en riesgo relativo (Mcdonald & Zhao, 2001).

Estos procesos son propuestos debido al proceso de toma de decisiones sobre la ubicación y el tiempo en que se debe rehabilitar una tubería es una tarea que requiere considerar una gran cantidad de fuentes de información técnica y operacional del sistema estudiado, e incluso otros sistemas, en algunos casos. Sin embargo, actualmente el enfoque de muchas empresas para llevar a cabo la rehabilitación priorizada de su infraestructura está basado en la consideración de información básica sobre los sistemas y está ampliamente influenciado por la experiencia y la intuición de los encargados de estos procedimientos (Carvalho, 2015; Van Riel et al., 2014a). Por lo anterior, es posible identificar la necesidad de estudiar las actividades requeridas y evaluar los beneficios y limitaciones a las cuales se enfrentan las agencias prestadoras del servicio al efectuar los diferentes tipos de mantenimiento en sus sistemas.

La clasificación comúnmente aceptada de los tipos de mantenimiento distingue tres clases: mantenimiento correctivo, mantenimiento preventivo y mantenimiento predictivo. En general, en los sistemas reales de alcantarillado la gestión de activos se lleva a cabo mediante una combinación de estos tres tipos de mantenimiento, siendo el objetivo la reducción de actividades correctivas y la priorización de actividades preventivas y predictivas. En los siguientes tres subcapítulos se presenta una descripción de estos tres tipos de mantenimiento, haciendo particular énfasis en la descripción del mantenimiento predictivo.

4.1 Mantenimiento correctivo

Este tipo de mantenimiento se realiza únicamente cuando alguno de los equipos o componentes del sistema presenta una falla o cuando se requiere llevar a cabo mantenimiento de emergencia, y corresponde a un enfoque reactivo (New England Interstate Water Pollution Control Commission, 2003). El mantenimiento correctivo consiste en el conjunto de actividades que se deben llevar a cabo para corregir algún problema y restaurar el desempeño de algún componente del sistema (Organizacion Panamericana de la Salud, 2005).

Estos problemas se pueden presentar bajo condiciones normales de operación o debido a situaciones extraordinarias a las cuales se ve sometido el sistema. Por un lado, los problemas comunes que ocurren por las condiciones normales de operación son los bloqueos, rupturas o colapsos de las tuberías, los cuales son fácilmente identificables debido a las consecuencias generadas en el desempeño del sistema. Por otro lado, las situaciones extraordinarias corresponden a eventos generados debido a lluvias de alta intensidad, huracanes, inundaciones o terremotos, y por lo tanto son eventos impredecibles.

La implementación del mantenimiento correctivo por las empresas prestadoras del servicio resulta en (New England Interstate Water Pollution Control Commission, 2003):

- Incapacidad para planificar y programar las actividades
- Incapacidad para realizar un presupuesto adecuado de las actividades requeridas
- Uso ineficiente de los recursos disponibles
- Alta tasa de fallas en equipos y componentes del sistema

4.2 Mantenimiento preventivo

Este tipo de mantenimiento se clasifica como proactivo y se define como el conjunto de actividades programadas y sistemáticas de mantenimiento. En el caso en que el mantenimiento proactivo corresponda a actividades preventivas, estas se realizan con cierta periodicidad, teniendo como base el conocimiento de áreas propensas a problemas específicos, el tiempo de operación de los equipos en diferentes partes del sistema o el paso de cierta cantidad de tiempo. Este tipo de mantenimiento generalmente resulta en la mejora del desempeño del sistema, excepto en los casos en los cuales se encuentren problemas crónicos debido a defectos de diseño o construcción de los componentes y estos no puedan corregirse (New England Interstate Water Pollution Control Commission, 2003).

Para que sea posible llevar a cabo una buena estructuración de actividades de mantenimiento preventivo, se requieren los siguientes elementos principales (New England Interstate Water Pollution Control Commission, 2003):

- Planificación y programación de actividades
- Contar con un Sistema de mapeo o Sistema de Información Geográfica (SIG) de la red
- Inventario y gestión de activos
- Gestión de la información registrada
- Gestión de la infraestructura de repuesto
- Control de costos y presupuesto
- Procedimientos de reparación en casos de emergencia
- Programas de entrenamiento de personal

Algunas de las estrategias de mantenimiento que se llevan a cabo como parte del mantenimiento preventivo corresponden a actividades de limpieza y labores hechas por los usuarios del sistema considerando recomendaciones para la obstrucción de los colectores. Algunas de las estrategias de la primera categoría son: limpieza de la trampas de grasas, mantenimiento de tanques interceptores y limpieza de los colectores; mientras que las estrategias de la segunda categoría son primordialmente buenas prácticas de la disposición de residuos en estos sistemas (Organizacion Panamericana de la Salud, 2005).

Entre los beneficios que se generan debido a la implementación de un enfoque de mantenimiento preventivo se encuentran (New England Interstate Water Pollution Control Commission, 2003):

- El mantenimiento puede ser planeado y programado
- Se puede identificar el trabajo atrasado
- Es posible presupuestar los recursos requeridos para llevar a cabo las actividades de mantenimiento
- Se hace un uso eficiente los recursos humanos y materiales

4.3 Mantenimiento predictivo

El mantenimiento predictivo también se clasifica como proactivo y corresponde a un método en el cual se realiza el mantenimiento de forma planificada y programada teniendo en cuenta la predicción de las fallas encontrada a partir de la observación del comportamiento del sistema durante un periodo de tiempo. Para esto, se establecen datos base del desempeño (estructural u operacional) como referencia, se monitorean los criterios de desempeño seleccionados durante un tiempo específico y se observan los cambios que se producen en el sistema para realizar la predicción de los componentes con mayor probabilidad de falla (New England Interstate Water Pollution Control Commission, 2003). El proceso descrito anteriormente, se encuentra en general, alineado con el enfoque sistemático que varios autores sugieren se requiere en la implementación de un mantenimiento proactivo en que primero se evalúe la condición, desempeño y capacidad de cada componente de la red; en segundo lugar se realicen predicciones del estado futuro de estos

componentes durante su ciclo de vida; y por último, se determine la prioridad de las intervenciones antes de los eventos de falla (EPA, 2009).

En general, se espera que la operación y mantenimiento de los sistemas se realice más eficientemente y con mayor facilidad al tener un monitoreo continuo del desempeño del sistema. Sin embargo, realizar esta transición desde un enfoque reactivo a uno proactivo puede representar un reto para las empresas prestadoras del servicio puesto que si históricamente se han realizado actividades correctivas para su mantenimiento, la mayoría de sus recursos se dirigen a estas actividades y se enfrentan con dificultades para enfocarse en acciones preventivas y predictivas (New England Interstate Water Pollution Control Commission, 2003).

No obstante, la realización de un buen mantenimiento predictivo permite gestionar los recursos humanos y materiales de la forma más efectiva, así mismo como mantener altos niveles de servicio a los usuarios del sistema. Algunos de los beneficios que se obtienen al implementar esta transición son (New England Interstate Water Pollution Control Commission, 2003):

- Garantizar la disponibilidad de infraestructura y equipos según este previsto
- Mantener la confiabilidad de las instalaciones al nivel de las condiciones de diseño
- Mantener el valor de la inversión realizada, teniendo en cuenta que los sistemas de alcantarillado representan uno de los activos de mayor inversión de capital de las ciudades.
- Aprovechar al máximo los componentes del sistema a lo largo de su vida útil
- Contar con justificaciones técnicas para la inversión de recursos financieros teniendo en cuenta la recolección de información y datos precisos.
- Disminuir los costos de mantenimiento de las redes, considerando que las actividades de mantenimiento preventivo y predictivo evitan la reparación de fallas de mayor magnitud en el sistema que pueden llegar a implicar mayores costos tanto por el reemplazo de los componentes como por los impactos sociales y ambientales que estas fallas implican.

Así, para dar respuesta a las necesidades del modelamiento predictivo en sistemas de alcantarillado, diversos modelos de deterioro se han propuesto en la literatura para modelar el proceso de deterioro con base en condiciones observadas en las redes de alcantarillado y los factores que influencian este proceso (Caradot et al., 2014). Estos modelos corresponden a una especificación de una relación matemática o probabilística que existe entre diferentes variables (Grus, 2015), y para el caso de sistemas de drenaje urbano se pueden categorizar en modelos a nivel de grupos de tuberías (cohortes) o modelos a nivel de tuberías (Ana & Bauwens, 2010).

Si se consideran los modelos de deterioro a nivel de tuberías, se debe tener en cuenta que la clasificación de las tuberías por su estado obtenida en el proceso de inspección corresponde a los datos que se requieren como información de entrada al modelo para realizar la predicción del estado de los componentes no inspeccionados o el estado futuro de los componentes ya

inspeccionados. Rokstad y Ugarelli (2015), plantean la descripción matemática general en la que se basan los modelos de deterioro a nivel de tuberías para predecir la condición de las redes:

Primero, se debe considerar la situación en que un sistema tiene un conjunto, S, que contiene N tuberías de drenaje, de las cuales se tiene un subconjunto R que contiene n tuberías que han sido inspeccionadas y se ha determinado su condición. Luego, el tiempo de instalación para cada tubería i es $t_{i,inst}$. Cada tubería j en el subconjunto R se inspeccionó en el tiempo $t_{i,insp}$ obteniendo como resultado la clasificación y_i que puede corresponder a cualquier clasificación c de CC (Condition Classes). Finalmente, se considera que un vector $\boldsymbol{Z_j}$ contiene la información relacionada con la variabilidad de las CCs de la red, y que $\boldsymbol{Z_j}$ puede usarse como los factores explicativos del modelo (covariables de entrada).

Entonces, un modelo de deterioro de redes de alcantarillado f utiliza las observaciones R para predecir la condición de cualquier tubería en S en un tiempo T, ya sea en términos determinísticos (tasa de falla) o probabilísticos (probabilidad de falla), de manera que:

$$S = \{\boldsymbol{Z_i}, i = 1, 2, \dots N\},$$
$$R = \{O_j, j = 1, 2, \dots n\}, R \subset S \qquad \text{Ecuación 4-1}$$
$$O_j = \{i, \boldsymbol{Z_j}, t_{j,insp}, y_j\}, y_j \in \{1, 2, \dots c\}$$

En donde, O_j corresponde a la observación del estado de cada tubería j inspeccionada en el subconjunto R.

Con lo cual, el resultado del modelo corresponde a $\hat{y}_i(T)$: la predicción de la CC para la tubería i en el tiempo T:

$$\hat{y}_i(T) = f(\boldsymbol{Z_i}, T|R) \vee \Pr(\hat{y}_i(T)) = f(\boldsymbol{Z_i}, T|R) \qquad \text{Ecuación 4-2}$$
$$\forall\, i \in S, \hat{y}_i \in \{1, 2, \dots c\}$$

Ahora bien, existen tres tipos principales de modelos para aproximarse al problema de modelación del deterioro de redes de alcantarillado: modelos determinísticos, estadísticos y de aprendizaje automático (Caradot et al., 2014). Los modelos determinísticos buscan entender los diferentes mecanismos físicos que generan el deterioro de las tuberías, pero generalmente incluso los modelos más complejos son demasiado simples para modelar el complejo proceso de deterioro. Los modelos estadísticos usan relaciones matemáticas para relacionar la caracterización histórica del estado de las tuberías con factores de deterioro de las tuberías y sus resultados se expresan como valores de probabilidad. Los modelos de aprendizaje automático permiten identificar relaciones complejas no lineales entre las entradas y las salidas, sin limitarse a una expresión predefinida vinculando las variables de entrada y los resultados obtenidos (Caradot et al., 2014).

En los siguientes cuatro subcapítulos, se presenta una breve descripción de algunas herramientas comúnmente utilizadas para apoyar el proceso de toma de decisiones, algunas metodologías basadas en la modelación hidráulica, los modelos estadísticos y los modelos de aprendizaje automático comúnmente utilizados.

4.3.1 Herramientas para el apoyo a la toma de decisiones

Actualmente existen diversas herramientas de apoyo a la decisión en el contexto de la gestión de activos de infraestructura que buscan realizar un manejo eficiente de la información analizada y mejorar la precisión y legitimidad de las decisiones. En el trabajo realizado por Ana & Bauwens (2007), los autores realizan una recopilación de los enfoques de las principales herramientas utilizadas al igual que sus alcances y limitaciones en la aplicación de las mismas en las diferentes etapas de la gestión de activos de infraestructura como se observa en la **Figura 4-2.**

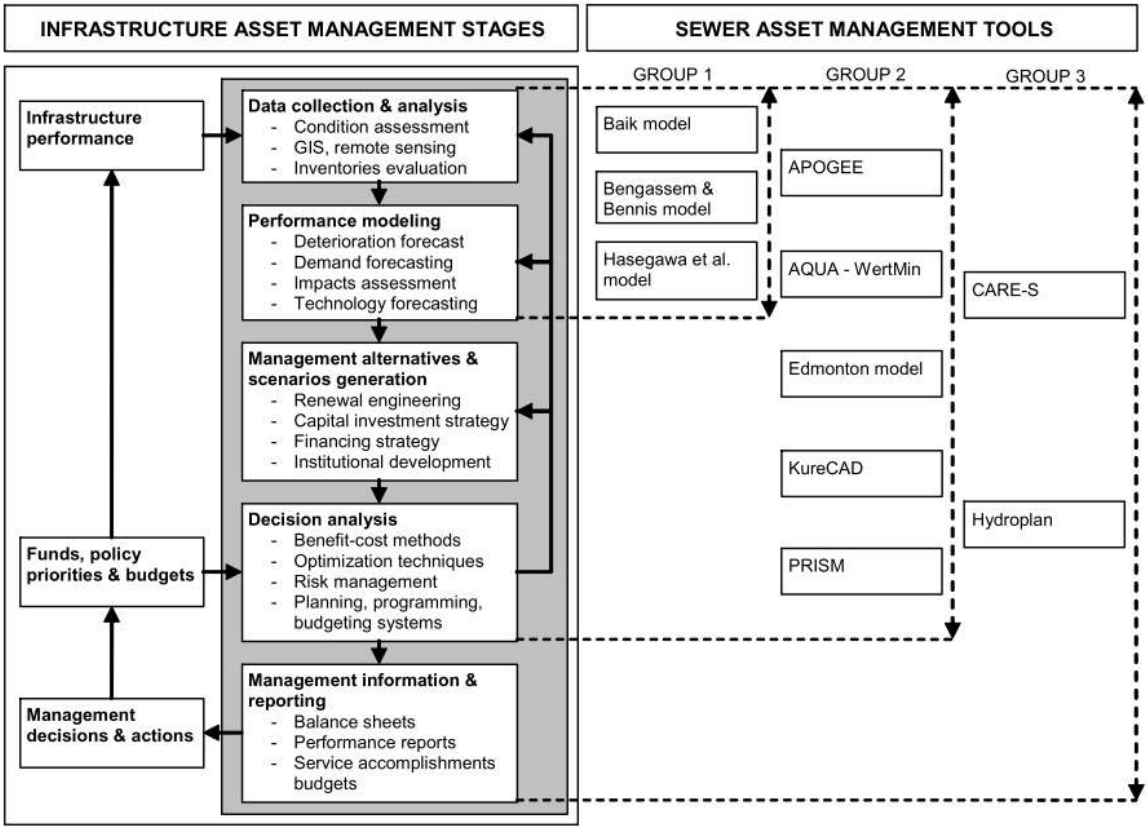


Figura 4-2. Aplicabilidad de las diferentes herramientas de gestión de alcantarillado a las etapas comunes del proceso de gestión de infraestructura. Tomado de Ana & Bauwens (2007)

Así, estas herramientas se clasifican en tres grupos, de acuerdo a su capacidad de utilizar información de diferentes fuentes y proporcionar la evaluación de diferentes escenarios para tomar

las decisiones más apropiadas para la gestión de los sistemas. A continuación se presenta una breve descripción de estos modelos clasificados por su capacidad (Ana & Bauwens, 2007):

Las herramientas categorizadas en el grupo 1 corresponden a aquellas que se enfocan principalmente en la modelación del desempeño de los sistemas, para que los encargados de los procesos de mantenimiento y rehabilitación en las empresas prestadoras del servicio puedan tomar decisiones respecto a las actividades requeridas. Entre estas herramientas se encuentran: el modelo de Cadenas de Markov para la predicción de la futura condición (estructural y operacional) de tuberías planteado por Baik et al. (2006); el modelo de inferencia difusa de Bengassem y Bennis (2000) para la predicción de la condición de tuberías considerando la inspección del estado estructural y modelación hidráulica; y el modelo de Hasegawa et al. (1999) que realiza una clasificación de su condición considerando tanto factores físicos de la tubería, los resultados de inspecciones mediante CCTV y las condiciones externas del ambiente (Ana & Bauwens, 2007).

Por otro lado, las herramientas categorizadas en el grupo 2 corresponden a aquellas que realizan tanto la modelación del desempeño de los sistemas como el análisis de las decisiones que se pueden tomar para la rehabilitación del sistema, considerando tanto los costos . Entre estas herramientas se encuentran los sistemas APOGEE, AQUA-WertMin, los modelos Edmonton, KureCAD y PRISM. De acuerdo con lo encontrado por Ana & Bauwens (2007), a continuación se presenta una breve descripción de cada una de estas herramientas:

- La herramienta APOGEE corresponde a un sistema de apoyo a la decisión desarrollado en Francia, que busca optimizar la planeación y rehabilitación de las redes de alcantarillado considerando tres componentes básicos: una base de datos, un sistema experto y un módulo de planeación. El segundo componente de este sistema corresponde al principal, y es en donde se realiza el diagnóstico del estado de la red con base en los datos ingresados en el primer componente, considerando datos hidrológicos e hidrogeológicos, cargas excesivas sobre la red, abrasión y agresividad del flujo, flujo presurizado en las tuberías y los métodos de construcción históricos.
- En segundo lugar, la herramienta AQUA-WerMin corresponde a un software desarrollado en Alemania para proveer ayuda a las agencias prestadoras del servicio con la planificación de inspecciones mediante CCTV y estrategias de renovación y construcción. Esta herramienta usa como principal modelo la aplicación de la distribución de Herz, la cual permite calcular la transición de una condición de las tuberías a otra que representa un peor estado a medida que pasa el tiempo, teniendo en cuenta 6 clasificaciones posibles. Así, es posible conocer el proceso de deterioro de las tuberías, las actividades de rehabilitación requeridas en el sistema, y un análisis de costos de las diferentes estrategias de rehabilitación que es posible llevar a cabo.
- En tercer lugar, los modelos Edmonton desarrollados en Canadá corresponden a tres modelos que realizan simulación basadas en reglas y análisis de probabilidad con el fin de

ayudar a la ciudad Edmonton a planificar sus gastos de mantenimiento de redes de alcantarillado, considerando únicamente la condición de las tuberías. Estos tres modelos constituyen la secuencia de actividades que se deben realizar para predecir el estado de las tuberías y realizar un análisis de los costos considerando el método de reparación y su costo. Así, el primer modelo determina la condición actuar de las tuberías con base en sus características físicas y su probabilidad de existencia (APE) mediante simulaciones de Montecarlo; el segundo modelo utiliza la teoría de Markov para predecir el estado futuro de las tuberías dentro de cinco años; y el tercero modelo pronostica los costos actuales y futuros de la renovación considerando los resultados de los dos primeros modelos.

- En cuarto lugar, KureCAD (Finlandia) corresponde a una herramienta basada en sistemas de información geográfica (SIG) con la cual es posible almacenar información de los activos de alcantarillado, priorizar la rehabilitación de las tuberías de alcantarillado y proveer información para la implementación de planes de rehabilitación. En este programa se requieren tres tipos de información básica: condición estructural, condición funcional/operacional y la tasa de fugas. A estas tres clasificaciones se les asigna un puntaje (1-4) y el programa se encarga entonces de calcular un índice ponderado que expresa la condición de las tuberías y lo presenta mediante el SIG.
- Finalmente, la herramienta PRISM desarrollada en Canadá es un programa que busca priorizar la rehabilitación de tuberías de alcantarillado considerando limitaciones de presupuesto establecidas para un futuro. PRISM se enfoca en minimizar los gastos de capital en un tiempo futuro de planificación considerando presupuestos anuales y la asignación de procesos de rehabilitación a tuberías de clases más importantes con el uso de programación lineal.

Por último, se encuentran las herramientas CARE-S (Europa) y Hydroplan (Bélgica) que clasifican dentro del grupo 3 de herramientas, en el cual se ubican las herramientas que permiten dar apoyo a todas las etapas del proceso de gestión de activos de infraestructura. La herramienta CARE-S tiene el objetivo de garantizar que la tubería adecuada sea rehabilitada en el momento correcto en el tiempo mediante la aplicación de la tecnología más apropiada. Esta herramienta consta de 4 pasos en las cuales se realiza una planeación inicial de los criterios de desempeño, se realiza un diagnóstico del estado actual de las redes, se desarrollan posibles escenarios para dar solución a los problemas y se monitorea la implementación de la solución más apropiada considerando modelación hidráulica y medidas de desempeño. De manera similar, Hydroplan consiste en una herramienta que tiene un enfoque integrado para la gestión estructural, hidráulica y ambiental de los elementos de las redes de alcantarillado.

De acuerdo a lo anterior, es esperado que dependiendo del alcance que tiene cada una de las herramientas mencionadas anteriormente, estas requieran diferentes niveles de información desde diversas fuentes para apoyar el proceso de toma de decisiones en la rehabilitación. Ana & Bauwens

(2007) muestran una recopilación de los requerimientos de cada una de estas herramientas, como se observa en la Figura 4-3. En esta, es posible identificar que las herramientas del grupo 3 son las que requieren la mayor cantidad de información, como es esperado; al igual que es posible identificar ciertos parámetros que se requieren transversalmente desde las herramientas más sencillas hasta las más complejas. Entre estos parámetros se encuentran el material, la edad, la longitud, la profundidad y la pendiente de las tuberías, al igual que información del tipo del suelo y los datos de inspección de la condición de las tuberías.

	TOOLS	DATA DESCRIPTION																								
		Pipe material	Pipe age	Pipe length	Pipe diameter	Pipe thickness	Pipe shape	Depth of pipe	Type of pipe	Pipe roughness	Pipe slope	Pipe location	Condition data	Defects/ History	Flow data	Leakage rates	Soil data	Land use	Traffic/ Road	Tree location	Groundwater data	Population	Geotechnical	Type of joints	Economic data	Rehab tech/cost
Group 1	Baik	x	x	x	x			x	x		x		x		x	x	x		x		x					
	Bengassem & Bennis			x	x			x			x	x	x	x	x		x						x			
	Hasegawa et al.	x	x	x	x			x			x		x		x	x	x		x		x				x	
Group 2	APOGEE	x	x	x	x		x	x	x	x	x	x	x	x		x	x	x	x		x		x				x
	Aqua-Wertmin	x	x	x	x		x		x		x	x	x	x		x	x		x					x	x	x
	Edmonton	x	x	x	x			x					x				x						x			x
	KureCAD	x	x	x	x						x	x	x	x		x	x		x	x				x	x	x
	PRISM	x	x	x	x			x	x				x													x
Group 3	CARE-S	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x
	Hydroplan	x	x	x	x		x	x	x		x	x	x		x		x	x	x	x		x				x

Figura 4-3. Información de entrada relevante para las diferentes herramientas de apoyo a la decisión en la gestión de redes de alcantarillado. Tomado de (Ana & Bauwens, 2010)

En general, se pueden resaltar algunas características importantes de estas herramientas que permiten reconocer los requerimientos del registro y manejo de la información, al igual que las necesidades computacionales a las que se enfrentan las agencias prestadoras del servicio.

En primer lugar, se observa que todas las herramientas tienen la capacidad de almacenar grandes cantidades de información, usando generalmente sistemas de información geográfica. Por otro lado, la mayoría de las herramientas tiene el análisis del desempeño de las redes y la predicción del proceso de deterioro como fundamento para la toma de decisiones. Así mismo, estas herramientas tienen típicamente una estructura compleja, lo cual puede dificultar su aplicación en empresas prestadoras del servicio más pequeñas (Ana & Bauwens, 2007).

Así, se identifica la necesidad de investigar y evaluar la aplicabilidad de herramientas menos rígidas y menos complejas que permitan apoyar la gestión de sistemas de alcantarillado mediante los datos

usualmente disponibles o que se puedan recolectar a partir de los métodos más usados de inspección de redes.

4.3.2 Modelación hidráulica e indicadores de servicio

Otro de los enfoques utilizados para la priorización de rehabilitación de activos de alcantarillado corresponde al uso de modelos hidráulicos e indicadores de servicio para la estimación del comportamiento de las tuberías bajo ciertos escenarios. En estos casos, se considera la utilidad de los modelos hidráulicos existentes de las redes de alcantarillado para simular diversas condiciones de servicio (problemas comunes en las tuberías) e interpretar la respuesta hidráulica de las tuberías, buscando la identificación de tubos propensos a las fallas y/o inundaciones. No obstante, este enfoque no ha sido ampliamente investigado, pues pocas investigaciones como las de Arthur & Crow (2007); Arthur, Crow, & Pedezert (2008); Arthur, Crow, Pedezert, & Karikas (2009) y Duncan & Arthur (2005).

Entre los anteriores estudios, la metodología planteada inicialmente por Duncan & Arthur (2005) y mejorada en Arthur & Crow (2007), corresponde a una de las primeras en que se considera este enfoque para realizar el mantenimiento proactivo de las activos de alcantarillado. En estos estudios, desarrollan una técnica para el establecimiento de las localizaciones en donde se requiere un mantenimiento proactivo bajo un escenario en el que se cuenta con poca información histórica de las fallas en las redes con el propósito de minimizar el riesgo de pérdida de servicio. El estudio se desarrolla considerando una pequeña cuenca costera en Edimburgo (122 ha y 16250 personas), en la cual se contara con un modelo hidrodinámico y existiera un registro de quejas de los usuarios de la red. La metodología planteada se puede resumir en 5 pasos, presentados en la **Tabla 4-1**.

Tabla 4-1. Metodología para el mantenimiento proactivo de activos en redes de alcantarillado. Adaptado de (Arthur & Crow, 2007; Duncan & Arthur, 2005)

Paso	Descripción	Resultado
1	Definición de los modos de falla de las tuberías	Modos de falla más probables en la red considerando el alcance del proyecto: desbordamientos, obstrucciones y colapsos
2	Identificar los nodos problemáticos en la red al simular los modos de falla en el modelo hidrodinámico	Base de datos de nodos donde pueden presentarse inundaciones debido a obstrucciones.
3	Estimar la severidad de las consecuencias de fallas	Puntaje de la severidad de las consecuencias considerando
4	Calcular la probabilidad de falla mediante la ponderación de factores	Puntaje de la probabilidad de falla mediante un cálculo multicriterio

Paso	Descripción	Resultado
5	Realizar la priorización de tuberías a rehabilitar considerando las consecuencias y probabilidad de falla de cada una	Tuberías con mayores probabilidad de falla o

Los resultados de la aplicación de la metodología anterior, permitieron a los autores establecer que no existe una relación espacial directa entre los nodos problemáticos encontrados en una red y las quejas reportadas por los clientes, por lo cual el mantenimiento proactivo de zonas para la renovación con base en quejas de los clientes puede tener resultados sesgados. Más aún, consideran que este indicador de servicio puede ser problemático debido a factores socioeconómicos, la dificultad de relacionar el número de quejas con la severidad de una falla, al igual que con la distribución espacial del activo correcto. En una aproximación final a esta metodología, Arthur et al. (2009), establecen una metodología detallada con la cual es posible detectar las tuberías problemáticas en una red mediante la simulación hidráulica, teniendo en cuenta el caudal máximo de las tuberías y la profundidad de flujo máxima respecto a las condiciones de sobreflujo. Con lo anterior, destacan la posibilidad de transferir esta metodología a otros casos de estudio y la ocurrencia de una detección de tuberías problemáticas en un 80% de los casos. Sin embargo, entre sus principales limitaciones encuentran la gran cantidad de tiempo que requieren las diferentes actividades a realizar en su metodología y la capacidad de aplicación a redes pequeñas (Arthur et al., 2009).

Por otro lado, entre sus principales aportes, Arthur & Crow (2007) resaltan la importancia de implementar enfoques cuantitativos y cualitativos para la minimización del riesgo de fallas, resaltando que utilizar únicamente uno de los dos puede resultar en la pérdida de ventajas significativas; en el caso de los enfoques cuantitativos, se puede esperar que estos sean más objetivos y no dependan significativamente del conocimiento especializado de los sistemas, mientras que, los enfoques cualitativos, a pesar de depender en alto grado de las habilidades y criterio de los tomadores de decisiones y ser dependientes de la zona de estudio, estos pueden contribuir al entendimiento del comportamiento de los sistemas cuando no existen grandes cantidades de datos disponibles.

Adicionalmente, Arthur et al. (2008) estudiaron la influencia de factores específicos en la generación de obstrucciones en redes de alcantarillado, con el propósito de realizar mantenimiento proactivo a tuberías que pueden encontrarse en un estado subcrítico. La metodología establecida por los autores se puede sintetizar en los siguientes tres pasos, en un caso de estudio al Sur de Inglaterra en el cual se obtuvieron registros de quejas de 10 empresas prestadoras del servicio.

1. Recopilar información de quejas de clientes y la acción realizada en cada evento (periodo de 4 años)

2. Clasificar la ocurrencia de bloqueos en tuberías distinguiéndolos por diferentes factores hidráulicos
3. Analizar la significancia que tienen los factores hidráulicos considerados como indicadores de una mayor probabilidad a la ocurrencia de obstrucciones.

La significancia de los factores se estudia mediante la prueba estadística χ^2 y los resultados encontrados para cada factor se presentan en la Tabla 4-2.

Tabla 4-2. Factores considerados como indicadores para la ocurrencia de obstrucciones. Adaptado de (Arthur et al., 2009)

No.	Factor	Significativa?
1	Tipo de alcantarillado	Si
2	Estado de sobrecarga	No
3	Riesgo de inundación	Si
4	Efecto de remanso	No
5	Confluencia de flujos	No
6	Velocidad de autolimpieza	Si
7	Densidad poblacional	Si
8	Tamaño de la tubería	N/A

A partir de sus resultados concluyen como analizar los registros de eventos de falla reportados puede ayudar a identificar factores que influencian la ocurrencia de obstrucciones en redes de alcantarillado, al igual que establecer estimaciones del incremento en la probabilidad de que ocurran obstrucciones en las tuberías al calcular entre las tasas de falla cuando un factor está presente y cuando no lo está (Arthur et al., 2008).

4.3.3 Modelos estadísticos

Los modelos estadísticos relacionan la información histórica obtenida a partir de la inspección de los sistemas de alcantarillado con el proceso de deterioro de las tuberías. Estos métodos comúnmente son subdivididos en modelos a nivel de grupos de tuberías o modelos aplicables a nivel de tuberías. Los primeros consideran redes enteras o cohortes que presentan características similares que afectan su deterioro para así estimar el proceso de deterioro de tuberías que se consideren homogéneas; mientras que los segundos evalúan las características de las tuberías de forma individual para establecer las variables que serán relevantes en su proceso de deterioro (Ana & Bauwens, 2010). Entonces, teniendo en cuenta las características del problema de predicción del proceso de deterioro de las tuberías en el tiempo, se consideran que métodos como los modelos de Markov o los modelos de supervivencia de cohortes pueden ser apropiados para la modelación de la condición en redes de alcantarillado (Caradot et al., 2014).

A continuación se presenta una breve descripción de los principales modelos estadísticos reportados en la literatura aplicados en el problema de deterioro de redes de alcantarillado (Ana & Bauwens, 2010):

1. Cadenas de Markov: Este modelo representa proceso estocástico que ocurre en un tiempo discreto, en el cual la probabilidad condicional de que una tubería se encuentre en un estado futuro en un tiempo $t + \Delta t$ se encuentra determinada por la condición actual de la tubería. Así, asume que es posible describir el proceso de deterioro considerando la información que se tiene en el estado actual. Los cambios en los estados de la tubería se denominan transiciones y existe una probabilidad asociada para cada transición posible, conformado así una matriz de transición P de tamaño $m \times m$, donde m es el número de estados posibles. Estas probabilidades de transición pueden ser independientes del tiempo, en cuyo caso se tiene un modelo homogéneo de Markov, o pueden ser dependientes del tiempo, teniendo así un modelo no homogéneo de Markov. Generalmente se asume para el caso del deterioro de tuberías en redes de alcantarillado, que no es posible que la condición de las tuberías mejore en el tiempo, por lo cual la matriz de transición comúnmente utilizada es la que se presenta en la Ecuación 4-3

$$P^{t,t+1} = \begin{bmatrix} P_{11}^{t,t+1} & P_{12}^{t,t+1} & \dots & P_{1m}^{t,t+1} \\ 0 & P_{22}^{t,t+1} & \dots & P_{2m}^{t,t+1} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & P_{mm}^{t,t+1} = 1 \end{bmatrix}$$

Ecuación 4-3

2. Modelo de supervivencia de cohortes: En este modelo se observa el comportamiento las tuberías de un cohorte a medida que su condición atraviesa los diferentes estados (desde un buen estado cuando las tuberías están nuevas) hasta el peor (cuando se presenta la falla) a lo largo de su vida útil. Mas aún, el número de años en que una tubería se encuentra en una condición específica tiene una probabilidad específica y la transición de las tuberías desde una condición a otra se representa mediante las funciones de transición, que está dada generalmente por la distribución de Herz. Estas funciones de transición deben ser calibradas para lo cual se requiere como mínimo información de: el año de instalación, el año de la última inspección, la condición de la tubería resultante de la inspección y la longitud de la tubería, pero es posible incorporar otros factores influyentes en el modelo.

3. Regresión logística: Este modelo, a diferencia de los anteriores, permite analizar las relaciones entre las variables independientes (factores de deterioro) y la variable dependiente (condición de las tuberías), y corresponde a la categoría de métodos estadísticos denominados modelos lineales generalizados. Este modelo asume que la variable dependiente y es categórica y dependiente de un conjunto de variables $\vec{X}$; sin embargo, el valor obtenido de y corresponde a una probabilidad de corresponder a una

clase o a otra dependiendo de los valores que tome y. En el caso en que solo existan dos categorías para realizar la clasificación, el modelo se denomina regresión logística binaria; mientras que si son más de dos se conoce como regresión logística multinomial. En el capítulo 5.1 de este documento se explica el modelo de regresión logística binaria en más detalle.

En la **Tabla 4-3** se presenta una recopilación de los modelos estadísticos implementados por diferentes autores hasta el momento y sus respectivos resultados obtenidos en la modelación del deterioro de tuberías en redes de alcantarillado. En esta se observa que uno de los modelos comúnmente utilizados corresponde a las cadenas de Markov, lo cual puede resultar esperado debido a su capacidad de predecir el estado futuro de las tuberías de las redes en el tiempo y no solo el estado actual de un conjunto de tuberías no inspeccionadas.

Tabla 4-3. Resumen de modelos estadísticos para la modelación del deterioro estructural en tuberías de redes de alcantarillado. Adaptado de Ana & Bauwens (2010).

Metodología	Metodología (inglés)	Referencias	Resultados principales
Modelo de supervivencia de cohortes	Cohort survival model	Baur et al. (2004)	- Tiempo de vida restante esperado para la tubería - Proporción esperada de tuberías en las diferentes clases de acuerdo a su condición
Cadenas de Markov	Markov chain	Wirahadikusumah et al. (2001)	Condición esperada de un grupo de tuberías
		Baik et al. (2006)	Vector de la condición esperada de una tubería individual
		Micevski et al. (2002)	Vector de la condición esperada de un grupo de tuberías
		Le Gat (2008)	Vector de la condición esperada de una tubería individual
		Caradot et al. (2017)	Vector de la condición esperada de una tubería individual
		Ugarelli et. Al (2013)	Vector de la condición esperada de una tubería individual
Semi-Markov	Semi-Markov	Kleiner (2001)	Vector de la condición esperada de un grupo de tuberías
Modelo de regresión logística	Logistic regression model	Ariaratnam et al. (2001)	Probabilidad de falla de una tubería individual
		Wright et al. (2006)	Probabilidad de falla de una tubería individual
		Salman & Salem (2012)	Probabilidad de falla de una tubería individual
		Ahmadi et al (2015)	Probabilidad de falla de una tubería individual

Metodología	Metodología (inglés)	Referencias	Resultados principales
Análisis discriminante múltiple		Tran et al. (2006)	Condición de una tubería individual

Es importante mencionar que los anteriores modelos son de gran utilidad pero generalmente todos se enfrentan al problema de la disponibilidad de la gran cantidad de datos que se requieren para construir modelos robustos y confiables con capacidad aceptable de predicción. Por otro lado, este tipo de modelos en particular pueden tener dificultades para encontrar relaciones más complejas entre los datos de entrada y las variables de salida debido a las restricciones existentes en la estructura de los modelos. En particular, los modelos de supervivencia de cohortes que requieren particionar los datos en grupos de tuberías de características similares pueden presentar problemas debido a que solo se puede encontrar un número limitado de grupos de un conjunto de datos pequeño; igualmente, en términos del costo de cálculos requeridos, los modelos de grupos de tuberías como los modelos semi-Markov pueden ser más tediosos de desarrollar en términos computacionales debido a que no consideran directamente variables predictoras relacionadas con las tuberías (Ana & Bauwens, 2010).

4.3.4 Modelos de aprendizaje automático

Estos modelos, también conocidos como modelos de "machine learning", a diferencia de los modelos estadísticos, permiten identificar relaciones no-lineales complejas entre los datos de entrada (que corresponderían en este caso a los factores de deterioro) y la variable de salida (estado de las tuberías de alcantarillado) (Caradot et al., 2014). Como su nombre lo indica, los métodos que se encuentran dentro de esta clasificación son considerados capaces de aprender patrones o relaciones a partir de los datos de entrada, mediante alguno de los tipos de aprendizaje con los que se puede entrenar una máquina.

Los algoritmos de aprendizaje automático pueden clasificarse en tres categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado (Lab41, 2019). Las técnicas de aprendizaje supervisado se utilizan cuando se requiere aprender la relación entre atributos independientes y un atributo dependiente que ya se encuentra designado (las clases); mientras que las técnicas de aprendizaje no supervisado agrupan instancias sin que haya un atributo dependiente especificado previamente (Kohavi & Provost, 1998).

Entre los algoritmos de aprendizaje supervisado se encuentran: arboles de decisión, el clasificador naive bayes, regresión logística, redes neuronales y las máquinas de soporte vectorial (Lab41, 2019). En el caso del problema del deterioro estructural de tuberías en redes de alcantarillado, los

principales tipos de modelos utilizados son las redes neuronales, bosques aleatorios y máquinas de soporte vectorial (Caradot et al., 2017).

5 MINERÍA DE DATOS APLICADA A FALLAS EN REDES DE ALCANTARILLADO

La minería de datos comprende la extracción de información de un conjunto de datos y su transformación en estructuras que sean comprensibles. Así, corresponde al proceso computacional de descubrir patrones en grandes conjuntos de datos mediante la intersección de métodos de aprendizaje automático, estadísticos y de sistemas de bases de datos (Gupta, Rawat, Jain, Arora, & Dhami, 2017). Este proceso debe ser automático o por lo menos semiautomático, como lo es más comúnmente, y los patrones encontrados deben ser significativos, en el sentido que dirigen hacia el descubrimiento de información no conocida (Witten, Frank, Hall, & Pal, 2017).

En muchos casos, el término minería de datos es usado indistintamente para hacer referencia a todo el proceso de extracción de conocimiento (Knowledge Discovery in Database – KDD) o se hace referencia a la minería de datos como uno de los pasos que se debe implementar para llevar a cabo KDD. En general, el proceso consta de diversos pasos que van desde la recolección y almacenamiento de la información hasta la interpretación y validación del conocimiento obtenido, como se observa en la **Figura 5-1**.

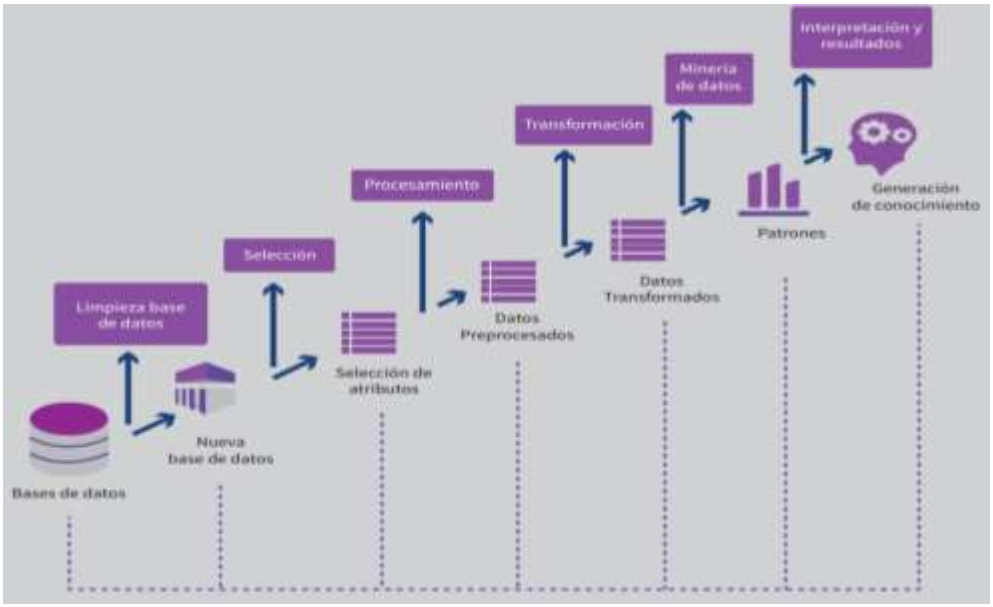


Figura 5-1. Etapas del proceso de extracción de conocimiento (KDD). Tomado de (UIAF, 2014)

Por otro lado, cuando se abordan problemas desde el campo de la minería de datos existen 5 clases de tareas o clases de minería de datos que se pueden implementar, siendo estas: detección de anomalías, aprendizaje de reglas de asociación, agrupación (clustering), clasificación y regresión. La clasificación es una función de la minería de datos en la que los algoritmos asignan elementos de

una colección a categorías o clases predeterminadas, y los métodos comúnmente utilizados para esta tarea son: arboles de decisión, redes bayesianas, redes neuronales y máquinas de soporte vectorial (Gupta et al., 2017).

Ahora bien, al considerar la aplicación de la minería de datos al problema de gestión de activos en redes de alcantarillado es posible identificar que, dado el tamaño de las redes y la gran cantidad de información que usualmente se registra sobre las mismas, este escenario presenta una oportunidad para la aplicación de técnicas de minería de datos para la identificación de patrones en el comportamiento del sistema y así, un mejor entendimiento del mismo (Bailey et al., 2015). En particular, es posible aproximarse al problema de la predicción de la condición de las tuberías en los sistemas de alcantarillado, teniendo en cuenta que, en la mayoría de los casos las empresas prestadoras del servicio cuentan con la tecnología y los recursos para llevar a cabo la inspección (parcial o total) de las redes y el uso de estos datos como información de entrada a los modelos de predicción representa un aprovechamiento más eficiente de los recursos invertidos. No obstante, la aplicación de estas técnicas a datos reales también implica una serie de retos relacionados con la selección de los conjuntos de datos apropiados, el manejo de datos erróneos o incompletos en las bases de datos y la necesidad de integrar información proveniente de múltiples fuentes, los cuales pueden limitar significativamente la calidad y cantidad de conocimiento que se puede obtener (Bailey et al., 2015).

Así, el problema de predicción de fallas descrito anteriormente se puede considerar como una tarea de clasificación supervisada de minería de datos, en el cual el conjunto de datos observados (inspecciones) corresponden a los datos de entrenamiento o datos de aprendizaje del modelo con los cuales se desarrollan relaciones para predecir la condición en la cual se encuentran instancias no inspeccionadas a partir de un conjunto de variables dependientes (Wright et al., 2006). Diversas técnicas han sido utilizadas en las últimas décadas para dar solución a este problema de clasificación como regresiones lineales, regresión logística binaria y multinomial, arboles de decisión, bosques aleatorios, máquinas de soporte vectorial y redes neuronales. Así, múltiples autores han estudiado la viabilidad de aplicar estos métodos a datos reales obtenidos a partir de las empresas prestadoras de servicio obteniendo en general buenos resultados, pero encontrando así mismo, múltiples problemas claves que pueden limitar la aplicabilidad de estas técnicas y su aceptación por parte de los encargados de la gestión de activos en redes de alcantarillado.

En los siguientes dos capítulos de este documento (5.1 y 5.2) se presenta la descripción de estos modelos, al igual que una recopilación de las medidas de desempeño comúnmente aceptadas para medir el desempeño de estos. Por otro lado, en el capítulo 0 se describen los principales resultados obtenidos en la aplicación de estos modelos en diversos casos de estudio.

5.1 Modelos de Minería de datos

5.1.1 Regresión lineal

La técnica de regresión lineal es un modelo para problemas de regresión que modela la relación existente entre un conjunto de variables independientes o predictoras x y una variable dependiente o de respuesta y (Rencher & Schaalje, 2007). Sin embargo, es posible utilizar esta técnica en algunos problemas de clasificación al considerar la asignación de rangos de valores a cada una de las clases que toma la variable y, e interpretando el puntaje obtenido de y según la regresión para asignar una clase.

5.1.1.1 Modelo

De manera general, un modelo lineal realiza la predicción computando la suma de los pesos por las variables predictoras, más un término α denominado el intercepto (Géron, 2017). El modelo es:

$$\hat{y} = \alpha + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \cdots + \beta_n \mathrm{x}_\mathrm{n} \qquad \text{Ecuación 5-1}$$

donde:

- n es el número de variables predictoras.
- $\hat{y}$ es el valor predicho
- x_i es el valor de un predictor donde $i = 1,2,3,\ldots,n$
- β_j es el j-esimo parámetro del modelo, donde $\mathrm{j} = 1,2,3,\ldots,n$

De manera más concisa el modelo se puede escribir como:

$$\hat{y} = h(x) = \alpha + \beta^T \mathrm{X} \qquad \text{Ecuación 5-2}$$

donde:

- β es un vector columna que contiene los parámetros del modelo
- X es el vector columna de variables predictoras, conteniendo desde x_1 hasta x_n
- h es una función de hipótesis usando el modelo parametrizado por α y β

5.1.1.2 Estimando los parámetros α y β's del modelo

En la práctica, los valores de α y los β son desconocidos, por lo cual lo primero que se debe hacer para realizar predicciones sobre el modelo es utilizar los datos para encontrar estos parámetros (James, Witten, Hastie, & Tibshirani, 2013). Sean $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_m, y_m)$ los datos de entrenamiento, donde m es el número total de los datos de entrenamiento y teniendo en cuenta que cada $\vec{x}_i$ es un vector que contiene el valor correspondiente a cada variable predictor:

$$\vec{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}$$ Ecuación 5-3

El objetivo es obtener los parámetros α y β's de modo que el modelo lineal se ajuste a los datos de entrenamiento lo mejor posible, es decir: $\hat{y} \approx y$, donde y es el vector de resultados de los datos de entrenamiento (James et al., 2013). Para conocer que tan bien se ajustan los parámetros α y β's de una predicción en específico se utiliza la función de pérdida $\ell(y_i, \hat{y}_i)$, la cual se encarga de determinar qué tan buena es la predicción de un modelo respecto al valor real (Daumé, 2017b). En general la función de pérdida más común para regresión lineal, es el criterio de mínimos cuadrados (James et al., 2013):

$$\ell(y_i\, \hat{y}_i) = (y_i - \hat{y}_i)^2$$ Ecuación 5-4

Para ajustar todos los datos de entrenamiento al modelo lineal, se buscan los parámetros α y β's que minimizan:

$$\min_{\alpha,\beta} \sum_{i=1}^{m} \ell(y_i, \hat{y}_i) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} (y_i - (\alpha + \beta^T \mathrm{x_i}))^2$$ Ecuación 5-5

En otros palabras, se busca minimizar la función de costo:

$$J(\alpha, \beta) = \sum_{i=1}^{m} \ell(y_i, \hat{y}_i)$$ Ecuación 5-6

Si el valor obtenido de $\hat{y}_i \approx y_i$ para todos los $i = 1, 2, \ldots, m$ entonces $J(\alpha, \beta)$ tomará un valor pequeño y se concluye que el modelo se ajusta bien a los datos de entrenamiento. Por otro lado, si $\hat{y}_i$ es un valor muy lejano a y_i para más de un dato de entrenamiento entonces $J(\alpha, \beta)$ puede tomar un valor muy grande, indicando que el modelo no se ajusta adecuadamente a los datos (James et al., 2013).

5.1.2 Regresión logística

La técnica de regresión logística es un modelo para problemas de clasificación y es comúnmente utilizada para estimar la probabilidad de una instancia (x_i, y_i) de pertenecer a una clase en particular. Si la probabilidad estimada es mayor al 50% el modelo predice que la instancia pertenece a la clase positiva 1, de lo contrario a la clase negativa 0. Por lo cual, generalmente se conoce como un clasificador binario (Géron, 2017). El modelo $\hat{y}$ realiza predicciones de valores únicamente entre [0, 1].

5.1.2.1 Modelo

Si se retoma el modelo de regresión lineal:

$$\hat{y} = \alpha + \beta^T \mathrm{X} \qquad \text{Ecuación 5-7}$$

y se aplica en regresión logística, se podría decir que para valores donde $\hat{y}_i$ es cercano o menor a 0 se clasifique a la instancia en la clase 0 y para valores mayores a un umbral se clasifique a la instancia en la clase 1:

$$\hat{y} = \begin{cases} 0\ si & \alpha + \beta^T \mathrm{X} < \text{umbral} \\ 1\ si & \alpha + \beta^T \mathrm{X} \geq \text{umbral} \end{cases} \qquad \text{Ecuación 5-8}$$

donde el $\text{umbral} > 0$. Siempre que se tenga una solución lineal para ajustar una respuesta binaria codificada como 0 y 1, se deben definir los umbrales para cada clasificación. Para evitar este problema se puede crear un modelo $\hat{y}$ usando una función que obtenga resultados entre $[0,1]$ para todos los valores de X (James et al., 2013). Para definir cuál es la probabilidad de X de ser 1 :

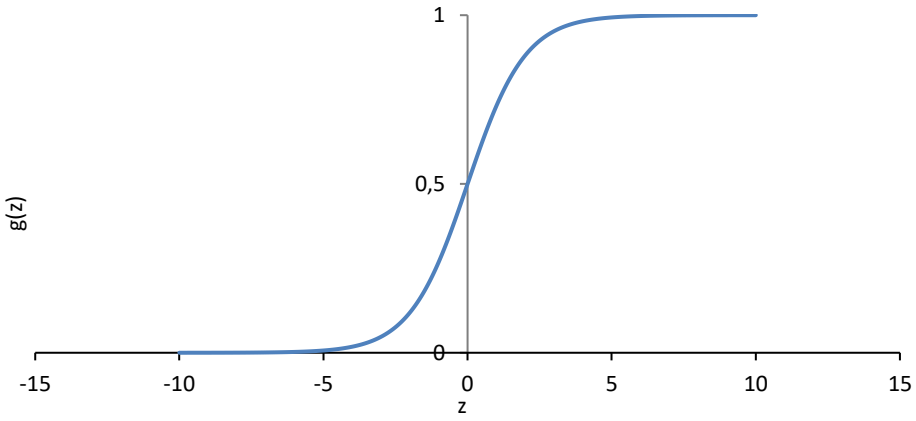
$$P(X) = \Pr(y = 1 \mid X;\ \alpha, \beta) = g(\alpha + \beta^T \mathrm{X}) \qquad \text{Ecuación 5-9}$$

Donde g es una función que mapea los valores de la regresión lineal entre $[0,1]$ y define la probabilidad de y de ser 1. De modo que, si $P(X) = 0.7$ se dice que X pertenece en un 70% a la clase 1 y en un 30% a la clase 0.

Existen muchas funciones que cumplen con esta característica; en regresión logística, se utiliza la función logística (James et al., 2013):

$$g(z) = \frac{e^z}{1 + e^z} \qquad \text{Ecuación 5-10}$$

El comportamiento de esta función se puede observar en la **Gráfica 5-1**. Se puede observar que entre mayor es el valor de z el valor de la función logística se aproxima a 1; por el contrario, entre más pequeño es z, su valor se aproxima a 0.



Gráfica 5-1. Función logística.

Utilizando la función logística en la probabilidad de obtener 1, dado X se obtiene finalmente:

$$P(x) = \frac{e^{\alpha + \beta^T \mathrm{X}}}{1 + e^{\alpha + \beta^T \mathrm{X}}} \qquad \text{Ecuación 5-11}$$

Finalmente, el modelo para la clasificación es:

$$\hat{y} = \begin{cases} 0 \; si \;\; P(x) < 0.5 \\ 1 \; si \; P(X) \geq 0.5 \end{cases} \qquad \text{Ecuación 5-12}$$

Si se realiza una manipulación de la ecuación de $P(x)$ se puede obtener:

$$\ln\left(\frac{P(x)}{1 - P(x)}\right) = \alpha + \beta^T \mathrm{X} \qquad \text{Ecuación 5-13}$$

La cantidad en que $P(x)$ cambia debido a una unidad de aumento en x depende del valor actual de x. Sin embargo, sin importar el valor de x si los $\beta's$ son positivos, entonces incrementar x estará asociado con incrementar $P(x)$ y si los $\beta's$ son negativos entonces incrementar x estará asociado con decrementar $P(x)$ (James et al., 2013).

5.1.2.2 Estimando los parámetros α y β's del modelo

Los valores de α y los β son desconocidos y deben ser estimados con base en los datos de entrenamiento disponibles. El objetivo del entrenamiento es encontrar los parámetros α y los β de modo que $P(x)$ sea alta para instancias de clase 1 y que $P(x)$ sea baja para instancias de clase negativa. La función de pérdida para un dato de entrenamiento se define (Géron, 2017):

$$\ell(y_i, P(x_i)) = \begin{cases} -\log P(x_i) & Si \; y_i = 1 \\ -\log(1 - P(x_i)) & Si \; y_i = 0 \end{cases} \qquad \text{Ecuación 5-14}$$

Esta función de pérdida toma sentido dado que $-\log z$ crece rápidamente cuando z se acerca a 0, así el costo será grande si el modelo estima una probabilidad cercana a 0 para instancias positivas y también lo será si el modelo estima una probabilidad cerca de 1 para instancias negativas. Por otro lado, $-\log z$ se aproxima a 0 cuando z se acerca a 1; así, el costo para instancias positivas con una probabilidad cercana a 1 será pequeña e igualmente para instancias negativas, lo cual es precisamente lo que se busca (Géron, 2017).

La función de costo considerando todos los datos de entrenamiento es el promedio entre la función de costo de todas las instancias de entrenamiento (Géron, 2017):

$$J(\alpha,\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i\ \log P(x_i) + (1-y_i)\ \log(1-P(x_i))] \qquad \text{Ecuación 5-15}$$

Entonces, lo que se debe hacer es minimizar la función de costo para obtener el α y los β que se ajustan mejor los datos de entrenamiento.

5.1.2.3 Clasificación múltiple

El modelo clásico de regresión logística es un modelo que permite realizar una clasificación binaria; sin embargo, existen muchos escenarios en los cuales se tienen más de dos clases por clasificar. Entonces, la manera convencional en la que es posible dar solución a este problema es tomar la solución aplicada al problema de dos clases e implementarla una serie de veces para resolver el problema de k clases. Así, se construyen un número k de clasificadores binarios, con los cuales se encontrará la clasificación más probable dado un conjunto de variables predictoras específico (Liu & Zheng, 2005).

El más común y ampliamente implementado es el método "Uno vs Todos", el cual construye K clasificadores $\hat{y}^{(i)}$ de regresión logística y cada clasificador toma a una clase $i = 1, 2, \dots, K$ como la clase positiva y las restantes como la clase negativa. Una vez se tienen los K clasificadores se dice que la instancia x pertence al clasificador que tenga la mayor $P^{(i)}(x)$ (Liu & Zheng, 2005), donde:

$$P^{(i)}(x) = \Pr(y = i \mid x;\ \alpha,\beta) \qquad \text{Ecuación 5-16}$$

En otras palabras, la clasificación de una instancia se obtiene seleccionando el clasificador que maximiza la probabilidad de que x pertenezca a la clase i:

$$\max_i P^{(i)}(x) \qquad \text{Ecuación 5-17}$$

5.1.3 Arboles de decisión

Los árboles de decisión se pueden aplicar en problemas de clasificación y regresión. Estos involucran la estratificación o segmentación del espacio predictor en un número de regiones simples. Su nombre proviene del hecho que el conjunto de reglas de particiones usados para segmentar un espacio predictor se puede resumir en un árbol. Los árboles de decisión son predictores transparentes, simples y útiles para la interpretación (James et al., 2013).

Un ejemplo de un árbol de decisión para clasificación se presenta a continuación, el cual predice una fruta con base en su color y su diámetro:

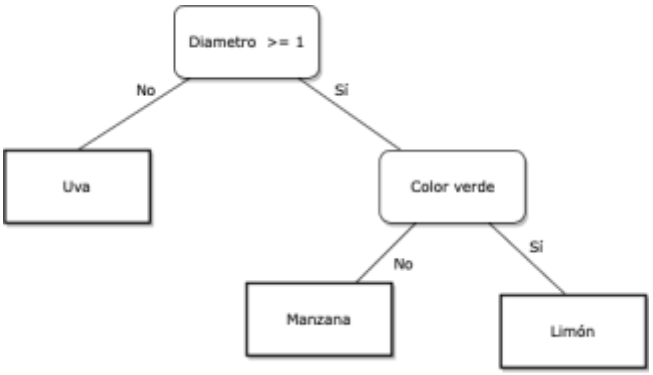


Figura 5-2. Ejemplo árbol de decisión

En un árbol las preguntas se encuentran en los nodos y la clasificaciones se escriben en hojas, los cuales son nodos que no tienen hijos. Cada nodo no terminal tiene dos hijos, el izquierdo indica qué hacer en caso de que la respuesta sea “no” y el derecho en caso de la respuesta sea “si”. En un árbol de decisión cada rama de un nodo representa una decisión y cada hoja representa una clasificación (Daumé, 2017a).

5.1.3.1 Modelo

El objetivo en los árboles de decisión es determinar qué preguntas hacer, el orden de las preguntas y qué predecir cuando ya se han hecho suficientes preguntas (Daumé, 2017a). Existen diversos algoritmos para construir los árboles de decisión, entre ellos: ID3 (Iterative Dichotomiser 3), C4.5, CART (Classification and regression trees), CHAID (Chi- squared automatic interaction detector) y MARS (Gupta et al., 2017).

- ID3

Es un algoritmo que construye el árbol de arriba a abajo utilizando los datos de entrenamiento, donde cada variable predictora en cada nodo es evaluada para seleccionar aquella que produce la mejor clasificación de los datos de entrenamiento; así, el atributo que obtiene la mayor ganancia de

información puede ser seleccionado en el nodo. Este proceso se hace de manera recursiva hasta obtener el árbol final. Para la construcción del árbol, ID3 solo acepta variables predictoras categóricas.

- C4.5

Es una extensión mejorada del algoritmo ID3 dado que permite trabajar con variables predictoras continuas y discretas. Para la división de datos categóricos realiza el mismo proceso de ID3 y para atributos continuos siempre genera divisiones binarias.

5.1.3.2 Medidas para la selección de atributos del árbol

El criterio de división que mejor separa los datos busca que al hacer una partición en D, cada partición resultante sea más pura, es decir que los datos que caen en cada partición pertenezcan a una misma clase (Gupta et al., 2017).

Si se supone que se tienen los siguientes datos D:

Tabla 5-1. Datos ejemplo – Medidas de selección de atributos

Color	Diámetro	Clasificación
Verde	3	Manzana
Amarillo	3	Manzana
Rojo	1	Uva
Rojo	1	Uva
Amarillo	3	Limón

El primer nodo del árbol recibe todos los datos D:

Color	Diámetro	Clasificación
Verde	3	Manzana
Amarillo	3	Manzana
Rojo	1	Uva
Rojo	1	Uva
Amarillo	3	Limón

Cada nodo va a realizar una pregunta sobre una de las características de los datos que permite hacer una división binaria de los datos. Como respuesta a esta pregunta los datos se dividen en dos subconjuntos D_1 y D_2; a continuación estos se vuelven la entrada los dos nodos hijos.

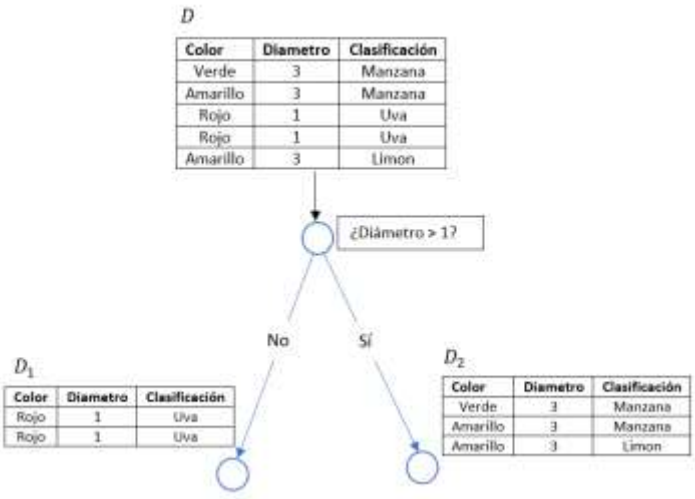


Figura 5-3. Ejemplo de la partición de datos en un árbol de decisión.

El objetivo de la pregunta es separar los datos de modo que a cada lado los subconjuntos sean lo más puros posibles, es decir pertenezcan a una misma clase. En el caso del ejemplo anterior, cuando la pregunta se responde con "No" los datos de D_1 unicamente pertenecen a la clase "Uva" por lo cual se pueden decir que estos datos son puros.

Las medidas de selección de atributos también conocidas como reglas de división y establecen cómo dividir los datos en cada nodo. Estas medidas se utilizan para determinar la impureza de unos datos y también, cuál es la mejor partición para hacer. Las más conocidas son: Entropía e Índice de Gini (para la impureza) y Ganancia de información y Radio de ganancia (para determinar cuál es el atributo por el cual se hace la mejor partición) (Gupta et al., 2017). En la Error! Reference source not found. se presenta una breve descripción de estas medidas.

Tabla 5-2. Medidas de selección de atributos en arboles de decisión

Medida	Descripción	Ecuación
Entropía	La entropía se utiliza como una medida de la homogeneidad de los datos D, se encarga de caracterizar su impureza. La entropía varía entre [0,1], entre más alta es la entropia significa que más impuros son los datos, se define así (Gupta et al., 2017)	$Entropia(D) = \sum_{i=1}^{c} - p_i \log_2 p_i$ Ecuación 5-18 Donde p_i es la probabilidad de un registro en D de pertenecer a la clase C y es estimada como: $p_i = \frac{\lvert C_{iD} \rvert}{\lvert D \rvert}$ Ecuación 5-19
Índice de Gini	El índice de Gini mide la impureza de D y es utilizado en el algoritmo CART.	$Gini(D) = 1 - \sum_{i=1}^{m} {p_i}^2$ Ecuación 5-20 Donde p_i es la probabilidad de un registro en D de pertenecer a la clase C y es estimada como: $p_i = \frac{\lvert C_{iD} \rvert}{\lvert D \rvert}$ Ecuación 5-21
Ganancia de información	La ganancia de información mide que tan buena es una partición; es la diferencia entre la entropía de los datos entrantes al nodo y la nueva entropía al realizar una partición. D se divide en v particiones donde D_j contiene aquellas tuplas resultantes de la división por el atributo A. El atributo con la mayor ganancia de información se utiliza en el nodo.	$GI(D,A) = Imp(D) - \sum_{j=1}^{v} \frac{\lvert D_j \rvert}{\lvert D \rvert} Imp(D_j)$ Ecuación 5-22 Donde • D es un conjunto de datos • A es el atributo por el cual se quiere hacer la partición • v es el número de divisiones que se hacen en D • D_j son los datos resutaltes de una división de D por el atributo A • $Imp(D)$ es una medida de impureza de los datos y se puede utilizar el índice de Gini o la entropía.
Radio de ganancia	La ganancia de información es una medida que prefiere seleccionar atributos que tienen un gran número de valores puesto que cada partición es pura y la ganancia de información por dicha partición es máxima. El radio de ganancia mejora la ganancia de información puesto que penaliza aquellas variables que toma muchos valores. El radio de ganancia aplica la ganancia de información pero introduce un nuevo concepto denominado: división de información, el cual determina que tan dispersos están los valores en esa variable. El atributo con mayor radio de ganancia se utiliza en el nodo.	$divisionInformacion_A(D) = -\sum_{j=1}^{v} \frac{\lvert D_j \rvert}{\lvert D \rvert} \log_2 \frac{\lvert D_j \rvert}{\lvert D \rvert}$ Ecuación 5-23 Donde • v es el número de posibles valores que puede tomar el atributo A • $\lvert D_j \rvert$ es el número de instancias en D que tienen el valor D_j en el atributo A • $\lvert D \rvert$ es el número de instancias Así el radio de ganancia se define como: $radioGanancia(D,A) = \frac{GI(D,A)}{divisionInformacion_A(D)}$ Ecuación 5-24

5.1.4 Bosques aleatorios

Los árboles de decisión afrontan problemas de sobreajuste (over-fitting) y alta varianza, dado que los datos se dividen hasta alcanzar pureza en los subconjuntos (James et al., 2013). Para evitar esto se crearon los bosques aleatorios, los cuales son una colección de árboles de decisión donde cada árbol produce una respuesta a una entrada X. Al final del proceso, la clasificación de la entrada X se obtiene por medio de un sistema de votación como aquella respuesta que haya sido obtenida un mayor número de veces (Gupta et al., 2017).

Para la construcción de cada árbol en el bosque aleatorio, se utiliza un método de muestreo (sampling) de los datos de entrenamiento, por ejemplo el método bootstrap para generar B nuevos conjuntos de datos a partir de los datos originales de entrenamiento entregados al modelo y con cada conjunto de datos b_i se crea un árbol (Hastie, Tibshirani, & Friedman, 2017).

Método de Bootstrap

Suponer que se tienen un conjunto de datos $Z = \{z_1, z_2, \dots, z_m\}$ donde $z_i = \{x_i, y_i\}$. La idea es crear nuevos conjuntos de datos con remplazo a partir de Z. Los nuevos conjuntos de datos deben tener el mismo tamaño de Z, se crean B nuevos conjuntos de datos como se desee. Produciendo B connjuntos de datos de bootstrap (Hastie et al., 2017).

5.1.5 Redes neuronales artificiales

Una red neuronal artificial (ANN por sus siglas en inglés) es un modelo matemático que trata de simular la estructura y funcionalidad de las redes neuronales biológicas. El componente básico de una ANN es una neurona la cual es simplemente un modelo matemático (una función) (Krenker, Bester, & Kos, 2011). Cada neurona es una unidad de procesamiento que ejecuta la $P(x)$ del modelo de regresión logistica. Las entradas se multiplican por unos pesos w (equivalente a los parámetros β mencionados en regresión logistica). Dentro de la neurona se realiza la suma de los pesos por las entradas más el sesgo b (equivalente al intercepto en regresión logística); finalmente la suma de los pesos por las entradas más el sesgo se pasan a través de la función de activación o función de transferencia que clasifica el resultado en términos categóricos (función logística en el caso de regresión logística).

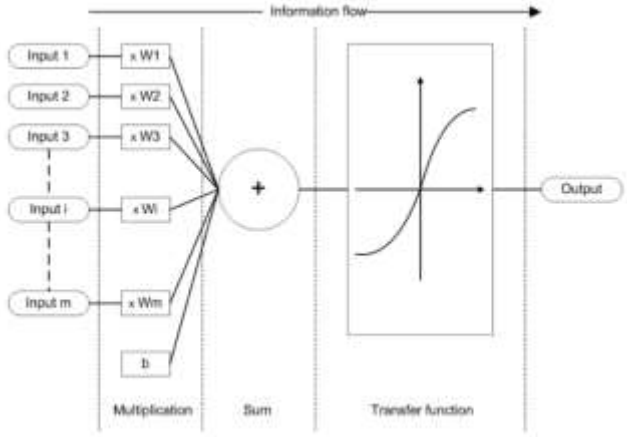


Figura 5-4. Funcionamiento de una neurona artificial. Tomado de Krenker et al., (2011)

Aunque los principios de funcionamiento y el sencillo conjunto de reglas de la neurona artificial no tienen nada de especial, el potencial y el poder de cálculo de estos modelos cobran vida cuando comenzamos a interconectarlos en redes neuronales artificiales. Al combinar dos o más neuronas artificiales se obtiene una ANN (Krenker et al., 2011). Sin embargo, estos modelos aún se consideran cajas negras debido a que no es posible tener un entendimiento completo de la estructura del modelo o de cómo se obtiene la predicción $\hat{y}$.

5.1.6 Máquinas de soporte vectorial

Las máquinas de soporte vectorial (Support Vector Machines – SVM) se utilizan en problemas de clasificación y regresión. Este modelo busca el hiperplano que mejor separa las clases de los datos de entrenamiento. La idea es encontrar un hiperplano que maximice la distancia a los puntos más cercanos de cada clase, de manera que la clasificación sea más confiable (Grus, 2015).

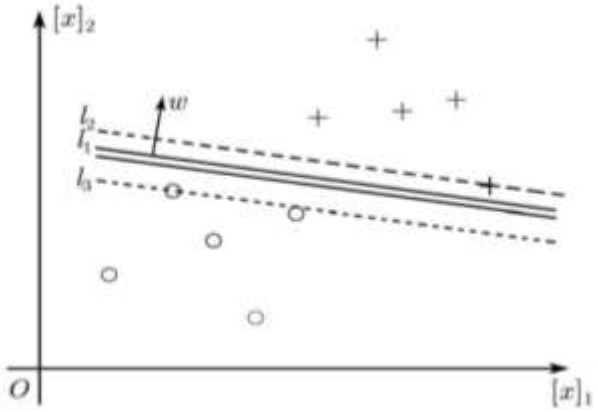


Figura 5-5. Clasificación mediante máquinas de soporte vectoriales (SVM). Tomado de (Deng, Tian, & Zhang, 2013)

Como se puede ver en la imagen anterior $l_1$es un hiperplano que separa perfectamente las clases positivas de las clases negativas; sin embargo ese hiperplano no es único pues cualquier hiperplano paralelo a l_1 es un candidato para el ejemplo. Los hiperplanos en líneas punteadas l_2 y l_3 se llaman lineas de soporte dado que cada uno pasa por una o más instancias de las diferentes clases. De todos los planos posibles entre l_2 y l_3 aquel que se encuentra en la mitad será la mejor opción (Deng et al., 2013).

5.1.6.1 Modelo

El hiperplano que se construye es:

$$b + \omega^T \mathrm{x} = 0 \qquad \text{Ecuación 5-25}$$

El objetivo de las máquinas vectoriales de soporte es definir el hiperplano y si se obtienen valores mayores a 0 se obtiene una clasificación y los menores a cero obtendrán la otra clasificación.

$$\hat{y} = \begin{cases} 1 \ si \ b + \omega^T \mathrm{x} \geq 0 \\ 0 \ otro \ caso \end{cases} \qquad \text{Ecuación 5-26}$$

Dado que el hiperplano seleccionado se encuentra exactamente en la mitad entre las dos líneas de soporte, las líneas de soporte se definen como:

$$b + \omega^T \mathrm{x} = 1$$
$$b + \omega^T \mathrm{x} = -1 \qquad \text{Ecuación 5-27}$$

5.1.6.2 Encontrando los parámetros b y w's del modelo

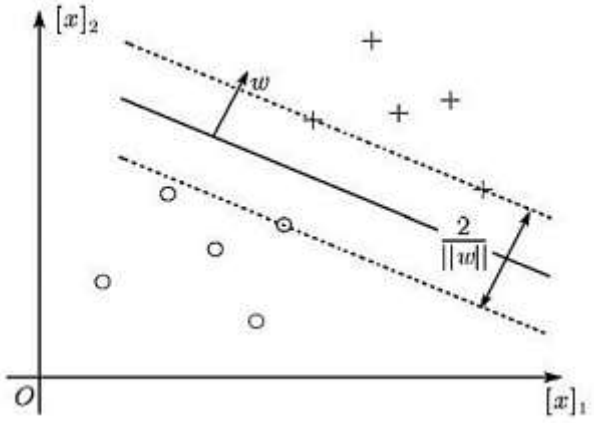


Figura 5-6. Vectores de soporte – SVM. Tomado de (Deng et al., 2013)

Geométricamente la distancia entre las líneas de soporte es $\frac{2}{||\omega||}$. Así el objetivo es maximizar la margen entre las dos clases, lo cual conlleva al siguiente problema de optimización para encontrar el hiperplano:

$$\max_{\omega,b} \frac{2}{||\omega||}$$ Ecuación 5-28

Lo anterior se utiliza cuando los datos son linealmente separables; sin embargo, este caso es poco común y las fronteras de los datos usualmente tienen comportamientos menos simplificados, por lo cual se ha establecido la técnica de los kernels. Esta técnica consiste en utilizar al interior de los modelos SVM una transformación de las variables que permita mapear el nuevo problema de clasificación mediante un modelo SVM lineal. Estas transformaciones o tipo de kernels pueden ser lineales, polinomiales, una función de base radial (RBF) o sigmoidal, entre otros.

5.1.7 Regresión polinómica evolutiva (EPR)

Regresión polinómica evolutiva (Evolutionary Polynomial Regression – EPR) es un modelo utilizado para problemas de regresión que integra los métodos de regresión simbólica y regresión numérica para desarrollar expresiones matemáticas que tienen estructura polinomial. La expresión general que representa EPR es (Giustolisi, Savic, & Laucelli, 2004):

$$y = \sum_{j=1}^{m} f(\mathbf{X}, a_j) + a_0$$ Ecuación 5-29

donde:

- y es el resultado estimado del sistema
- a_j es el valor de una constante
- f es una función construida mediante el proceso de EPR
- $\mathbf{X}$ es la matriz de las variables de entrada
- m es el número de términos de la expresión polinomial

Entonces, EPR es una técnica de dos pasos para construir modelos simbólicos, en que primero se realiza la identificación de la estructura del modelo $\boldsymbol{f}$ y en segundo lugar, se estiman los parámetros del modelo $\boldsymbol{a_j}$. Las estructuras simbólicas en el primer paso se buscan a partir de Algoritmos Genéticos, mientras que las constantes del modelo se encuentran solucionando un problema lineal de mínimos cuadrados.

Esta técnica tiene un alto nivel de flexibilidad, lo cual implica que las expresiones encontradas mediante la aplicación de EPR pueden ajustarse en gran medida a los datos de entrenamiento y pueden generarse problemas de sobreajuste (over-fitting), evitando que la capacidad del modelo

para la predicción en otros conjuntos de datos no sea óptima. Sin embargo, se han investigado técnicas para evitar este sobreajuste entre las cuales se encuentran la penalización de la complejidad del modelo, varianza de las constantes a_j y varianza de los términos $a_j Z_j$ (Giustolisi & Savic, 2006). Los detalles matemáticos del proceso que se debe llevar a cabo para encontrar la expresión simbólica, al igual que la descripción de los métodos para evitar sobreajuste se describen en detalle en Giustolisi & Savic (2006).

5.2 Medidas de desempeño de los modelos

Para evaluar el desempeño de los modelos descritos anteriormente es posible establecer diferentes medidas de desempeño, las cuales pueden ser indicativas de la capacidad predictiva de los modelos a nivel de la red o a nivel de las tuberías (Caradot, Riechel, et al., 2018). A nivel de la red, las medidas de desempeño son indicadores de la capacidad del modelo para predecir la distribución de las tuberías que se encuentran en las diferentes clases que caracterizan su condición; mientras que a nivel de tuberías, las medidas de desempeño buscan evaluar la capacidad del modelo para predecir correctamente la condición inspeccionada de cada tubería.

Caradot et al., (2018) menciona la importancia del uso de ambas medidas de desempeño puesto que la información que ambas proporcionan sirve para propósitos diferentes. Las medidas a nivel de red muestran la importancia del modelo para apoyar la planeación estratégica de la rehabilitación a largo plazo, puesto que permiten observar el estado general de la red para cada condición y las actividades e inversiones requeridas para lograr una distribución de condiciones que garantice un buen desempeño; mientras que, las medidas a nivel de tuberías muestran la importancia del modelo para apoyar las estrategias de inspección en la red al identificar las tuberías más críticas que deben ser priorizadas.

De igual manera, es importante mencionar que al estudiar el problema de predicción de la condición en redes de alcantarillado, es posible que se requieran algunas medidas de desempeño no convencionales para garantizar el buen desempeño de los modelos incluso cuando la distribución de tuberías en las diferentes clases que categorizan su condición estructural (Carvalho, 2015).

Para desarrollar estas medidas de desempeño de un modelo en un conjunto de datos con clases etiquetadas conocidas se utiliza la matriz de confusión:

5.2.1 Matriz de confusión

Esta matriz corresponde a la comparación de los resultados obtenidos mediante el modelo (Predicted condition class) y las observaciones resultantes de la inspección de las tuberías (Actual condition class). Luego, es una matriz de $m \times m$ cuyos elementos c_{jk} corresponden al número de instancias que tienen una condición observada $o_i = j$ y son pronosticadas por el modelo en una condición $p_i = k$. Así, los elementos de la diagonal de esta matriz representan las instancias para

las cuales la condición real coincide con la predicción del modelo; mientras que los elementos por fuera de la diagonal representan las instancias para la cuales el modelo pronostica incorrectamente la condición, siendo así errores de clasificación (Mashford et al., 2011).

Luego, esta matriz puede construirse cuando la clasificación es binaria, es decir que solo existen dos posibles clases para la condición de las tuberías como se muestra en la Tabla 5-3 (Harvey & McBean, 2014), o puede realizarse para las diferentes clases en las que se categorice la condición de las tuberías.

Tabla 5-3. Matriz de confusión para clasificación binaria.

		Predicted condition class	
		Poor	Good
Actual condition class	Poor	True Positive (TP)	False Negative (FN)
	Good	False Positive (FP)	True Negative (TN)

Se pueden obtener cuatro resultados posibles al comparar la clasificación real de las tuberías y la predicción realizada por el modelo (Harvey et al., 2015):

- True positive – Positivo verdadero (TP): Una tubería que se sabe **está en mal estado** y se pronostica correctamente que se encuentra en **mal estado por el modelo**.

- False positive – Falso positivo (FP): Una tubería que se sabe **está en buen estado** y se pronostica incorrectamente que se encuentra en **mal estado por el modelo**

- True negative – Negativo verdadero (TN): Una tubería que se sabe **está en buen estado** y se pronostica correctamente en **buen estado por el modelo**.

- False negative – Negativo falso (FN): Una tubería que se sabe **está en mal estado** y se pronostica incorrectamente en **buen estado por el modelo**.

Teniendo en cuenta los resultados posibles, se pueden establecer una serie de medidas alternativas, que permiten cuantificar el desempeño del modelo cuando la distribución del conjunto de datos en las clases no es uniforme; es decir, existe una clase que tiene un número significativamente mayor que otra clase, como es el caso esperado en las tuberías de redes de alcantarillado. En estas, se espera que haya una mayor cantidad de tuberías en buen estado y un número mucho menor de tuberías en mal estado. Estas son (Carvalho, 2015; Harvey et al., 2015):

$$Accuracy - Exactitud = \frac{TP + TN}{TP + TN + FN + TN} = P(\hat{Y} = Y) \qquad \text{Ecuación 5-30}$$

$$True\ positive\ rate = \ Sensitivity - Sensibilidad = \frac{TP}{TP + FN} = P(\hat{Y} = 1|Y = 1) \quad \text{Ecuación 5-31}$$

$$True\ negative\ rate = Specificity - Especificidad = \frac{TN}{FP + TN} = P\left(\hat{Y} = 0\middle|Y = 0\right) \quad \text{Ecuación 5-32}$$

$$False\ positive\ rate = 1 - TNR = \frac{FP}{FP + TN} \quad \text{Ecuación 5-33}$$

$$Precision - Precisión = \frac{TP}{TP + FP} = P\left(Y = 1\middle|\hat{Y} = 0\right) \quad \text{Ecuación 5-34}$$

En las ecuaciones anteriores, Y corresponde a la clase real según la condición de la tubería y $\hat{Y}$ corresponde a la clase pronosticada por el modelo para la condición de la tubería; por otro lado, 1 corresponde a la clase mal estado (Poor) y 0 corresponde a la clase buen estado (Good).

Luego, al utilizar todas las medidas anteriores y no solo la exactitud del modelo, es posible entender en más detalle la forma en la que el modelo realiza predicciones y si se presentan resultados sesgados debido al entrenamiento del modelo con un conjunto de datos con desequilibrio de clases.

La exactitud es una medida que asume que los falsos positivos y los falsos negativos tienen el mismo costo o las mismas consecuencias, a pesar de que se sabe que los falsos negativos tienen mayor costo debido a los problemas que se pueden generar al pronosticar una tubería en buen estado, cuando en realidad tiene una alta probabilidad de falla. Así mismo, al usar solo esta medida para evaluar el desempeño del modelo, es posible encontrar que un modelo trivial es apropiado cuando la clase que contiene las tuberías en buen estado tiene un mayor número de instancias pues predice la mayoría de las tuberías en esta clase.

La sensibilidad, por otro lado, es una medida de la confianza que se puede tener en el modelo para predecir instancias en una clase de mal estado considerando el total de instancias que se encuentran realmente en esa clase; y la especificidad consiste en la medida contraria, es decir, mide la confianza que se puede tener en el modelo para predecir instancias en una clase de buen estado considerando el total de instancias que se encuentran realmente en esa clase.

5.2.2 Medidas a nivel de red y a nivel de tuberías

Ahora bien, en el trabajo realizado por Caradot et al. (2018) se desarrollan, en conjunto con la Empresa prestadora del servicio de agua de Berlín (Berlin Water Company), una serie de medidas

más generales que buscan evaluar el desempeño de los modelos teniendo en cuenta los conceptos anteriores y la necesidad de distinguir entre las medidas a nivel de red y a nivel de tuberías:

Medidas a nivel de red:

Estas medidas describen la desviación entre la distribución de tuberías en las clases de acuerdo a su condición obtenida mediante los pronósticos del modelo y la distribución real de estas clases de acuerdo a la inspección de las redes. De acuerdo con (Caradot, Riechel, et al., 2018), se pueden considerar dos tipos de medidas:

- Desviación de la distribución de la condición – *Todas* las tuberías: En este caso, se debe definir un parámetro K_i que cuantifique la desviación absoluta entre los porcentajes de tuberías pronosticadas e inspeccionadas en cada clase i en que se encuentre categorizada la condición de todas las tuberías en el conjunto de datos analizado. Así:

$$K_i = |\%Tuberías\ inspeccionadas\ C_i - \%Tuberías\ pronosticadas\ C_i| \quad \text{Ecuación 5-35}$$

- Desviación de la distribución de la condición – Grupo de tuberías en un rango de edad: Igual que en la medida anterior, se define un parámetro K_i que representa lo mismo que en el caso anterior, pero únicamente para un grupo de tuberías cuya edad esté en un rango específico. Esto se realiza debido a que las tuberías que se encuentran en un rango de edad mayor pueden generar sesgo en los resultados del modelo debido a que es probable que estas tuberías ya hayan sido rehabilitadas a lo largo de su vida útil, introduciendo sesgo de selección de supervivencia.

Medidas a nivel de tuberías:

Estas medidas son necesarias un modelo puede garantizar resultados excelentes al nivel de red pero aun así fallar en la predicción de la condición apropiada para cada tubería de la red; obteniendo así, los porcentajes apropiados de tuberías en cada clase según la condición pero las tuberías equivocadas en cada condición.

En general, estas medidas son muy similares a las derivadas en las ecuaciones Ecuación 5-31Ecuación 5-34, pero pueden definirse de manera que sean aplicables para un número de clases superior a 2. Así, para la clase C_i (Caradot, Riechel, et al., 2018):

$$True\ positive\ rate_{C_i} =$$

$$Sensitivity_{C_i} = \frac{número\ de\ predicciones\ correctas\ en\ la\ clase\ C_i}{número\ de\ observaciones\ en\ la\ clase\ C_i} \quad \text{Ecuación 5-36}$$

La sensibilidad indica el porcentaje de tuberías inspeccionadas en la condición i que han sido pronosticadas correctamente en la condición i.

$$False\ negative\ rate_{C_i} =$$

$$Miss\ rate_{C_i=} \frac{\begin{matrix} número\ de\ tuberías\ observadas\ en\ condición\ i \\ pero\ pronosticadas\ en\ condición\ j \end{matrix}}{número\ de\ observaciones\ en\ la\ condición\ j}\ para\ j < i \qquad \text{Ecuación 5-37}$$

La tasa de falla (miss rate) indica el porcentaje de tuberías inspeccionadas en la condición i que han sido pronosticadas en una mejor condición j. Esto quiere decir, que es un indicador de cuanto se sobreestima la condición inspeccionada de las tuberías.

$$False\ positive\ rate_{C_i} = False\ alarm\ probab_{C_i}$$

$$= \frac{\begin{matrix} número\ de\ tuberías\ observadas\ en\ condición\ i \\ pero\ pronosticadas\ en\ condición\ j \end{matrix}}{número\ de\ observaciones\ en\ la\ condición\ j};\ para\ j > i \qquad \text{Ecuación 5-38}$$

La tasa de falsos positivos o probabilidad de falsa alarma indica el porcentaje de tuberías inspeccionadas en la condición i que han sido pronosticadas incorrectamente en una peor condición j. Por lo tanto, esta medida es un indicador que cuanto se subestima la condición inspeccionada de las tuberías.

5.2.3 Curva de características operativas del receptor (ROC)

La curva de características operativas del receptor (receiver-operating characteristics – ROC) es una herramienta comúnmente utilizada para evaluar la relación que existe entre la tasa de positivos verdaderos (TPR) y la tasa de positivos falsos (FPR), con el propósito de identificar la compensación que existe entre verdaderos positivos y falsos positivos pronosticados por el modelo.

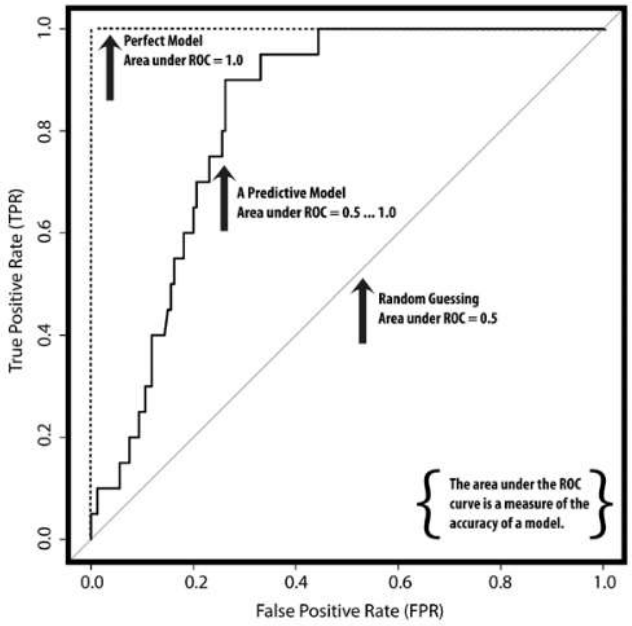


Figura 5-7. Curva de características operativas del receptor (ROC). Tomado de Harvey et al. (2015)

La medida de desempeño utilizada teniendo en cuenta la curva ROC corresponde al área bajo la curva de la misma. Como se observa en la Figura 5-7, un modelo perfecto se ve reflejado en un área bajo la curva ROC igual a 1, de manera que la tasa de positivos verdaderos sea máxima y la tasa de falsos positivos sea mínima. Lo primero implica que el modelo clasificador tiene un buen comportamiento al categorizar las tuberías que se encuentran en buena condición en la clase apropiada; mientras que lo segundo significa que la tasa de falsos positivos es baja (o que la tasa de negativos verdaderos es alta), logrando también una clasificación de las tuberías que se encuentran en mala condición en la clase apropiada. Así, se previene que la exactitud evaluada del modelo únicamente tome un valor alto debido a que una clase tiene significativamente más instancias que otra. Para entender esto más fácilmente, se puede pensar en un conjunto de datos hipotético en el cual 750 tuberías se encuentran en buen estado y 250 en mal estado, un modelo sesgado podría alcanzar una exactitud del 75% al asignar todas las tuberías a la primera clase y ser considerado como un modelo aceptable, pero el modelo no sería apropiado para predecir las tuberías en mala condición, lo cual puede resultar significativamente más costoso (Harvey et al., 2015).

El modelo con el área más grande bajo la curva ROC puede ser considerado el más efectivo para realizar la clasificación. Valores del área bajo la curva ROC mayores a 0.7 en un conjunto de datos estratificado se consideran aceptables y el umbral inferior de 0.5 es considerado cuando las predicciones de las clases tienden a ser inadecuadas en datos con clases desequilibradas (Harvey & McBean, 2014).

5.3 Casos de estudio

5.3.1 Modelos aplicados a la detección de fallas en tuberías de Redes de Alcantarillado

5.3.1.1 Regresión logística

Los modelos de regresión logística son utilizados frecuentemente en problemas de clasificación y en particular, se ha considerado su aplicación para el caso de la predicción de fallas en sistemas de alcantarillado por diversos autores como se observa en la Tabla 5-4. En el contexto de redes de alcantarillado, el resultado obtenido por los modelos de regresión logística corresponde a la probabilidad de falla de una tubería $P(x)$ con la cual se realiza la clasificación de pertenencia a una clase mediante el modelo descrito en la sección 5.1.2.1 de este documento. Ahora bien, hay tres tipos de modelos de regresión logística que pueden ser utilizados, dependiendo del número de clases en que se puede categorizar la condición de las tuberías: (1) Regresión logística binaria, (2) Regresión logística multinomial y, (3) Regresión ordinal (Salman & Salem, 2012).

Al observar las investigaciones realizadas por diferentes autores, es posible notar que la aplicación de esta técnica de minería de datos a la predicción de fallas en redes de alcantarillado ha sido ampliamente estudiada y aún continua siendo relevante para realizar comparaciones de la capacidad predictiva de otros modelos. Entre las primeras investigaciones, Ariaratnam et al. (2001) desarrollaron un modelo de regresión logística considerando los efectos de las variables edad, diámetro, material, tipo de tubería y profundidad media de cobertura; encontrando que es posible reducir la subjetividad al encontrar valores de probabilidad de falla de las tuberías en lugar de utilizar calificaciones numéricas.

Por otro lado, Salman & Salem (2012) analizan los resultados obtenidos al aplicar los tres tipos de regresión logística a un mismo conjunto de datos, encontrando que el modelo multinomial permite obtener mayores valores de sensibilidad (sensitivity) pero con el modelo binario se obtienen mayores valores de la especificidad (specificity); por lo cual, este último modelo debería preferirse al trabajar con conjuntos de datos desbalanceados.

Tabla 5-4. Modelos de regresión logística para la predicción de fallas en redes de alcantarillado.

No.	Año	Título	Autores	Resultado de la metodología	Nivel de predicción
1	2001	Assessment of infrastructure inspection needs using logistic models	Ariaratnam, Samuel El-Assaly, Ashraf Yang, Yuqing	Probabilidad de falla individual por tubería	Tubería
2	2006	Prioritizing Sanitary Sewers for Rehabilitation Using Least-Cost Classifiers	Leonard T. Wright; James P. Heaney; and Shawn Dent	Probabilidad de falla individual por tubería	Tubería

No.	Año	Título	Autores	Resultado de la metodología	Nivel de predicción
3	2012	Modeling failure of wastewater collection lines using various section-level regression models	Baris Salman and Ossama Salem	Probabilidad de falla individual por tubería	Tubería
4	2015	Benefits of using basic, imprecise or uncertain data for elaborating sewer inspection programmes	Mehdi Ahmadi, Frederic Cherquic,d1, Jean-Christophe De Massiaca and Pascal Le Gauffre	Probabilidad de falla individual por tubería	Tubería
5	2018	Sewer Condition Prediction and Analysis of Explanatory Factors	Tuija Laakso, Teemu Kokkonen, Ilkka Mellin and Riku Vahala	Probabilidad de falla individual por tubería	Tubería

En general, estos modelos son de gran utilidad debido a su fácil interpretación y a la capacidad de incluir variables predictoras continuas y categóricas. Sin embargo, debido a que estos modelos se encuentran basados en relaciones lineales entre las variables, pueden enfrentar dificultades al explicar relaciones en problemas altamente no lineales, como es el caso de la predicción de fallas en redes de alcantarillado.

5.3.1.2 Arboles de decisión

El uso de los árboles de decisión o de la técnica mejorada bosques aleatorios, se ha estudiado con mayor frecuencia en el caso del problema de clasificación de tuberías según su en redes de alcantarillado, obteniendo en general resultados buenos debido a su capacidad para modelar relaciones no lineales (Laakso, Kokkonen, et al., 2018).

Múltiples investigaciones se han llevado a cabo utilizando registros de inspección de las redes de alcantarillado de diversas empresas, como se observa en la Tabla 5-5. En general, el resultado al implementar esta técnica corresponde a una estructura de árbol mediante la cual es posible clasificar las tuberías de acuerdo a sus atributos y las preguntas realizadas en los nodos del árbol. Así, se construye un árbol con las variables predictoras del modelo en el cual las variables de los nodos superiores corresponden a los factores de mayor importancia para el proceso de clasificación.

En el trabajo realizado por Jung et al. (2012) utilizan un modelo de árbol de decisión para identificar las variables predictoras más relevantes en su caso de estudio, encontrando que el diámetro y el material de las tuberías son los factores que se ubican en los nodos principales. En línea con esto, el trabajo de Laakso, Kokkonen, et al. (2018) también utiliza esta técnica para encontrar las variables más relevantes y predecir la condición de las tuberías obteniendo una clasificación correcta en el 62% de los casos al fijar la $\mathrm{FNR} = 20\%$.

Por otro lado, Harvey & McBean (2014) comparan esta técnica con la predicción obtenida mediante SVM en un caso de estudio de Canadá, en el que encuentran un modelo dependiente de la edad, la profundidad, la longitud, el diámetro y el número de fallas cercanas a las tuberías, que tiene una exactitud global de 76% y un área bajo la curva ROC igual a 0.78 y supera el desempeño del modelo SVM. De manera similar, los autores anteriores aplican esta técnica a otro conjunto de datos de una red en Gelph y encuentran relevantes las variables año de instalación, longitud, pendiente y diámetro de las tuberías, logrando un área bajo la curva ROC igual a 0.77.

Tabla 5-5. Modelos de árboles de decisión para la predicción de fallas en redes de alcantarillado.

No.	Año	Título	Autores	Caso de estudio	Resultado de la metodología	Nivel de predicción
1	2012	Application of Classification Models and Spatial Clustering Analysis to a Sewage Collection System of a Mid-Sized City	I-S. Jung, J. H. Garrett Jr., L. Soibelman and K. Lipkin	RedZone Robotics network	Identificación de los factores predictores del deterioro de tuberías	Tubería
2	2014	Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure	Robert Richard Harvey and Edward Arthur McBean	City of Guelph, Ontario, Canada	Probabilidad de falla de una tubería dados sus atributos	Tubería
3	2015	A Data Mining Tool for Planning Sanitary Sewer Condition Inspection	R. Harvey and E. McBean	City of Guelph, Ontario, Canada	Probabilidad de falla de una tubería dados sus atributos	Tubería
4	2015	Predictive risk modelling of real-world wastewater network incidents	James Bailey, Edward Keedwell, Slobodan Djordjevic, Zoran Kapelan, Chris Burton and Emma Harris	Network of Dŵr Cymru Welsh Water (DCWW)	Probabilidad de falla de una tubería dados sus atributos	Tubería
5	2018	Sewer Condition Prediction and Analysis of Explanatory Factors	Tuija Laakso, Teemu Kokkonen, Ilkka Mellin and Riku Vahala	-	Probabilidad de falla de una tubería dados sus atributos	Tubería

Así, se observa que la aplicación de este tipo de modelos resulta apropiado para realizar predicciones de fallas en sistemas de alcantarillado y puede ser de gran utilidad para otras tareas de rehabilitación debido a la interpretabilidad de la estructura del árbol obtenida. Los bosques aleatorios constituyen una técnica más avanzada que aún no han sido aplicada extensamente en este problema de clasificación particular pero podría brindar resultados más confiables para otros conjuntos de datos debido a la disminución de sobreajuste que se logra mediante esta técnica.

5.3.1.3 Máquinas de soporte vectorial

Ya que esta técnica de minería de datos permite realizar tareas de clasificación y regresión, su estudio es de interés para el problema de clasificación de tuberías según su condición en la gestión de redes de alcantarillado. Pocos autores como Mashford et al. (2011) y Harvey & McBean (2014) han implementado modelos de SVM aplicados al caso de redes de alcantarillado; sin embargo, la aplicación de estas técnicas ha sido estudiada en otras áreas como bioinformática, categorización de texto, segmentación de imágenes y análisis financieros, entre otros (Mashford et al., 2011).

Mashford et al. (2011) investigaron la aplicación de estos modelos al caso de redes de alcantarillado como alternativa a la aplicación de modelos ANN en la red de drenaje de Adelaida, Sur de Australia. Los autores consideran la aplicación de este modelo debido a ventajas como: Entre las consideraciones para la aplicación de este modelo, los autores mencionan el hecho de que no se requiere la especificación de la estructura interna, la habilidad de ser entrenados con conjuntos de datos más pequeños y su habilidad de adaptarse a predicciones no lineales en problemas complejos, como puede ser el caso de la predicción de la condición de tuberías de alcantarillado. En su investigación consideran la aplicación de cuatro modelos diferentes de SVM modificando la cantidad de variables predictoras incluidas en el modelo, siendo entrenados con un conjunto de datos de 1441 casos; encontrando que SVM's permiten modelar adecuadamente el proceso de deterioro y que el modelo más adecuado corresponde al que involucra el menor número de variables, lo cual se explica en detalle en la sección 8.1.2 de este documento.

Por otro lado, Harvey & McBean, (2014) comparan la aplicación de un árbol de decisión con SVM considerando resultados de inspección desde 2008 hasta 2011 en la red de alcantarillado de Gelph, Ontario, Canadá. Los autores encuentran, en este caso, que la capacidad de predicción es mejor al utilizar el árbol de decisión por encima del modelo SVM, a pesar de encontrar un área bajo la curva ROC con SVM es igual a 0.69, con una exactitud (accuracy) de 89%. Además, los autores consideran otro escenario para la implementación de SVM en que se calibre este modelo considerando el problema de desbalance de clases, mejorando el desempeño del mismo en conjuntos de datos desbalanceados ($\text{Área ROC} = 0.72$) pero disminuyendo la exactitud general del modelo a un 58%.

El uso de esta técnica de minería de datos puede representar una buena alternativa al uso de modelos más complejos como redes neuronales en los que es posible explicar relaciones altamente complejas pero se debe sacrificar la interpretabilidad del modelo y además se requieren conjuntos

de datos muy grandes para un entrenamiento satisfactorio del modelo. Sin embargo, es importante tener en cuenta que cuando las SVM tratan con problemas de clasificación no lineales es necesario incluir el uso de kernels que pueden dificultar el entendimiento de los modelos. Por otro lado, debido a los pocos trabajos encontrados en los cuales se aplica esta técnica al problema de fallas de redes de alcantarillado, su uso en otros casos de estudio sería de gran utilidad para identificar en más profundidad sus ventajas y limitaciones.

5.3.1.4 Regresión polinómica evolutiva

Esta técnica de minería de datos se introdujo por Savic et al. (2006) y consiste en un modelo que permite desarrollar expresiones matemáticas para la predicción de colapsos y bloqueos en sistemas de alcantarillado a partir de las bases de datos resultantes de inspecciones de las tuberías. Así, permite identificar relaciones interpretables entre los atributos de las tuberías y la tasa de falla por unidad de longitud de las mismas, facilitando no solo la detección de fallas en un sistema de drenaje particular estudiado sino también la identificación de la relación entre las variables independientes (predictoras) y la variable independiente (tasa de falla) (Savic et al., 2006).

En particular, Savic et al. (2006) encuentran expresiones para la tasa de bloqueos y la tasa de colapsos por unidad de longitud en un sistema de alcantarillado del Reino Unido. A partir de sus resultados, obtienen que la tasa de colapsos se incrementa a medida que el radio entre la profundidad mínima y el diámetro equivalente (C_e/D_e) incrementa, y a medida que la longitud de las tuberías L_m es menor; lo cual puede representar una confirmación de que las tuberías de mayor tamaño (D_e) se instalan más cuidadosamente y que la tasa de colapsos es mayor en redes más fragmentadas (un mayor número de tuberías de mayor longitud) respectivamente. Vale la pena mencionar que el modelo seleccionado no corresponde a la expresión que maximiza el coeficiente de determinación al evaluar el desempeño del modelo (71.75%) pero se prefiere sobre las otras expresiones debido a que permite una fácil interpretación de las relaciones entre variables. En este caso particular, los autores mencionan la importancia de incluir la variable edad en los modelos para la predicción de las fallas, a pesar de que en su caso de estudio no se encuentra como una variable predictora debido a que no existe registro de esta variable.

De manera similar, (Ugarelli, Kristensen, Røstum, Sægrov, & Di Federico, 2009) encuentran que la tasa de bloqueos en una red de alcantarillado en Oslo tiene una relación directa con la edad de las tuberías y una relación inversa con el diámetro y la pendiente de las mismas; sin embargo, encuentran que la relación con la longitud de las tuberías no es fácilmente interpretable en la mayoría de expresiones obtenidas.

Tabla 5-6. Modelos EPR para la predicción de fallas en sistemas de alcantarillado.

No.	Año	Título	Referencia	Resultado de la metodología	Nivel de predicción
1	2006	Modelling sewer failure by evolutionary computing	D. Savic, O. Giustolisi, L. Berardi, W. Shepherd, S. Djordjevic, and A. Saul	Tasa de falla por unidad de longitud de tuberías	Tubería
2	2009	An effective multi-objective approach to prioritization of sewer pipe inspection	L. Berardi, O. Giustolisi, D. A. Savic and Z. Kapelan	Tasa de falla por unidad de longitud de tuberías	Tubería
3	2009	Asset deterioration analysis using multi-utility data and multi-objective data mining	Savic, D Giustolisi, O Laucelli, D	Tasa de falla por unidad de longitud de tuberías	Tubería
4	2007	Multi-Case EPR strategy for the development of sewer failure performance indicators	Berardi, L Kapelan, Zoran	Tasa de falla por unidad de longitud de tuberías	Tubería
5	2009	Statistical analysis and definition of blockages-prediction formulae for the wastewater network of Oslo by evolutionary computing	Ugarelli, Rita Kristensen, Stig Morten Røstum, Jon Sægrov, Sveinung Di Federico, Vittorio	Tasa de falla por unidad de longitud de tuberías	Tubería

En general, el uso de esta técnica de minería de datos resulta de gran utilidad para el entendimiento de la influencia de los parámetros seleccionados para el modelo y la tasa de falla de las tuberías, al igual que es posible garantizar medidas de desempeño (coeficiente de determinación del ajuste) que indican que las funciones se ajustan muy bien a los datos de entrenamiento del modelo. Sin embargo, debido a la flexibilidad en los modelos que se construyen a partir de EPR, es necesario considerar que pueden ocurrir problemas de sobreajuste a los datos y por lo tanto estos no se ajustaran apropiadamente a otros conjuntos de datos. Para evitar este problema se requiere la implementación de otras técnicas que penalicen la flexibilidad del modelo, como se menciona en Savic et al., (2009).

De igual manera, para la construcción de estos modelos se requiere, en todos los casos, realizar un paso de procesamiento de los datos adicional a los mencionados en la sección 8.1.2, debido a la necesidad de seleccionar los rangos de las variables predictoras en los cuales se han registrado fallas, de manera que sea posible encontrar una expresión matemática que relacione estas variables con el colapso o bloqueo de las tuberías. Luego, si se tuviera un conjunto de datos en el cual las tuberías con diámetros mayores a 700 mm no han presentado fallas, el proceso de construcción del modelo

requiere eliminar estas instancias de los datos, para que las relaciones encontradas sean significativas estadísticamente (Savic et al., 2006).

Debido a lo anterior, es posible identificar que para construir modelos confiables a partir del proceso de EPR se requieren conjuntos de datos grandes, al igual que la inclusión de procesos previos y posteriores a la aplicación de la técnica de minería de datos.

5.3.2 Variables predictoras utilizadas en la detección de fallas de tuberías en redes de alcantarillado

Con el propósito de diagnosticar la disponibilidad de información utilizada en los casos de estudio anteriores y las variables explicativas encontradas en diferentes modelos de deterioro aplicados a redes de alcantarillado, se registró las variables reportadas como disponibles en 20 modelos de deterioro diferentes, y también las variables encontradas relevantes en la calibración de los modelos (ver **Tabla 5-8**).

Lo anterior permitió observar la gran variabilidad que existe no solo en la disponibilidad de información en los diferentes casos de estudio en los cuales se han aplicado modelos de deterioro, sino también la gran variabilidad de las variables encontradas como relevantes para la predicción del comportamiento de las tuberías. Además, se observó que únicamente en pocos casos se cuenta con información actualizada y completa sobre variables del entorno que pueden afectar los procesos de deterioro; y en particular, que las variables de las tuberías/sistema que son más comúnmente registradas corresponden al material, edad, longitud, diámetro, pendiente y profundidad de las tuberías (ver **Figura 5-8** y la **Tabla 5-7**).

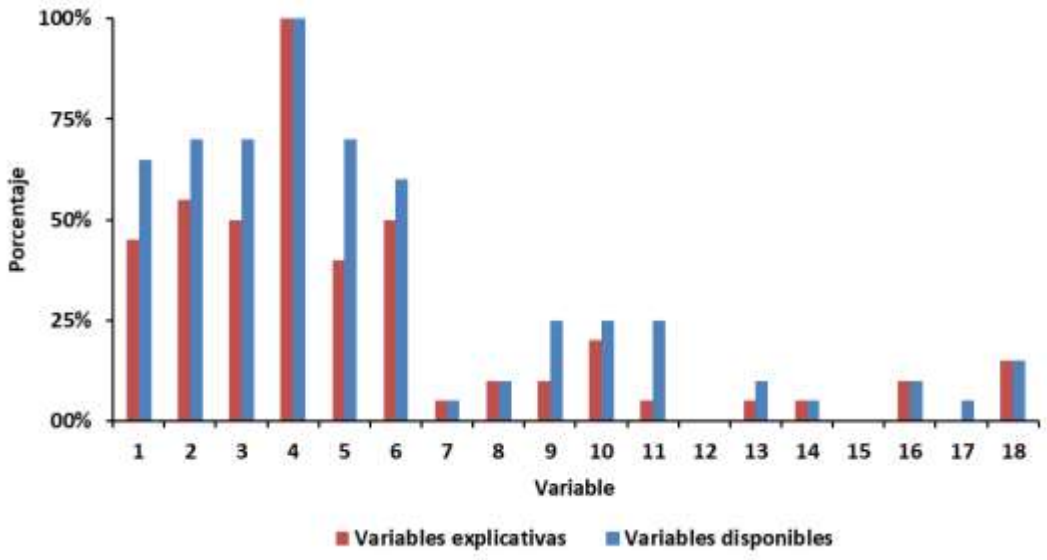


Figura 5-8. Frecuencia de variables disponibles y explicativas en 20 modelos de deterioro en redes de alcantarillado.

Tabla 5-7. Frecuencia de variables disponibles y explicativas en 20 modelos de deterioro en redes de alcantarillado.

No.	Variable	Frecuencia (Variable disponible)	Frecuencia (Variable explicativa)
1	Material	65%	45%
2	Edad de la tubería/Año de instalación	70%	55%
3	Longitud	70%	50%
4	Tamaño/ Diámetro	100%	100%
5	Pendiente	70%	40%
6	Profundidad	60%	50%
7	Tipo de material de lecho	5%	5%
8	Elevación del borde del pozo aguas arriba	10%	10%
9	Elevación del borde del pozo aguas abajo	25%	10%
10	Tipo de sistema	25%	20%
11	Tipo de tubería (Principal o secundaria	25%	5%
12	Rugosidad	0%	0%
13	Localización	10%	5%
14	Número de fallas de tuberías cercanas	5%	5%
15	Historial de falla	0%	0%
16	Flujo/velocidad	10%	10%
17	Proximidad al tranvía	5%	0%
18	Número de propiedades/metro	15%	15%

Por otro lado, se observa que a pesar de encontrar porcentajes similares en la frecuencia de las variables disponibles y las variables determinadas como explicativas para el deterioro de las tuberías, se presentan casos de modelos en los cuales realizar la distinción del mecanismo de falla representa un cambio en las variables predictoras encontradas como relevantes (ej. Savic et al. (2006)). Así mismo, existen diversos casos de estudio en los cuales a pesar de contar con la disponibilidad de información, se realizó una selección preliminar de las variables que podrían afectar el deterioro de las tuberías considerando previas investigaciones en las que se ha evaluado los factores que influyen en este proceso.

Es de gran importancia recordar que la influencia de estos factores en el proceso de deterioro no se puede generalizar, pues a pesar de que comúnmente se encuentran patrones similares en diferentes casos de estudio respecto al incremento de la probabilidad de falla de tuberías cuando se incrementa el valor de una variable (ej. Longitud), también se han reportado excepciones en los cuales la relación encontrada con el estado estructural de las tuberías es inversa, como se presentó en el capítulo 3.3 de este documento.

Tabla 5-8. Covariables disponibles y explicativas para diferentes modelos de deterioro en redes de alcantarillado.

No.	Título	Referencia	Metodología	Covariables del sistema/tubería consideradas	Covariables explicativas
1	Markov Model for Storm Water Pipe Deterioration	(Micevski et al., 2002)	Cadenas de Markov	- Material - Tamaño/Diámetro - Localización	- Material - Tamaño/Diámetro

No.	Título	Referencia	Metodología	Covariables del sistema/tubería consideradas	Covariables explicativas
2	Estimating Transition Probabilities in Markov Chain-Based Deterioration Models for Management of Wastewater Systems	(Baik et al., 2006)	Cadenas de Markov	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente
3	Prioritizing Sanitary Sewers for Rehabilitation Using Least-Cost Classifiers	(Wright et al., 2006)	Regresión logística y análisis discriminante	- Tamaño/Diámetro - Elevación del pozo aguas abajo	- Tamaño/Diámetro - Elevación del pozo aguas abajo
4 (4a)	Modelling sewer failure by evolutionary computing	(Savic et al., 2006)	EPR	- Longitud - Tamaño/Diámetro - Pendiente - Profundidad	- Longitud - Tamaño/Diámetro - Profundidad
5 (4b)	Modelling sewer failure by evolutionary computing	(Savic et al., 2006)	EPR	- Longitud - Tamaño/Diámetro - Pendiente - Profundidad	- Longitud - Tamaño/Diámetro - Pendiente
6 (6a)	An effective multi-objective approach to prioritization of sewer pipe inspection (a. Colapso)	(Berardi et al., 2009)	EPR	- Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Número de propiedades/metro	- Edad - Tamaño/Diámetro - Profundidad - Número de propiedades/metro
7 (6b)	An effective multi-objective approach to prioritization of sewer pipe inspection (b. Obstrucciones)	(Berardi et al., 2009)	EPR	- Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Número de propiedades/metro	- Longitud - Tamaño/Diámetro - Número de propiedades/metro
8	Prediction of Sewer Condition Grade Using Support Vector Machines	(Mashford et al., 2011)	SVM	- Material - Edad - Tamaño/Diámetro - Pendiente - Elevación del pozo aguas arriba -Elevación del pozo aguas abajo	- Edad - Tamaño/Diámetro - Pendiente - Elevación del pozo aguas arriba -Elevación del pozo aguas abajo

No.	Título	Referencia	Metodología	Covariables del sistema/tubería consideradas	Covariables explicativas
9 (9a)	Application of Classification Models and Spatial Clustering Analysis to a Sewage Collection System of a Mid-Sized City	(Jung et al., 2012)	Árbol de clasificación	- Material - Tamaño/Diámetro - Profundidad	- Material - Tamaño/Diámetro - Profundidad
10 (9b)	Application of Classification Models and Spatial Clustering Analysis to a Sewage Collection System of a Mid-Sized City	(Jung et al., 2012)	Redes bayesianas	- Material - Tamaño/Diámetro - Profundidad	- Material - Tamaño/Diámetro - Profundidad
11	Modeling failure of wastewater collection lines using various section-level regression models	(Salman & Salem, 2012)	Regresión logística	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Tipo de sistema	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Tipo de sistema
12	Wastewater pipes in Oslo: from condition monitoring to rehabilitation planning	(Ugarelli et al., 2013)	Cadenas de Markov	- Edad - Tamaño/Diámetro - Tipo de material de lecho - Tipo de sistema	- Tamaño/Diámetro - Tipo de material de lecho - Tipo de sistema
13	Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure	(Harvey & McBean, 2014)	Árbol de clasificación	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Elevación del pozo aguas abajo - Tipo de tubería - Número de fallas de tuberías cercanas	- Edad - Longitud - Tamaño/Diámetro - Profundidad - Número de fallas de tuberías cercanas
14	A Data Mining Tool for Planning Sanitary Sewer Condition Inspection	(Harvey et al., 2015)	Árbol de clasificación	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Tipo de tubería	- Edad - Longitud - Tamaño/Diámetro - Pendiente
15	Predictive risk modelling of real-world wastewater network incidents	(Bailey et al., 2015)	Árbol de clasificación	- Edad - Longitud - Tamaño/Diámetro	- Longitud - Tamaño/Diámetro

No.	Título	Referencia	Metodología	Covariables del sistema/tubería consideradas	Covariables explicativas
				- Pendiente - Velocidad -Número de conexiones/metro	- Pendiente - Velocidad -Número de conexiones/metro
16	Benefits of using basic, imprecise or uncertain data for elaborating sewer inspection programs	(Ahmadi et al., 2015)	Regresión logística	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Tipo de sistema - Localización	- Material - Edad - Longitud - Tamaño/Diámetro - Profundidad - Localización
17	Physical characteristics of pipes as indicators of structural state for decision-making considerations in sewer asset management	(López Kleine et al., 2016)	Análisis de componentes principales (PCA)	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Elevación del pozo aguas arriba - Elevación del pozo aguas abajo - Tipo de tubería	- Edad - Tamaño/Diámetro - Pendiente - Profundidad - Elevación del pozo aguas abajo
18	Identificación de factores de riesgo para la gestión patrimonial óptima de sistemas de drenaje urbano: estudio piloto en la ciudad de Bogotá	(Angarita et al., 2017)	Regresión lineal	- Material - Edad - Longitud - Tamaño/Diámetro - Tipo de sistema - Tipo de tubería	- Material - Tamaño/Diámetro - Tipo de sistema
19	The relevance of sewer deterioration modelling to support asset management strategies	(Caradot, Sonnenberg, et al., 2017)	Cadenas de Markov	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Tipo de sistema	- Material - Edad - Tamaño/Diámetro - Profundidad - Tipo de sistema
20	Sewer Condition Prediction and Analysis of Explanatory Factors	(Laakso, Kokkonen, et al., 2018)	Bosques aleatorios	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Elevación del pozo aguas abajo - Tipo de tubería - Velocidad	- Material - Edad - Longitud - Tamaño/Diámetro - Pendiente - Profundidad - Tipo de tubería - Velocidad

No.	Título	Referencia	Metodología	Covariables del sistema/tubería consideradas	Covariables explicativas
				-Número de propiedades/metro	

5.4 Calidad y cantidad de la información para la construcción de modelos

Una de las principales preocupaciones para la aplicación de métodos de gestión del riesgo cualitativos en diferentes problemas de ingeniería consiste en la capacidad de los métodos de generalizar los resultados en diferentes condiciones a las estudiadas inicialmente y de la alta dependencia que estos pueden tener de la opinión de expertos. Por esto, podría considerarse favorable la utilización de métodos de gestión del riesgo con un enfoque cuantitativo, en los cuales los resultados son más objetivos y no son altamente dependientes de un conocimiento especializado del sistema analizado (Arthur & Crow, 2007). Como se ha presentado en los capítulos 4 y 5 de este documento, en la última década se han aplicado una gran cantidad de métodos cualitativos con el propósito de realizar un mantenimiento proactivo de activos de redes de alcantarillado, debido a la fácil y rápida implementación de la mayoría de las metodologías cuando se cuenta con una base de datos amplia del comportamiento estructural y/o operacional de las redes en los últimos años.

Sin embargo, el uso de este tipo de metodologías encuentra limitaciones cuando se busca su aplicación en zonas o ciudades en las cuales no se cuenta con un registro exhaustivo del estado de los activos de las redes de alcantarillado, y adicionalmente, existe incertidumbre acerca de la calidad de los datos registrados en campo para su futuro uso en estos modelos. Así, al plantear estas metodologías para la realización efectiva del mantenimiento proactivo en redes de alcantarillado en ciudades con bajas tasas de inspección o pocos registros históricos, es altamente importante tener en cuenta la cantidad de datos disponibles y las características específicas que pueden tener los conjuntos de datos de un caso de estudio particular. Teniendo en cuenta lo anterior, es posible establecer tres tipos de inconvenientes que pueden presentarse para el uso adecuado de la información proveniente de inspecciones CCTV en las redes de alcantarillado (Carvalho, Amado, Brito, Coelho, & Leitão, 2018; Rokstad & Ugarelli, 2016; Rokstad, Ugarelli, & Sægrov, 2015; Scheidegger & Maurer, 2012; Van der Steen, Dirksen, & Clemens, 2014):

1. Detalle y estandarización de los códigos de inspección utilizados.
2. Representatividad del conjunto de datos utilizados para la calibración de los modelos respecto a los datos totales de las redes.
3. Cantidad mínima de registros necesarios para la generalización de los resultados.

A continuación se presenta una breve descripción de los casos en que se han investigado los efectos de cada uno de estos inconvenientes y las principales repercusiones que estos tienen sobre la modelación predictiva en redes de alcantarillado.

1. Detalle y estandarización de los códigos de inspección utilizados.

El detalle y la estandarización de los códigos de inspección utilizados para la clasificación estructural de las tuberías en redes de alcantarillado ha sido investigado por autores como Ahmadi, Cherqui, De Massiac, & Le Gauffre (2014b, 2014a); Dirksen et al. (2013); Van der Steen et al. (2014) y Rokstad & Ugarelli (2016), entre otros. En primer lugar, es importante tener en cuenta que la recolección de información mediante inspecciones CCTV es una de los técnicas más utilizadas para evaluar la calidad y funcionabilidad de los sistemas de alcantarillado (Van der Steen et al., 2014), permitiendo dar apoyo a los procesos de toma de decisiones para llevar a cabo mantenimiento reactivo, preventivo y proactivo en las redes (Butler, D., Davies, 2011). Sin embargo, es esperado que la efectividad de la toma de decisiones sea inherentemente dependiente de la precisión y confiabilidad de las predicciones realizadas a partir de los modelos, los cuales se construyen a partir de los registros de las clasificaciones estructurales asignadas a los activos en las inspecciones (Rokstad & Ugarelli, 2016).

Por lo tanto, diversas investigaciones se han enfocado en establecer y cuantificar las repercusiones de la calidad de información registrada en la capacidad de predicción de los modelos de deterioro al ser aplicados a sistemas de alcantarillado. Para esto, los enfoques adoptados han consistido en evaluar la consistencia de los resultados de inspecciones al incrementar el detalle de los códigos de inspección (Van der Steen et al., 2014), evaluar la influencia de la heterogeneidad de los datos en la capacidad predictiva (Rokstad & Ugarelli, 2016) y cuantificar la propagación de incertidumbre que puede generarse desde las variables predictoras recolectadas en la precisión de los modelos de deterioro (Ahmadi et al., 2015; Scheidegger & Maurer, 2012).

En el trabajo realizado por Van der Steen et al. (2014), los autores proponen realizar la comparación de los resultados de inspecciones realizados a partir del código antiguo (NEN33991992) y el nuevo código (NEN33992004) del sistema de alcantarillado holandés, con el propósito de establecer si al incrementar el detalle de los códigos de inspección resulta en una mayor variabilidad en la clasificación estructural asignada a las tuberías. Al analizar la clasificación obtenida para los mismos defectos mediante la evaluación de tuberías con los dos códigos, y comparar los casos en los cuales los defectos son correctamente identificados, los autores encuentran que realizar una caracterización demasiado detallada de los defectos que se pueden identificar en las tuberías conlleva a que ciertos defectos pasen desapercibidos por los inspectores pues se incrementa la dificultad de encontrar un defecto que se ajuste correctamente a las descripciones más específicas establecidas en los códigos. Más aún, resaltan que la caracterización de tuberías mediante defectos

más específicos no conlleva la generación de más información respecto al estado de las mismas sino que provoca que la misma información sea recopilada con mayor variabilidad.

Por otro lado, Rokstad & Ugarelli (2016) analizan la influencia de la ponderación de defectos correspondientes a diferentes mecanismos de falla para establecer la clasificación estructural de las tuberías durante las inspecciones. En particular, destacan que el uso de la condición de las tuberías determinada a partir del criterio anterior puede inhibir la precisión de los modelos de deterioro debido a que: (1) esta condición puede corresponder a la agregación de diferentes modos de falla y (2) la condición de las tuberías puede contener información no relacionada con el deterioro estructural de las mismas. Para su análisis, simulan la condición de un grupo de tuberías de acuerdo a dos modos de falla diferentes, generando un conjunto de datos sintéticos de las inspecciones de la tuberías. Para estudiar el efecto de tener una mayor heterogeneidad en los conjuntos de datos, establecen tres conjuntos de datos para el entrenamiento de un modelo de deterioro, disminuyendo la homogeneidad en cada uno de ellos al agregar progresivamente registros con diferentes modos de falla, realizar la calibración de los modelos y finalmente predecir la condición para todo el conjunto de datos. A partir de esto, los autores establecen como efectivamente los niveles de heterogeneidad e incertidumbre determinan la calidad de las predicciones, ya que se obtienen peores medidas de desempeño, y por lo tanto podría considerarse más importante reducir la heterogeneidad de los registros utilizados para la calibración de los modelos que incluir una mayor cantidad de datos (heterogéneos) (Rokstad & Ugarelli, 2016). Es decir, concluyen que la reducción del desempeño al utilizar modelos de deterioro no se debe principalmente a la incertidumbre de los parámetros o la estructura analítica del modelo sino del uso de una variable respuesta (condición de las tuberías) holística que incorpora diferentes mecanismos de falla.

Finalmente, otros autores como Carvalho et al. (2018), al estudiar la importancia de las variables predictoras en la predicción de fallas de alcantarillado, establecen que futuras investigaciones deben realizarse para evaluar más a fondo si diferentes tipos de falla tienen diversas variables explicativas, puesto que los procesos físicos que conducen a obstrucciones de las tuberías pueden ser significativamente diferentes a los que generan el colapso estructural de la misma.

Dado lo anterior, se observa que es posible establecer una dependencia directa de las técnicas de inspección y evaluación de tuberías de alcantarillado sobre la eficiencia de la modelación predictiva como herramienta para el mantenimiento predictivo de las redes de alcantarillado. Por lo cual, al evaluar la aplicación de modelos de gestión del riesgo, y en particular, de modelos de deterioro para realizar el mantenimiento proactivo de activos en redes de alcantarillado se debe: (1) realizar un diagnóstico de las técnicas de recolección y clasificación de la condición de las tuberías, (2) evaluar los resultados anteriores respecto a las mejores prácticas para el registro de información a partir de la cual sea posible la exploración de patrones, y (3) modificar las normativas y metodologías de evaluación de la condición de tuberías para cuantificar el impacto de esto en la gestión proactiva de

las redes. Estas modificaciones para la mejora de los códigos de inspección deben asegurar (Rokstad & Ugarelli, 2016; Stanić, Langeveld, & Clemens, 2014; Van der Steen et al., 2014):

a. La diferenciación de los diferentes mecanismos de falla en los códigos de inspección
b. La determinación y recolección de las variables necesarias para explicar cada mecanismo de falla
c. La desagregación de los indicadores holísticos utilizados para establecer la condición de las tuberías
d. Evitar detallar excesivamente los defectos, realizar descripciones vagas de cada defecto o combinar defectos que pueden no presentarse conjuntamente.

2. **Representatividad del conjunto de datos utilizados para la calibración de los modelos respecto a los datos totales de las redes.**

El segundo inconveniente o limitación que se puede presentar para el uso eficiente de la información recolectada de las inspecciones corresponde a la representatividad del conjunto de datos utilizados para la calibración de los modelos respecto al comportamiento global de todos los activos de las redes de alcantarillado. Es decir, concierne a la capacidad de los modelos de deterioro utilizados para generalizar el comportamiento de todos los activos, considerando las características del conjunto de datos utilizados para el entrenamiento de modelos. El problema anterior ha sido principalmente abordado (1) teniendo en cuenta si todos los patrones que podrían observarse en el conjunto completo de datos pueden capturarse a partir de un conjunto de datos limitados (es decir, preocupación por los defectos y tipos de falla disponibles en los registros de inspecciones) y (2) considerando el rango que toman las características de las tuberías inspeccionadas y clasificadas respecto al rango de cada una de esas características en el conjunto de datos completos (es decir, preocupación por las variables explicativas (Ahmadi, Cherqui, Aubin, & Le Gauffre, 2016; Scheidegger & Maurer, 2012)

La primera preocupación, busca entender si la naturaleza de los procesos de inspección de las tuberías logran la recolección de un conjunto de datos característico de los activos de las redes, considerando que en la mayoría de los casos, los costos de inspección de toda la red son muy altos para las empresas prestadoras del servicio y por lo tanto, únicamente se cuenta con un porcentaje inspeccionado de la red (Ahmadi et al., 2016). En general, este conjunto inspeccionado será lo que se conoce como una muestra de la red y se espera que refleje de la mejor manera las características de todos los activos de la red. Se considera que una muestra que es una imagen apropiada del conjunto total de activos corresponde a una muestra representativa (Cochran, 1977; Lohr, 2010). Pocos autores como Ahmadi et al. (2016) han investigado lo anterior a partir de la aplicación a información de tuberías en redes de alcantarillado; en su caso, llevando a cabo un ejercicio numérico en el cual se realiza la selección de la muestra del conjunto de datos total mediante tres métodos de muestreo diferentes. A partir de sus resultados, los autores encuentran que la selección de una muestra a partir de muestreo aleatorio simple no presenta diferencias significativas respecto a la

selección mediante un método de muestreo más avanzado como muestreo aleatorio estratificado. Así mismo, establecen como existe una dependencia de las variables explicativas que pueden encontrarse relevantes para los modelos de deterioro y el tamaño de la muestra con el cual se realice la calibración de los modelos; por lo cual, las conclusiones obtenidas respecto a los factores que afectan el deterioro de las tuberías deben ser interpretados con cautela, buscando la posible calibración de los modelos con diferentes tamaños de los conjuntos de datos (Ahmadi et al., 2016).

Por otro lado, en cuanto al rango de valores que toman las características de la tuberías inspeccionadas respecto al total de activos de las redes, autores como Scheidegger & Maurer (2012) investigaron las repercusiones de la calibración de modelos de deterioro al cambiar la edad promedio de las tuberías que componen el conjunto de datos sintético utilizado para el entrenamiento de estos modelos. Entre sus hallazgos, encuentran que la calibración de modelos para sistemas de alcantarillado con una edad de construcción menor (en los cuales existe una menor cantidad de activos en condiciones de falla o cercanos a la falla) se presenta más incertidumbre respecto a la estimación de los parámetros de los modelos. No obstante, los autores también resaltan que esta dependencia puede estar ligada a la estructura de los modelos y los métodos de estimación de parámetros (Scheidegger & Maurer, 2012).

3. Cantidad mínima de registros necesarios para la generalización de los resultados.

Investigaciones recientes en la línea de investigación de la modelación predictiva en redes de alcantarillado han buscado entender y cuantificar la aplicabilidad efectiva de modelos de deterioro considerando las limitaciones de la cantidad de registros de inspecciones CCTV en estos sistemas. En particular, autores como Ahmadi et al. (2016, 2015) y Scheidegger & Maurer (2012) han evaluado las implicaciones del uso de diferentes tamaños de muestra para la calibración de los modelos en el desempeño final de estos. Sin embargo, estudios similares se han realizado en bases de datos de clientes de una empresa o conjuntos de datos de pacientes con diabetes y cáncer (Morgan, Dougherty, Hilchie, & Carey, 2003; Udhayakumarapandian & Chandrasekaran, 2016).

En estos últimos estudios, el principal objetivo ha consistido en determinar la relación entre la capacidad de predicción de modelos de minería de datos (arboles de decisión) y el tamaño de la muestra de entrenamiento mínima requerida para obtener una precisión adecuada. Morgan et al. (2003) aplican modelos de árboles de decisión a una base de datos de clientes en una empresa de computación utilizando tres técnicas de muestreo diferentes: Muestreo aritmético progresivo, muestreo geométrico progresivo y muestro dinámico progresivo. Los autores encuentran que, en general, la precisión de los modelos incrementa a una tasa decreciente con el aumento del tamaño de la muestra y que es posible ajustar una curva potencial a la relación entre la precisión de los modelos y el tamaño de la muestra (ver **Figura 5-9**). Así, destacan que, dadas las características de su base de datos, la precisión de los modelos tiende a tomar un valor asintótico a medida que incrementa el tamaño de la muestra. Sin embargo, también establecen como la variación del

desempeño de los modelos en función del tamaño de la muestra es dependiente de las características del conjunto de datos utilizado y la variable predictora, y por lo tanto, la evaluación del tamaño óptimo de una muestra de datos debe realizarse para cada caso de estudio (Morgan et al., 2003). La investigación realizada por Udhayakumarapandian & Chandrasekaran (2016), encuentra resultados similares en cuanto al aumento de la precisión a medida que se incrementa el tamaño de la muestra en una base de datos de pacientes con diabetes y cáncer; pero en su caso no es posible ajustar esta relación a una curva potencial.

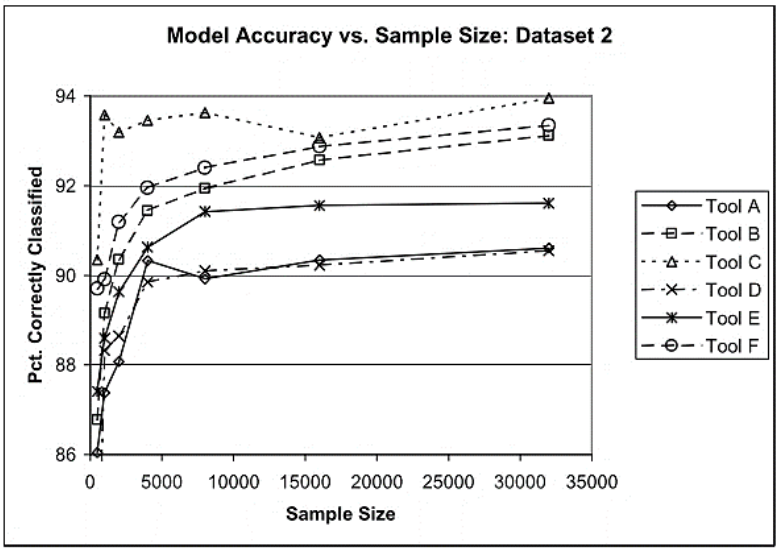


Figura 5-9. Precisión del modelo vs. Tamaño de la muestra para herramientas de minería de datos de árboles de decisión. Tomado de (Morgan et al., 2003).

Ahora bien, en el contexto del problema de clasificación de tuberías de acuerdo a su estado estructural, Ahmadi et al. (2016) investigaron la relación entre el tamaño de la muestra para la calibración de modelos de regresión logística y la precisión de los modelos. En su estudio, la precisión es comparada mediante la estimación de la proporción de las tuberías en mal estado realizada con los modelos de regresión logística en comparación con el valor real conocido. Para realizar esto, generan un conjunto de datos sintéticos a partir de una base de datos similar a la red de alcantarillado de Greater Cincinnati, EEUU. Mediante el uso de tres métodos de selección de la muestra (muestreo aleatorio simple, muestreo aleatorio estratificado y muestro aleatorio estratificado con asignación proporcional) estiman la proporción de tuberías en mal estado (p) para todo el conjunto de datos, mediante la creación de 1000 modelos diferentes con simulaciones de Montecarlo, para cada tamaño de muestra (Ahmadi et al., 2016). En sus resultados encuentran que a partir de una muestra aproximadamente de 1000 registros es posible estimar p con un error de $\pm 3\%$ y una confiabilidad del 95%. Así mismo, concluyen que para las tres técnicas de muestreo

utilizadas, al incrementar el tamaño de la muestra se genera una disminución de la dispersión de los valores de p estimados.

Más aún, los autores resaltan que debido a que se encuentran resultados similares con las tres técnicas de muestreo y, considerando la dificultad de aplicar los dos métodos de muestreo más avanzados, se puede considerar que el muestreo aleatorio simple de los datos es la mejor técnica de selección (Ahmadi et al., 2014a).

Así, si se quisiera estimar el tamaño mínimo de la muestra requerida de acuerdo a la proporción de tuberías en mal estado que se estima tiene el conjunto de datos total, se pueden utilizar las siguientes ecuaciones (Lohr, 2010):

$$n = \frac{p(1-p)}{V(p) + \frac{p(1-p)}{N}} \qquad \text{Ecuación 5-39}$$

$$V(p) = \left(\frac{e}{1.96}\right)^2 \qquad \text{Ecuación 5-40}$$

Donde:

- $p \coloneqq proporción\ real\ de\ tuberías\ e\ mal\ estado$
- $N \coloneqq Número\ de\ tuberías\ de\ todo\ el\ conjunto\ de\ datos$
- $e \coloneqq margen\ de\ error\ con\ el\ cual\ se\ quisiera\ estimar\ la\ proporción\ real\ de\ tuberías\ en\ mal\ estado$

Generalmente, se utiliza un valor de $e = 3\%$ con una confiabilidad asociada de 95% (Lohr, 2010). La **Figura 5-10** muestra el comportamiento del tamaño mínimo de la muestra requerido al utilizar muestreo aleatorio simple en función de la proporción estimada de tuberías en mal estado. Este comportamiento corresponde a la **Ecuación 5-39**, y se observa que los valores máximos se obtienen cuando la proporción de tuberías es igual a 0.5.

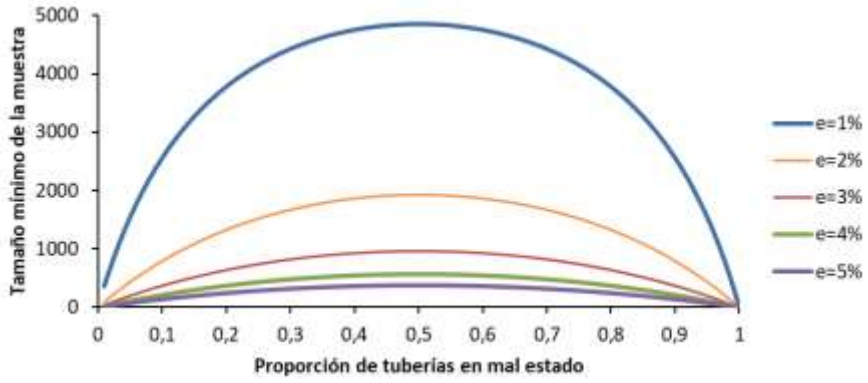


Figura 5-10. Tamaño de la muestra n en función de la proporción estimada de tuberías en mal estado p para diferentes valores del margen de error e.

Teniendo en cuenta las estudios y hallazgos descritos en los capítulos 5.15.3 y 5.4, se podría considerar que en zonas en las cuales se tienen condiciones de pocos registros históricos y una necesidad latente de realizar mantenimiento proactivo en las redes de alcantarillado para garantizar la prestación de un servicio de alta calidad en todo momento, se podrían considerar alternativas en las cuales se establezca un marco de modelación que permita a los directores y/o tomadores de decisiones de las empresas prestadoras de servicio interpretar el comportamiento a nivel global del sistema, y así, verificar y relacionar información especializada de los sistemas de alcantarillado con las estimaciones de patrones obtenidas a partir de la información histórica. Así mismo, sería necesario establecer directrices que garanticen el uso de los resultados obtenidos en la modelación para la priorización de rehabilitación de tuberías y que a su vez, estos últimos también contribuyan activamente a la retroalimentación de los modelos predictivos evitando el uso ineficiente de los mismos a medida que se realizan modificaciones a los sistemas inherentemente relacionadas a los cambios en las ciudades.

6 GESTIÓN ACTUAL DE ACTIVOS DE ALCANTARILLADO DE LA EAB

El presente capítulo de este documento presenta un análisis de la gestión actual que realiza la Empresa de Acueducto y Alcantarillado de Bogotá (EAB) a los activos de sus redes de alcantarillado, al igual que se realiza un análisis de viabilidad de la aplicación de modelos de minería de datos, considerando la frecuencia de inspecciones en las redes de alcantarillado, la información disponible (topológica y de inspecciones CCTV) y finalmente, se estudia el efecto de la cantidad de información utilizada para el entrenamiento de modelos de minería de datos para la predicción general del estado estructural de tuberías de alcantarillado mediante un caso de estudio sintético. Para esto, se considera la información reportada en los informes de gestión de la EAB, las normas técnicas de servicio bajo las cuales se rigen los procedimientos de inspección y rehabilitación de redes de alcantarillado y las investigaciones científicas recientes en las cuales se ha estudiado el comportamiento estructural de acuerdo a sus características físicas.

Mediante la realización del análisis anterior se busca estudiar la viabilidad de la aplicación de modelos de minería de datos para la estimación del comportamiento estructural de tuberías de alcantarillado en la ciudad de Bogotá, considerando las limitaciones de la cantidad y calidad de la información.

6.1 Descripción general de la EAB y sus métodos de gestión

La ciudad de Bogotá es la capital del departamento de Cundinamarca y a su vez capital de la República de Colombia. Se encuentra constituida por 20 localidades y cuenta con una población aproximada de 8’000.000 de habitantes, de acuerdo a las estimaciones del Departamento Administrativo Nacional de Estadística (DANE) y la Secretaria Distrital de Planeación (El Tiempo, 2017).

La Empresa de Acueducto y Alcantarillado de Bogotá (EAB) es la empresa pública encarga de la prestación de los servicios de Acueducto y Alcantarillado sanitario y pluvial en la ciudad desde aproximadamente 1955. La prestación de los servicios de la empresa sigue un modelo empresarial haciendo del concepto de zonas operativas, según el cual, la ciudad se encuentra divida en cinco (5) zonas, con el propósito de facilitar la asistencia de las necesidades de los ciudadanos. De esta manera, cada zona se encarga de la operación y mantenimiento de las redes menores de acueducto y alcantarillado en un área específica de la ciudad (Empresa de Acueducto de Bogotá E.S.P., 2019b). La disposición inicial del actual sistema de alcantarillado de la ciudad corresponde a la implementación del Plan Maestro de Acueducto y Alcantarillado de 2006 (Empresa de Acueducto de Bogotá E.S.P., 2006). Adicionalmente, los informes de gestión de la empresa a partir del año 2008, registran el progreso y logros alcanzados de las actividades trazadas para cumplir con las

metas del plan de desarrollo y el plan general estratégico (Empresa de Acueducto de Bogotá E.S.P., 2019a).

De acuerdo a los informes de gestión de la EAB de los últimos tres años (2016-2018) se observa que los procesos de gestión de las redes de alcantarillado se encuentran principalmente enfocados en la construcción, renovación, rehabilitación o reposición de los activos de sus redes. En particular, respecto a las actividades de renovación y rehabilitación de alcantarillado, se distingue que en el año 2016 se enfocan principalmente los recursos en el mantenimiento correctivo de las redes (Empresa de Acueducto de Bogotá E.S.P., 2016); mientras que en los años 2017 y 2018 se ha dado paso a medidas preventivas para la gestión de los activos de alcantarillado, principalmente encaminadas en la identificación y corrección de conexiones erradas, la optimización de tiempos de atención a solicitudes de quejas y reclamos de usuarios y, mediante programas sistemáticos de inspección del sistema de alcantarillado con equipo CCTV para conocer el estado de las redes de alcantarillado (Empresa de Acueducto de Bogotá E.S.P., 2017, 2018).

Dado lo anterior, se puede observar que la gestión de activos en las redes de alcantarillado en la ciudad de Bogotá se ha movilizado en los últimos tres años, de un mantenimiento correctivo de sus redes hacia un mantenimiento más proactivo, con el cual se busca optimizar los servicios de alcantarillado pluvial y sanitario y así, mitigar las reclamaciones de los usuarios. Sin embargo, aún es posible identificar oportunidades de mejora en las metodologías de gestión de redes de alcantarillado que permitan interpretar y comprender el comportamiento de las tuberías en los sistemas, considerando los recientes esfuerzos en incrementar los porcentajes inspeccionados de sus redes.

Ahora bien, considerando la alta importancia que tiene la evaluación de la condición de tuberías en el desempeño de los diferentes modelos predictivos mencionados en los capítulos 4 y 5, a continuación se realiza una breve descripción de las normas técnicas utilizadas para la clasificación de las tuberías y las variables registradas en los procesos de inspección mediante CCTV.

6.2 Datos disponibles de la red de alcantarillado de Bogotá

Con el propósito de identificar la aplicabilidad de los modelos predictivos estudiados anteriormente y la viabilidad de predecir correctamente el estado estructural de las tuberías, es necesario establecer las características registradas de los activos en las redes de alcantarillado.

6.2.1.1 Características físicas y/o topológicas

Estos datos corresponden a las características de las tuberías que son registrados y almacenados en el sistema de información geográfico de la EAB para las redes de alcantarillado sanitario y combinado.

Tabla 6-1. Características físicas y/o topológicas disponibles EAB. Creado a partir de la documentación de la EAB.

Variable	Unidades o valores posibles
Diámetro	[m]
Longitud	[m]
Profundidad media	[m]
Pendiente	[%]
Edad	Fecha de instalación [DD/MM/AAAA]
Material	Desconocido, (1) Concreto sin refuerzo, (2) Concreto reforzado, (3) Concreto extra reforzado, (4) Concreto reforzado revestido con lámina de polietileno, (5) PVC, (6) PVC Perfil cerrado, (7) PVC Perfil abierto, (8) Gres, (9) Poliéster reforzado con fibra de vidrio (GRP), (10) Polietileno, (11) Acero, (12) Ladrillo, (18) Revestimiento en concreto con baldosas, (19) Polietileno de alta densidad con pared estructural corrugada, (22) Otro
Material espacio público	Desconocido, (1) Sin pavimento, (2) Pavimento flexible, (3) Pavimento rígido, (4) Pavimento articulado, (5) Zona verde, (6) Otro, (99) No aplica
Tipo de sistema	(0) Sanitario, (1) Pluvial, (2) Combinado
Tipo de sección transversal	Ovoide, (1) Herradura, (2) Bóveda, (3) Elipse horizontal, (4) Elipse vertical, (5) Circular, (6) Trapezoidal, (7) Rectangular, (8) Rectangular triangular, (9) Rectangular redondeado, (10) Triangular, (11) Natural rectangular, (12) Box culvert
Estado	(0) Desconocido, (1) Fuera de servicio, (2) Abandonado, (3) En servicio

6.2.1.2 Información de inspecciones CCTV

Estos datos corresponden a las características físicas y/o topológicas que se especifican en el formato de inspección de redes de alcantarillado según la norma técnica NS-058, que se describe en más detalle en la sección 6.3 de este documento. A partir de este formato se realiza la recolección de los datos físicos de las tuberías inspeccionadas y se asignan los correspondientes grados operacionales y estructurales de cada tramo caracterizado de acuerdo a los defectos presentes.

Tabla 6-2. Variables registradas en las inspecciones mediante CCTV. Creado a partir de la documentación de la EAB.

Variable	Unidades o valores posibles
Id tramo	[-]
Fecha de inspección	[DD/MM/AAAA]
Tipo de sistema	Sanitario, (1) Pluvial, (2) Combinado
Material	Concreto, Gres, Mampostería, Polietileno, PVC, GRP
Estado de la vía	Afirmada, Pavimento asfaltado, Pavimentada concreto, Verde
Longitud	[m]
Diámetro	[m]
Profundidad	[m]
Clima	Lluvioso, Seco
Sentido de inspección	Fujo, Contraflujo
Puntaje operacional	[-]
Grado operacional	[-]
Puntaje estructural	[-]
Grado estructural	[-]

La tabla anterior permite observar que las variables registradas en el proceso de inspección corresponden principalmente a las características físicas de las tuberías y registran pocas variables relacionadas con el entorno y/o las condiciones ambientales que se presentan en la Figura 3-9.

6.3 Códigos para la evaluación de la condición de las tuberías en la ciudad de Bogotá

Para la evaluación de la condición de los activos en las redes de alcantarillado de la ciudad de Bogotá existen principalmente dos normas técnicas que establecen la terminología y los procedimientos que se deben llevar a cabo para la determinación de los indicadores de desempeño de estos activos. Estas normas son:

- NS-058: Aspectos técnicos para inspección de redes y estructuras de alcantarillado.
- NS-061: Aspectos técnicos para la rehabilitación de redes y estructuras de alcantarillado.

En la primera, NS-058 se establecen los lineamientos bajo los cuales se debe realizar la inspección de las estructuras para determinar su estado, al igual que los tipos de defectos que pueden ser identificados y la severidad de cada uno de ellos. Mientras que la NS-061 establece la terminología de los tipos de fallas que pueden generarse a partir de los defectos identificados en la inspección, al igual que los criterios para el diagnóstico y selección de la técnica de rehabilitación considerando, entre otros factores, la afectación del entorno para establecer la prioridad de ejecución de las labores (Empresa de Acueducto de Bogotá E.S.P., 2001, 2010).

A continuación se presenta una breve descripción de los aspectos más relevantes de cada una de las normas anteriores, enfatizando principalmente en la codificación de los defectos registrado por

la EAB debido a la importancia que estos tienen en el desempeño posterior de diferentes modelos predictivos (Ana & Bauwens, 2010; Rokstad et al., 2015; Van der Steen et al., 2014).

De acuerdo con la norma técnica NS-058, la evaluación de la condición de las tuberías se realiza diferenciando los aspectos estructurales y los aspectos operacionales, con lo cual se obtienen dos clasificaciones del estado de las tuberías: el estado estructural y el estado operacional. El estado estructural de las tuberías evalúa los defectos observados durante la inspección relacionados con deformaciones existentes o el estado límite de la capacidad estructural del sistema y evalúa mediante un sistema de puntajes el grado o nivel de deterioro con respecto a la probabilidad de colapso del mismo, obteniendo una clasificación de 1 a 5, en donde 1 corresponde a tubos sin la presencia de defectos o pequeños defectos y 5 corresponde a tubos que se encuentran colapsados o a punto de colapsar. Por otro lado, el estado operacional de las tuberías evalúa los defectos que puedan disminuir la capacidad en la conducción de los flujos establecidos en el diseño de las tuberías debido a la reducción de la sección transversal, obteniendo una clasificación de 1 a 5, en donde 1 corresponde a la menor cantidad de obstrucciones por unidad de longitud y 5 hace referencia a la mayor cantidad de obstrucciones por unidad de longitud.

Las siguientes tablas presentan un resumen de los defectos estructurales y operacionales considerados para la evaluación de la condición de tuberías. La Tabla 6-3y la Tabla 6-4 presentan los valores registrados y asignados en la evaluación de los defectos estructurados, y la

Tabla 6-3. Resumen de los defectos registrados para la clasificación estructural de las tuberías. Adaptado de (Empresa de Acueducto de Bogotá E.S.P., 2010)

Defecto estructural	Descripción	Código	Calificación/ Severidad	Información adicional
Deformación o deflexión	Variación en la dimensión vertical u horizontal del tubo. La sección transversal de la tubería se ha deformado	1.1.1.1	10, 20, 80, 165	Orientación de la deformación: A, B, C
Fisura/Grieta/ Fractura	- Fisura: Separación superficial <=50% - Grieta: Separación superficial <50% y <100% - Fractura: Rotura >= 100%	1.1.1.2	2, 10, 40, 80	Naturaleza de la observación: A, B, C
Rotura o Colapso	Hueco, abertura o partes ausentes	1.1.1.3	80, 165	A, B
Material de sello	Todo o parte del material usado para sellar una junta entre dos tubos esta entre la tubería	1.1.1.4	1, 2, 5, 8	Tipo de material de sello: A, B, C

Defecto estructural	Descripción	Código	Calificación/ Severidad	Información adicional
introducido en la tubería				
Junta desplazada	Las tuberías adyacentes se desplazan de su posición prevista. Los desplazamientos longitudinales de menos de 10 mm no se registran.	1.1.1.5	1, 2, 80	Tipo de desplazamiento: A, B
Daños superficiales	La superficie de la tubería se ha dañado por acción química o mecánica	1.1.1.6	5, 20, 120, 165	Tipo de daño: A – O y Z

De acuerdo a la caracterización de cada defecto de acuerdo a su información adicional (orientación de la deformación, tipo de material de sello, etc.) y a la severidad del defecto encontrado, se asigna un puntaje a la tubería para cada defecto y se calcula la calificación estructural final como el máximo puntaje asignado entre todos los defectos. Así, el grado estructural se asigna de acuerdo a la Tabla 6-4.

Tabla 6-4. Asignación del grado estructural según el puntaje obtenido. Adaptado de (Empresa de Acueducto de Bogotá E.S.P., 2010)

Puntaje máximo	Grado Estructural
< 10	1
10 – 39	2
40 – 79	3
80 – 164	4
165 +	5

Tabla 6-5. Resumen de los defectos registrados para la clasificación operacional de las tuberías. Adaptado de (Empresa de Acueducto de Bogotá E.S.P., 2010)

Defecto operacional	Descripción	Código	Calificación/ Severidad	Información adicional
Obstrucción por conexión	Tubo conector proyectado en la tubería, que obstruye la sección transversal de ésta.	1.1.2.1	1, 2, 5, 8, 10	Reducción en el área transversal en porcentaje
Raíces	Raíces de árboles u otras plantas crecen en la tubería por causa de los defectos en las conexiones o juntas.	1.1.2.2	1, 2, 4, 5, 10	Tipo de raíz: A, B y C. Reducción en el área transversal en porcentaje

Defecto operacional	Descripción	Código	Calificación/ Severidad	Información adicional
Depósitos pegados, sedimentados o ingreso de suelo	Material pegado a la pared de la tubería, depositado en la batea o suelo que se introduce en la tubería.	1.1.2.3	1, 2, 5, 8, 10	Tipo de material: A, B y C. Grados de obstrucción de la tubería: A, B y C
Otros obstáculos	Objetos en la tubería que obstruyen el área de la sección transversal.	1.1.2.4	10	Descripción de la observación
Infiltración	Ingreso de agua a través de las paredes de la tubería, de las uniones o de defectos	1.1.2.5	3, 5, 10	Alcance del flujo: A, B, C y D

Tabla 6-6. Asignación del grado operacional según el puntaje obtenido. Adaptado de (Empresa de Acueducto de Bogotá E.S.P., 2010)

Puntaje medio (total tramo/long)	Puntaje máximo	Grado Operacional
< 0.5	< 1	1
0.5 – 0.9	1 – 1.9	2
1 – 2.4	2 – 4.9	3
2.5 – 4.9	5 – 9.9	4
5+	10+	5

En la tabla anterior, la calificación final se obtiene como la sumatoria de los defectos observados en el tramo por unidad de longitud considerando la caracterización y severidad de cada defecto encontrado para la asignación de un puntaje (Empresa de Acueducto de Bogotá E.S.P., 2010).

Ahora bien, de acuerdo a los grados operacionales y estructurales determinados para cada uno de los tramos inspeccionados, la EAB recomiendan las acciones a tomar y los periodos de tiempo en que deben realizarse. Así, a partir del grado 3 la EAB sugiere la ejecución de acciones correctivas y/o preventiva, de la siguiente manera: para las tuberías que se encuentren en el grado 3 se deben realizar reparaciones puntuales de acuerdo a los defectos encontrados, para las tuberías que se encuentren en el grado 4 se deben tomar medidas preventivas o correctivas con el fin de evitar el colapso, y para las tuberías en el grado 5 se deben tomar medidas de emergencia y rehabilitar inmediatamente para evitar daños adicionales (Empresa de Acueducto de Bogotá E.S.P., 2010).

Finalmente, con la evaluación de la condición obtenida a partir de la inspección de las tuberías y la afectación del entorno se debe establecer el método de rehabilitación a utilizar y definir las zonas de intervención. La priorización de labores de acuerdo a la afectación del entorno se realiza

considerando una matriz de priorización, la cual permite evaluar, por sectores, la prioridad de atención del trabajo de rehabilitación (Empresa de Acueducto de Bogotá E.S.P., 2001).

Tabla 6-7. Aspectos considerados para la priorización de actividades según la afectación al entorno. Adaptado de (Empresa de Acueducto de Bogotá E.S.P., 2001)

Agente	Afectación
Socioeconómico	- Espacio público - Comunidad - Salud y seguridad
Técnico	- Condiciones normales de funcionamiento Sistema de alcantarillado - Alteración al normal funcionamiento de otros sistemas de servicios
Contaminación	- Suelo - Agua - Aire
Área	- Densidad poblacional - Uso del suelo - Topografía - Entorno

6.4 Frecuencia de inspecciones de tuberías en las redes de alcantarillado

Estudiar la cantidad y la frecuencia de las inspecciones de los activos en las redes de alcantarillado en la ciudad de Bogotá, también es de relevancia al analizar la viabilidad de implementar un mantenimiento proactivo en las redes considerando modelos de deterioro de los activos. La cantidad y calidad de la información registrada de los activos de alcantarillado es un aspecto que no puede ser ignorado al considerar la implementación de este tipo de modelos como ha sido mencionado por diversos autores como Ahmadi, Cherqui, Aubin, & Le Gauffre (2016), Rokstad & Ugarelli (2016) y Scheidegger & Maurer (2012), entre otros.

Así mismo, analizar el tipo de actividades de mantenimiento y la frecuencia con que las realiza la EAB permite identificar los cambios en la planificación y estrategias de gestión adoptadas en los últimos años y analizar la viabilidad económica, técnica y de cultura empresarial de implementar la modelación predictiva como una medida eficiente, eficaz y de fácil uso para la gestión de sus redes de alcantarillado.

A partir de los informes de gestión de la EAB de los últimos cinco años se identificó que las actividades para la medición de la eficiencia operacional de los años 2017 y 2018 incluyó la verificación de la calidad de la prestación del servicio considerando el índice de reclamación operativa de alcantarillado. De acuerdo a estos informes, el cumplimiento del 0.17% y 0.16% para los años 2017 y 2018 respectivamente, de la meta de 0.3% se cumple gracias a verificación de PQR's, al proceso de reposición de tapas de pozos de inspección y al programa sistemático de inspección del sistema de alcantarillado con equipo CCTV.

Tabla 6-10 se presentan los porcentajes planeados y realmente inspeccionados por la EAB en cada una de sus zonas respecto a su longitud. Para determinar estos porcentajes se establecieron las longitudes de las redes de alcantarillado sanitario y pluvial a partir de la información de catastro registrada en el sistema de información geográfica de la EAB en el año 2018. Estas longitudes se presentan en la Tabla 6-9.

Tabla 6-8. Longitudes de inspección planeados y reales – Red de alcantarillado de Bogotá. Tomado de (Empresa de Acueducto de Bogotá E.S.P., 2017, 2018)

	Nov 2017		Dic 2018	
Zona	**Long inspección planeado [Km]**	**Long inspección real [Km]**	**Long inspección planeado [Km]**	**Long inspección real [Km]**
1	21	32.1	21	6.8
2	26.6	31.4	14	13.9
3	14	22.6	7	22.7
4	39.5	31.6	14	14.1
5	10.8	13.9	21	4.5
Total	111.9	131.6	77	62

Tabla 6-9. Longitud redes de alcantarillado por zonas.

Zona	**Longitud [Km]**		
	Alc. Sanitario	**Alc. Pluvial**	**Total**
1	1349.59	826.45	2176.05
2	1279.17	429.94	1709.11
3	1550.04	468.34	2018.38
4	1465.32	584.55	2049.87
5	1111.41	584.88	1696.29
Total	6755.53	2894.16	9649.69

Tabla 6-10. Porcentajes de inspección planeados y reales – Red de alcantarillado de Bogotá. Cálculos propios a partir de (Empresa de Acueducto de Bogotá E.S.P., 2017, 2018)

Año		**Nov 2017**		**Dic 2018**	
Zona	**Longitud [Km]**	**% inspección planeado**	**% inspección real**	**% inspección planeado**	**% inspección real**
1	2176.05	0.97 %	1.48 %	0.97 %	0.31 %
2	1709.11	1.56 %	1.84 %	0.82 %	0.81 %
3	2018.38	0.69 %	1.12 %	0.35 %	1.12 %
4	2049.87	1.93 %	1.54 %	0.68 %	0.69 %
5	1696.29	0.64 %	0.82 %	1.24 %	0.27 %
Total	9649.69	1.36 %	1.16 %	0.8 %	0.64 %

Como se observa en las tablas anteriores, los porcentajes de inspección planeado y realmente inspeccionados para toda la red son similares para ambos años, a pesar de no cumplirse los porcentajes planeados. Sin embargo, como se presenta en la

Tabla 6-10 y la Figura 6-1, estos porcentajes no se cumplen de manera uniforme para cada zona, sino que por el contrario se presentan porcentajes muy desiguales de las inspecciones planeadas y las inspecciones reales, lo cual resulta en un porcentaje total inspeccionado de la red cercano al planeado.

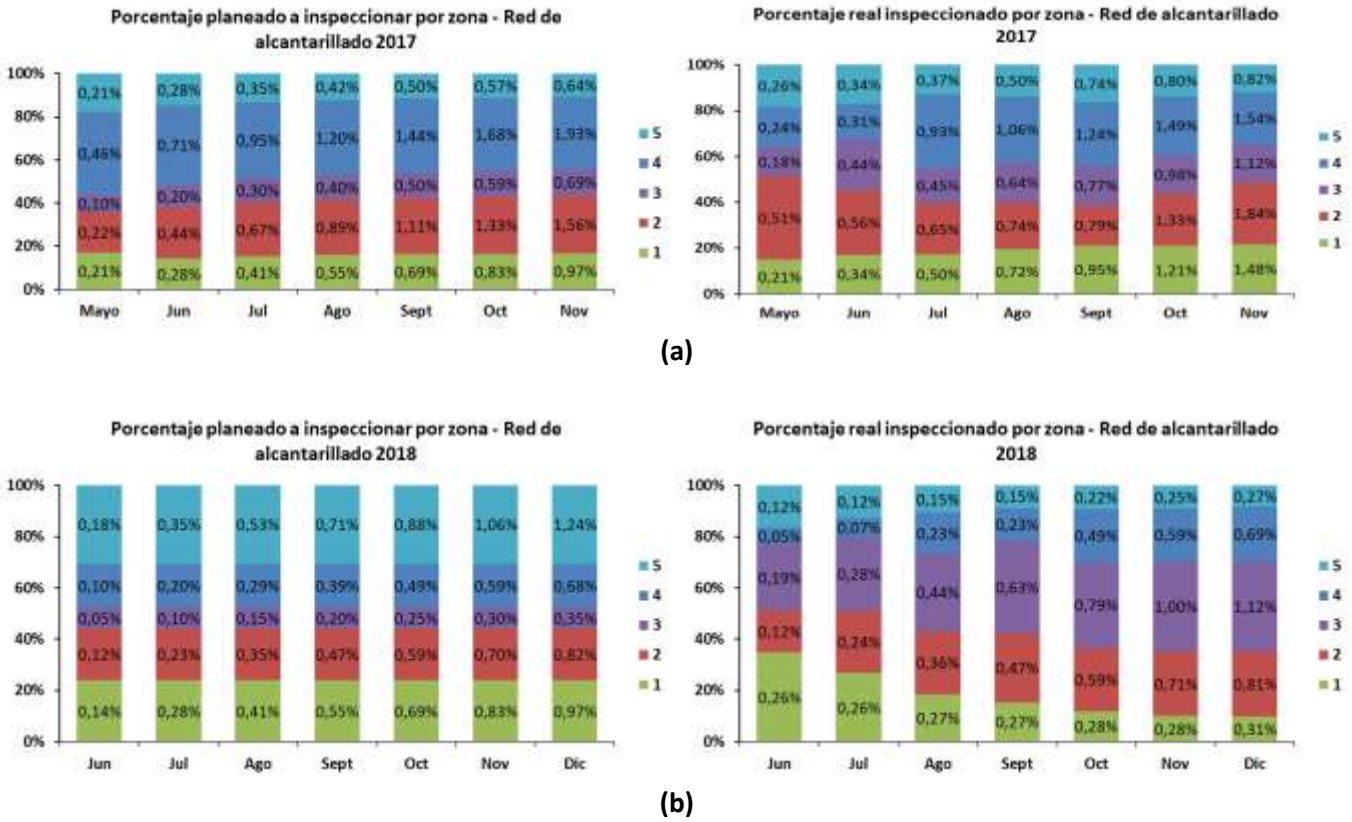


Figura 6-1. Inspección por zonas de la red de alcantarillado de Bogotá respecto a su longitud. (a) 2017, (b) 2018.

Así mismo, también es relevante considerar la magnitud inspeccionada de las redes en comparación con los valores reportados en otros países, los cuales suelen tener tasas de inspección significativamente más altas. Estos valores de inspección reportados para los años 2017 y 2018 son consistentes con lo esperado, pues se ha estimado que la cobertura de inspección por año en la red de alcantarillado de Bogotá corresponde a 2%, lo cual significa que el tiempo aproximado entre dos inspecciones es 50 años (López Kleine et al., 2016).

7 APLICACIÓN DE METODOS DE MINERÍA DE DATOS A UN CASO DE ESTUDIO SINTÉTICO EN LA CIUDAD DE BOGOTÁ

Las empresas prestadoras de servicios de agua potable y alcantarillado en Colombia tradicionalmente han realizado el mantenimiento y operación de sus sistemas con un enfoque reactivo de mantenimiento, es decir, atendiendo el problema después de que este se presenta (López Kleine et al., 2016). Sin embargo, teniendo en cuenta el contexto anterior, se puede observar que en la ciudad de Bogotá, las técnicas de mantenimiento han buscado en los últimos años dar pasos hacia la implementación de actividades preventivas y proactivas que permitan la adecuada gestión de sus servicios y el incremento de indicadores de buen servicio a sus usuarios en los sistemas de alcantarillado.

Lo anterior podría interpretarse como un primer paso hacia el interés y aceptación de los buenos resultados económicos y de servicio que pueden garantizarse al realizar una gestión proactiva de sus sistemas, en reemplazo o como complemento, a sus enfoques reactivos. Algunas investigaciones como las de Angarita et al. (2017); Caradot, Hernandez, Sonnenberg, Torres, & Rouault (2018); López Kleine et al. (2016) y Torres, Rodríguez, & Leitão (2017) han buscado demostrar la capacidad de aplicar técnicas estadísticas y algunos modelos de deterioro a partir de la información de alcantarillado registrada actualmente en Bogotá, encontrando que la utilización de estas técnicas puede resultar en la identificación de factores que afectan el proceso de deterioro en las redes locales y una buena estimación del estado en que pueden encontrarse las tuberías dadas sus características físicas a pesar de contar con información limitada (López Kleine et al., 2016).

No obstante, estos estudios también han resaltado la ausencia de variables importantes para la aplicación de otro tipo de metodologías y la limitaciones que pueden presentar los enfoques implementados debido a la disponibilidad de información. Adicionalmente, Ahmadi et al. (2015) y Rokstad & Ugarelli (2016) han resaltado que mejorar los procesos de recolección de información de las redes y evolucionar de una gestión reactiva a una gestión proactiva requiere no solo de la disponibilidad de información y herramientas para implementar estos métodos sino también de la demostración de los beneficios de registrar esta información para incentivar a las organizaciones y establecer la capacidad de las mismas de utilizar esta información.

Más aún, según la revisión bibliográfica realizada, se observa que no ha sido objeto de estudio la cantidad de información que pueden requerir los procesos de calibración de modelos de deterioro para la adecuada estimación generalizada de todos los activos en las redes de la ciudad; siendo esto un factor clave para la aceptación y utilización efectiva de estas técnicas como métodos de gestión en la EAB, puesto que en ciudades con bajas tasas de inspección de sus redes, las preocupaciones respecto a la gestión mediante el uso de técnicas de modelación predictivas tienen fundamento en la cantidad y la calidad de información registrada. Debido a estas limitaciones, los tomadores de

decisiones pueden adoptar actitudes de rechazo o escepticismo respecto a la capacidad de estas técnicas para estimar adecuadamente el comportamiento general de sus redes a partir de un conjunto limitado de datos.

Por consiguiente, para contribuir a esta línea de investigación y cuantificar los efectos de la cantidad de información requerida para la generalización de resultados de modelos de deterioro en redes de alcantarillado en la ciudad de Bogotá, se llevó a cabo un ejercicio académico a partir de un conjunto de datos sintéticos, en el cual se cuantificó la capacidad de estimar la proporción de tuberías en mal estado en una base de datos con información de una zona de la red de alcantarillado sanitario de la ciudad de Bogotá mediante tres modelos de minería de datos con un número limitado de variables predictoras al incrementar el tamaño de la muestra disponible para su calibración. Mediante la aplicación de este ejercicio numérico se busca establecer un marco de referencia a partir del cual se puedan desarrollar futuros estudios en que se establezcan los beneficios y costos de la destinación de recursos a actividades de inspección de redes y modelación predictiva, con el propósito de incentivar el uso efectivo de estas técnicas en el contexto de países en desarrollo.

Así, en los siguientes tres subcapítulos de este documento se describe la metodología utilizada y los resultados obtenidos para el establecimiento de las necesidades de la cantidad de información para el uso de tres modelos de minería de datos en un conjunto de datos sintético generado a partir del comportamiento estimado de las tuberías en la red de alcantarillado de Bogotá.

7.1.1 Metodología para la generación del caso de estudio sintético

A partir de las características físicas y/o topológicas del alcantarillado sanitario de la zona 1 del alcantarillado de la ciudad de Bogotá se generó una base de datos sintéticos, en la cual se asignó a cada tubería (según sus características) una condición estructural (0 o 1). Lo anterior se llevó a cabo con el propósito de poder estimar la precisión de diferentes modelos de minería de datos de clasificar correctamente las tuberías que se encuentran en mal estado en toda la red de la zona 1, al calibrar los modelos con una muestra de datos de tamaño incremental. La información de las características físicas y/o topológicas se obtuvo a partir de las capas de ArcGIS en las cuales la EAB almacena y gestiona la información de sus redes para el año 2018.

El proceso para la generación de la condición estructural de los datos sintéticos se describe en el diagrama de flujo que se presenta en la Figura 7-1. En este, se observan los 4 pasos principales realizados para la asignación estructural de cada tubería de acuerdo a sus características. A continuación se realiza una breve descripción de cada uno de estos pasos, los cuales se implementaron utilizando la herramienta de programación R en conjunto con Excel y VBA.

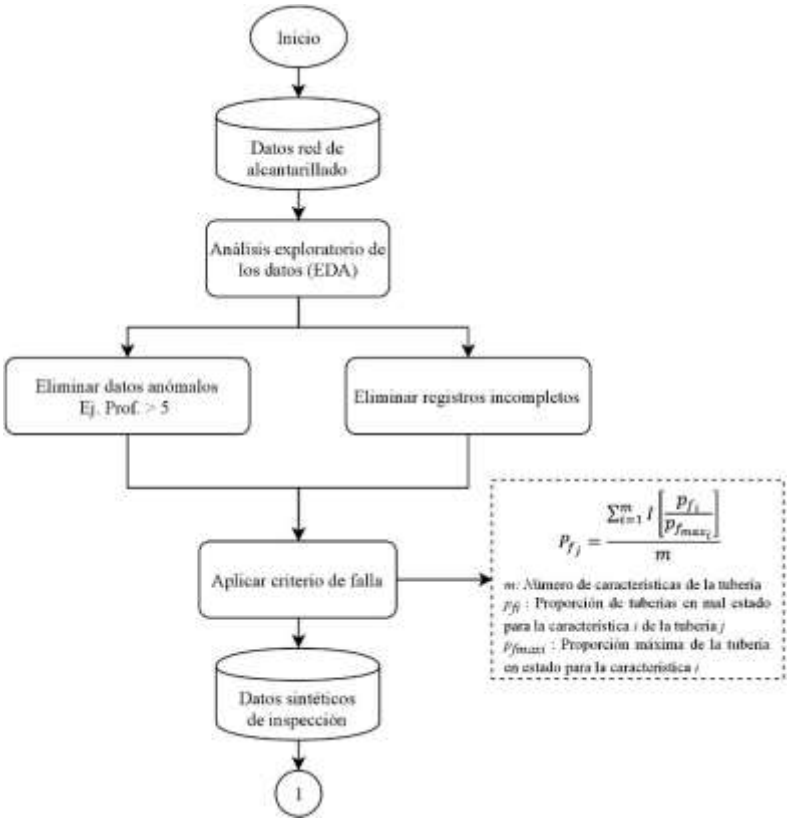


Figura 7-1. Metodología para la generación de datos sintéticos.

- Paso 1. Datos red de alcantarillado

Este paso corresponde a cargar los datos al software R para su posterior procesamiento. Para esto se requiere crear un archivo .xlsx en el cual únicamente se encuentren los nombres de las columnas y los registros de las diferentes variables para cada tubería.

- Paso 2. Análisis exploratorio de los datos

Este paso corresponde a la identificación de las variables continuas y las variables categóricas presentes en la base de datos. Así mismo, en este paso se determinan las distribuciones que presentan las variables continuas, y los valores que pueden tomar las variables categóricas. Así, es posible identificar si existen valores anómalos de las variables o registros incompletos de las variables de las tuberías.

- Paso 3. Pre-procesamiento de los datos

En este paso se realiza la eliminación de los datos anómalos o datos que no se consideren de interés y se pueden establecer estrategias para el manejo de datos faltantes. En este caso, al

contar con un registro de datos considerablemente grande, se decidió eliminar todos los datos anómalos, al igual que el registro de tuberías con características incompletas. En total se eliminaron 2524 registros (1600 corresponden a registros incompletos) de un total inicial de 27947 datos. Finalmente, el conjunto de datos total utilizado tiene 25423 registros.

La Figura 7-2 y la Figura 7-3 presentan la distribución de las variables en el conjunto de datos después de haber realizado su pre-procesamiento.

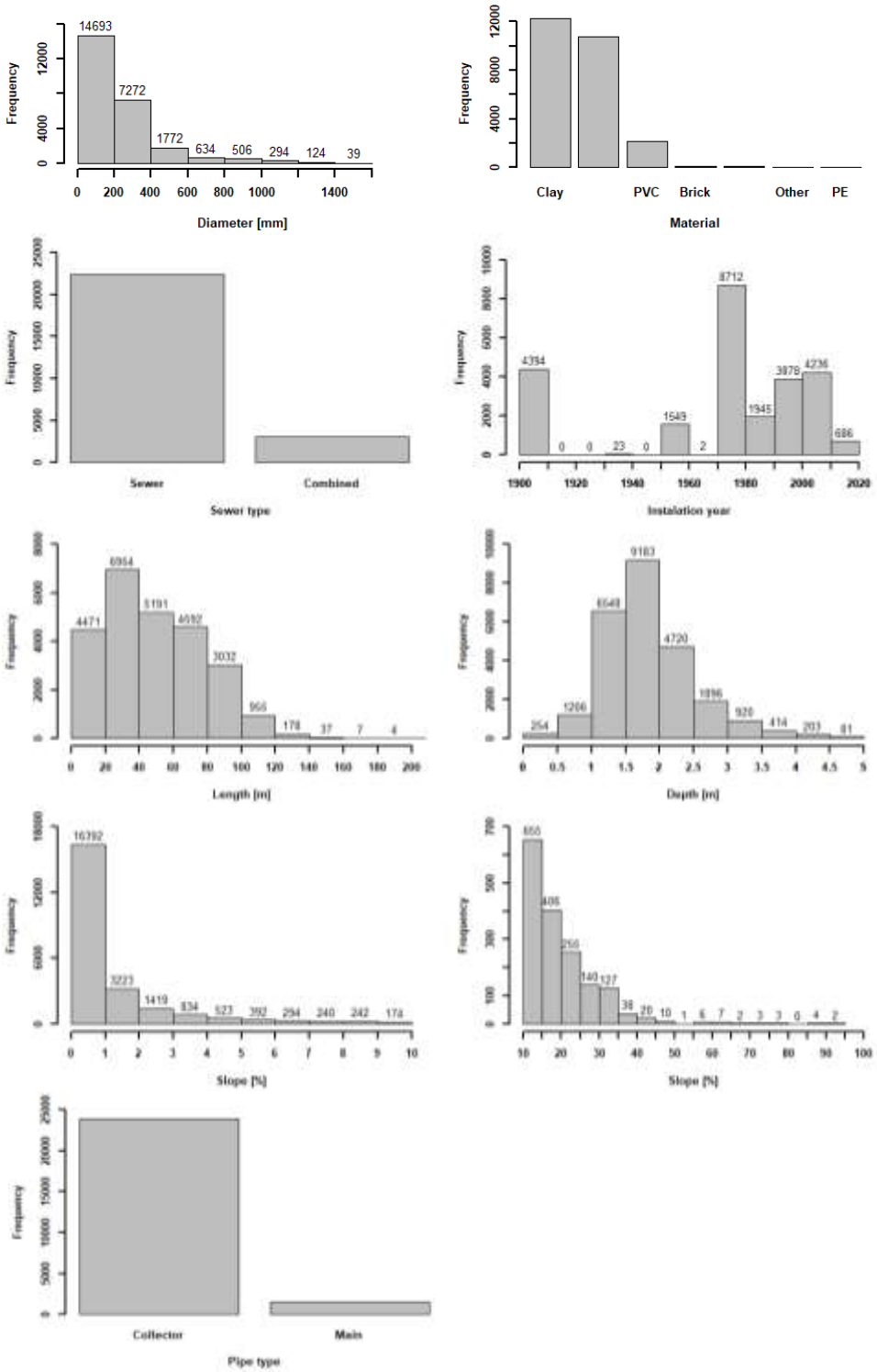


Figura 7-2. Histogramas de frecuencia para la distribución de valores de las variables del caso de estudio..

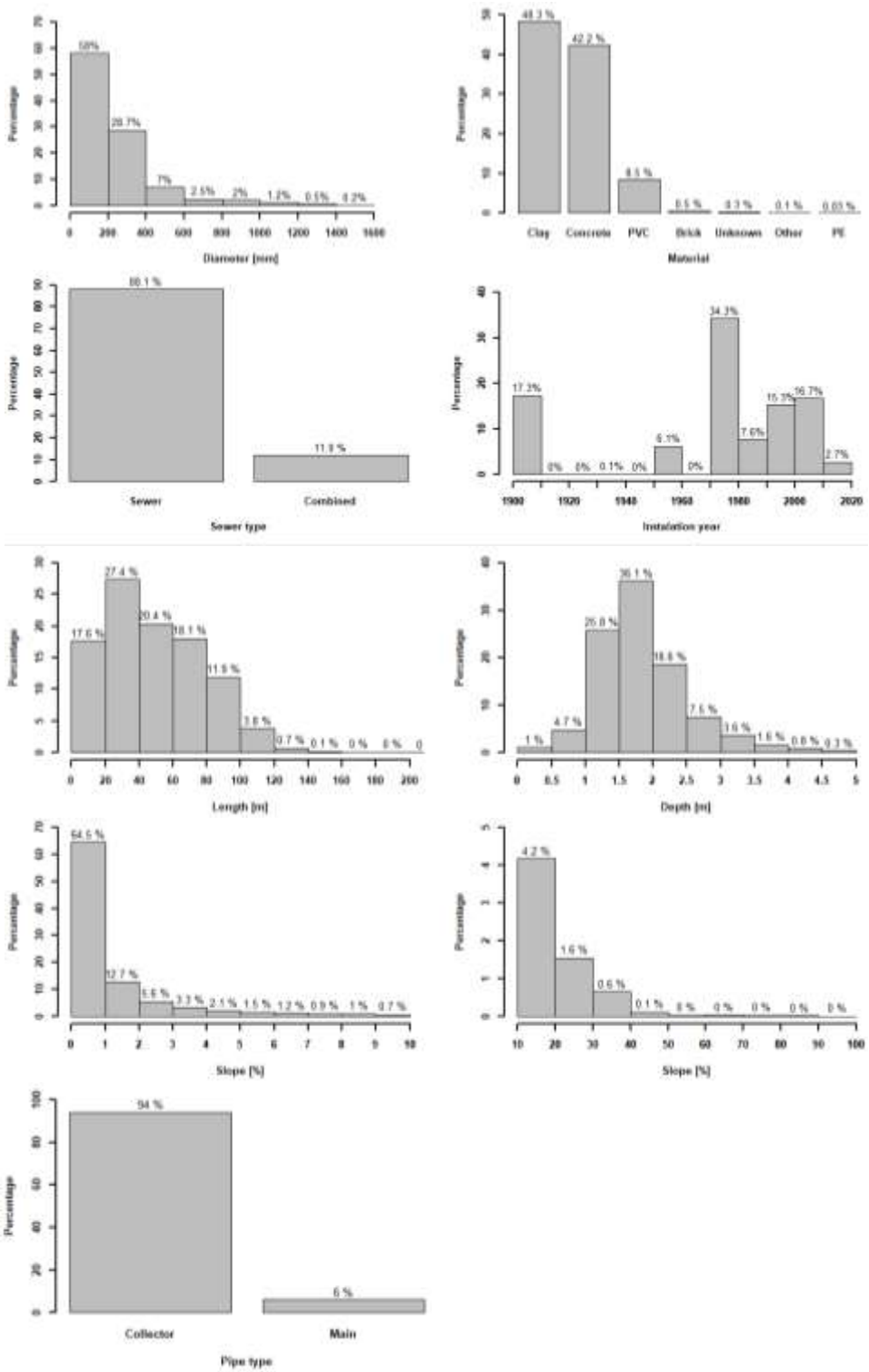


Figura 7-3. Histogramas de densidad para la distribución de valores de las variables del caso de estudio.

- Paso 4. Aplicar criterio de falla

Este paso consiste en la determinación del estado estructural de las tuberías, mediante una clasificación binaria. Para esto, se estableció un criterio de falla con base en las distribuciones de clases estructurales encontradas en el estudio realizado por Caradot et al., 2018 sobre un registro de 5.076 tuberías inspeccionadas en la ciudad de Bogotá.

Las distribuciones del conjunto de datos utilizado por estos autores se presenta en la Figura 7-4, en la cual se observa la clasificación estructural de las tuberías agrupada por cada característica física. Para los efectos de su investigación, los puntajes originales 3 y 4 de las tuberías se agruparon para tener solo 4 clases.

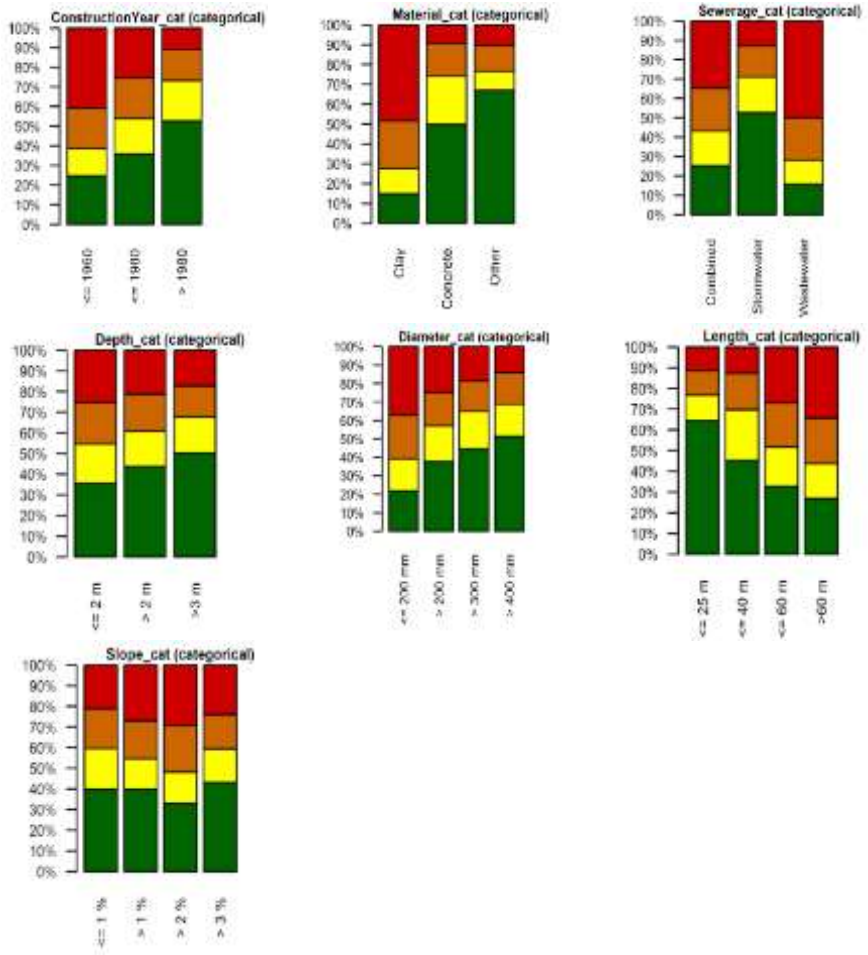


Figura 7-4. Distribución de las condiciones estructurales del conjunto de tuberías inspeccionadas. Las condiciones varían de 1 (verde) a 4 (rojo). Tomado de (Caradot, Hernandez, et al., 2018)

Con el propósito de establecer únicamente una clasificación binaria, se consideró que las clases 3 y 4 de las distribuciones anteriores corresponden a una tubería que se encuentra en mal estado (1) y

las clases 1 y 2 corresponden a una tubería que se encuentra en buen estado (0). Así, el criterio de falla definido para una tubería j se presenta en las siguientes ecuaciones:

$$Pf_j = \frac{\sum_{i=1}^{m} I\left(\frac{p_{fij}}{p_{f_{maxi}}}\right)}{m} \qquad \text{Ecuación 7-1}$$

$$I(x) = \begin{cases} 1 \; si \; x = 1 \\ 0 \; si \; x \neq 1 \end{cases}$$

$$Condición\ estructural\ (SC) = \begin{cases} 1 \; si \quad P_{f_j} \geq 0.7 \\ 0 \; si \quad P_{f_j} < 0.7 \end{cases} \qquad \text{Ecuación 7-2}$$

donde:

- $m \coloneqq Número\ de\ caracteristicas\ consideradas\ para\ el\ criterio\ de\ falla = 7$
- $p_{f_{ij}} \coloneqq Proporción\ de\ tuberías\ en\ mal\ estado\ para\ la\ caracteristica\ i\ de\ la\ tubería\ j$
- $p_{f_{max_i}} \coloneqq Proporción\ máxima\ de\ la\ de\ la\ tubería\ en\ mal\ estado\ para\ la\ caracteristica\ i$
- $P_{f_j} \coloneqq Probabilidad\ de\ falla\ para\ la\ tubería\ j$

El criterio de falla descrito en las ecuaciones anteriores intenta simular el comportamiento aproximado de las tuberías que pueden tener simultáneamente las características (diámetro, pendiente, material, longitud, profundidad media y año de instalación) que tienen generalmente una mayor proporción de tuberías en mal estado, y por lo tanto podrían interpretarse como las tuberías más propensas a encontrarse en un mal estado estructural. Un criterio de $P_{f_i} \geq 0.7$ corresponde a las tuberías que tienen simultáneamente 5 o más características con la proporción de tuberías en mal estado máxima, del total de 7 características utilizadas.

Es decir, para una tubería j con las características presentadas en la Tabla 7-1, las cuales corresponden a los valores que presentan la mayor proporción de tuberías en mal estado estructural de acuerdo con la Figura 7-4, se obtendría la siguiente clasificación:

Tabla 7-1. Características de una tubería j

Año de instalación	Material	Tipo de sistema	Profundidad	Diámetro	Longitud	Pendiente
1950	Gres	Sanitario	1.5 m	150 mm	70 m	2.5 %

$$Pf_j = \frac{I\left(\frac{0.6}{0.6}\right)_{año} + I\left(\frac{0.7}{0.7}\right)_{mater} + I\left(\frac{0.45}{0.45}\right)_{prof} + I\left(\frac{0.6}{0.6}\right)_{diam} + I\left(\frac{0.55}{0.55}\right)_{long} + I\left(\frac{0.5}{0.5}\right)_{pend}}{7}$$

Ecuación 7-3

$$P_{f_j} = \frac{1+1+1+1+1+1+1}{7} = 1$$

$$Condición\ estructural\ (SC) = 1$$

Ecuación 7-4

Es importante resaltar que el criterio anterior se define con el propósito de generar la condición estructural del conjunto de tuberías de toda la red de alcantarillado correspondiente a la zona 1 de la ciudad de Bogotá, y únicamente considera la distribución de tuberías en los diferentes estados estructurales reportados en el estudio de Caradot, Hernandez et al. (2018). Así, se obtiene un conjunto de datos de inspección sintéticos para toda la red de alcantarillado de la zona 1 con los cuales es posible comparar la predicción realizada por los diferentes modelos de minería de datos, lo cual no sería posible con datos reales debido a que se requerirían registros de inspección de todas las tuberías de esta zona.

7.1.2 Caso de estudio sintético

Los resultados obtenidos con el criterio de falla descrito anteriormente se pueden observar para cada una de las características en la Figura 7-5. Al aplicar este criterio se obtiene una proporción de tuberías en mal estado correspondiente a $p = 0.145$, lo cual se puede considerar un valor razonable considerando que representa la agrupación de las clases estructurales 3, 4 y 5, y además, que en las redes de alcantarillado generalmente la cantidad de tuberías que se encuentra en buen estado es significativamente mayor a la cantidad de tuberías en buen estado (Ahmadi et al., 2015; Bailey et al., 2015; Caradot, Riechel, et al., 2018; Carvalho, 2015; Harvey & McBean, 2014; Harvey et al., 2015; Laakso, Kokkonen, et al., 2018; Mashford et al., 2011; Rokstad & Ugarelli, 2015).

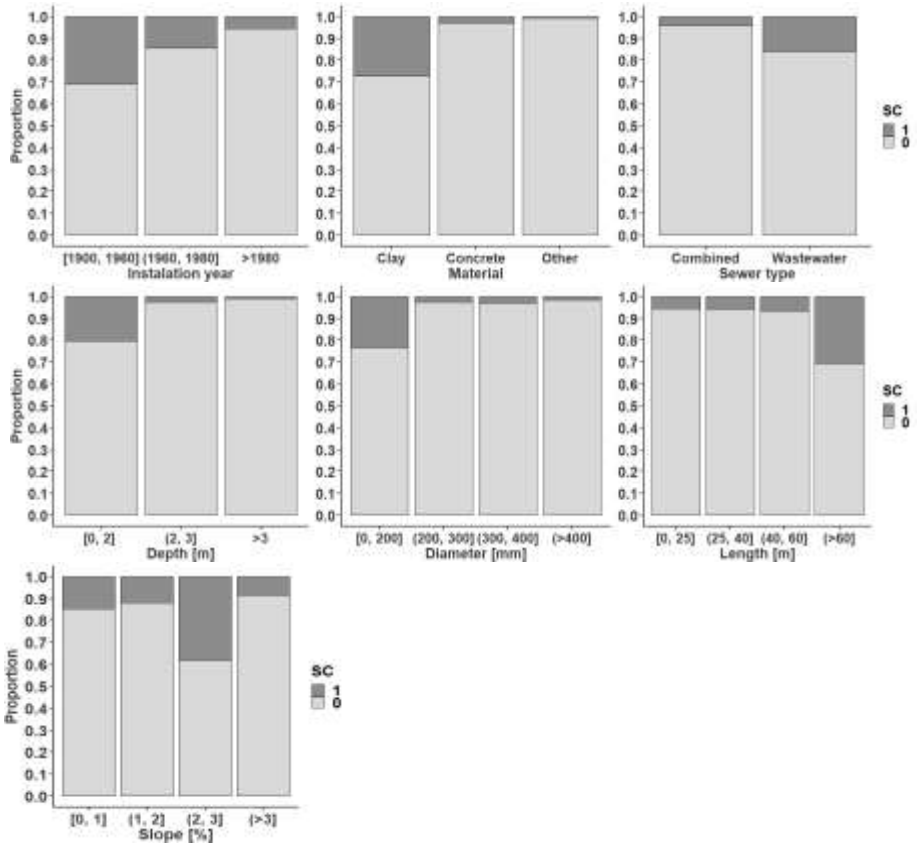


Figura 7-5. Distribución de las condiciones estructurales de la red de alcantarillado sanitario de la zona 1 de Bogotá. Datos sintéticos a partir de la aplicación del criterio de falla.

Además, a partir de las distribuciones de clases resultantes para cada una de las variables, es posible identificar que la tendencia de las proporciones de tuberías en mal estado es igual o muy similar a la presentada por Caradot, Hernandez, et al. (2018) para cada una de las características. Por ejemplo, se observa que la proporción de tuberías en mal estado según la característica "Año de instalación" presenta el mismo comportamiento decreciente, lo cual también es consistente con la influencia esperada de este factor en el deterioro de las tuberías. Al analizar en detalle las distribuciones de tuberías en mal estado agrupadas para las demás características se observa que también siguen las tendencias reportadas por los autores de referencia, por lo cual se podría considerar que la base de datos sintética corresponde a una aproximación adecuada para la aplicación del ejercicio numérico.

7.1.3 Modelos de minería de datos a un caso de estudio sintético

Previo al estudio del efecto del tamaño de la muestra de calibración en la capacidad de predicción de los modelos de deterioro, se realizó la calibración y estimación de parámetros relevantes para cada uno de los modelos considerando un tamaño de muestra apropiado, que se determinó a partir de la Ecuación 5-39 y Ecuación 5-40 teniendo en cuenta que la selección de la muestra para la

calibración se lleva a cabo mediante muestreo simple aleatorio (Ahmadi et al., 2016; Lohr, 2010). De acuerdo a estas ecuaciones, el tamaño de la muestra mínima (n) se selecciona con el propósito de que la calibración de los modelos con este conjunto de datos limitado permita estimar la proporción de tuberías en mal estado en todo el conjunto de datos (N) con un margen de error determinado. El valor aceptable para este margen de error es relativo a la proporción real de tuberías en mal estado en todo el conjunto de estado estimada preliminarmente (conocimiento del sistema) o conocerse (en el caso de haber inspeccionado toda la red) (Ahmadi et al., 2016).

Debido a que se cuenta con un conjunto de datos para el cual se conoce la condición estructural de todas las tuberías, es posible conocer el valor real de esta proporción, igual a $p = 0.145$. La Tabla 7-2 presenta los diferentes valores obtenidos para el tamaño de muestra mínimo (n), considerando la variación del margen de error (e).

Tabla 7-2. Tamaño mínimo de la muestra para diferentes valores del margen de error.

Tamaño total del conjunto de datos	Proporción real de tuberías en mal estado	Margen de error	V(p)	Tamaño mínimo de la muestra
N	p	e	-	n
25423	0.1453	0.01	$2.603 \cdot 10^{-5}$	4017
25423	0.1453	0.02	$1.041 \cdot 10^{-4}$	1139
25423	0.1453	0.03	$2.343 \cdot 10^{-4}$	519
25423	0.1453	0.04	$4.165 \cdot 10^{-4}$	294
25423	0.1453	0.05	$6.508 \cdot 10^{-4}$	189

Entonces, teniendo en cuenta el valor de p para todo el conjunto de datos, podría considerarse que una estimación $\hat{p} \pm 3\%$ sería aceptable; siendo además el valor más comúnmente utilizado (Lohr, 2010). Sin embargo, con el fin de evidenciar las diferencias que se presentan en dos modelos con una menor y mayor cantidad de datos, se realizaron las calibraciones de los modelos para $e = 3\%$ $(n = 519)$ y $e = 1\%$ $(n = 4017)$.

La Figura 7-6 presenta el diagrama de flujo de la metodología utilizada para la calibración de los diferentes modelos y la estimación de la proporción de tuberías en mal estado para toda la red.

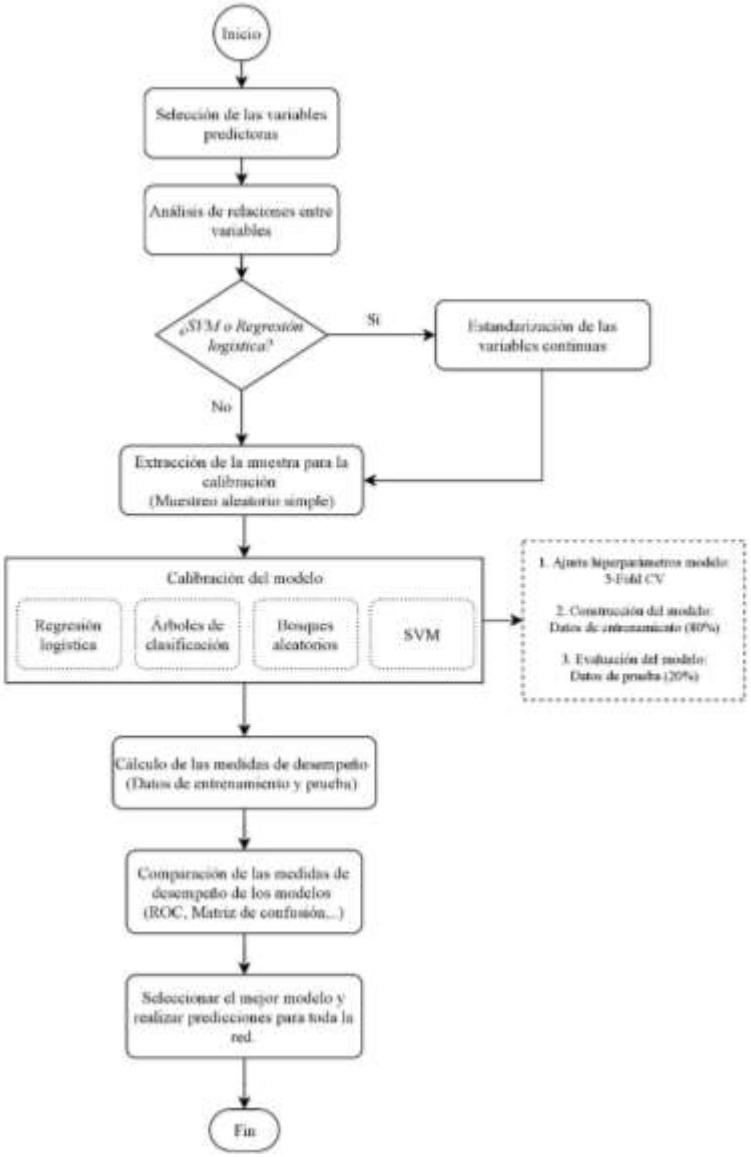


Figura 7-6. Calibración de los modelos de minería de datos y estimación de p para la red.

- El primer paso corresponde a la selección de las variables de entrada que se utilizaran para la predicción del estado estructural de las tuberías. Considerando que las tasas de inspección de la ciudad de Bogotá son bajas y no existe un registro extensivo de las características y/o el entorno, se decidió utilizar las variables más comúnmente registradas y que a su vez, son generalmente encontradas como explicativas en la modelación predictiva (ver Figura 5-8). Así, las variables predictoras son el material, la edad/año de instalación, la longitud, el diámetro, la pendiente y la profundidad de las tuberías (ver Tabla 7-3).

Tabla 7-3. Variables predictoras consideradas para la modelación.

No.	Variable	Tipo
1	Material	Categórica
2	Edad de la tubería/Año de instalación	Numérica
3	Longitud	Numérica
4	Tamaño/ Diámetro	Numérica
5	Pendiente	Numérica
6	Profundidad	Numérica

- El segundo paso corresponde al estudio de las relaciones entre las variables predictoras. Para esto se determinan los valores de correlación entre las variables numéricas y se analiza gráficamente las relaciones entre todas las variables (para incluir la variable categórica material). La Figura 7-7 y la Figura 7-8 presentan las relaciones existentes entre las variables; a partir de estas se observa que, en general, las variables se encuentran poco relacionadas entre sí, dándose valores de correlación inferiores a 0.2 con excepción de la longitud y la pendiente de las tuberías las cuales tienen una correlación negativa de -0.25.

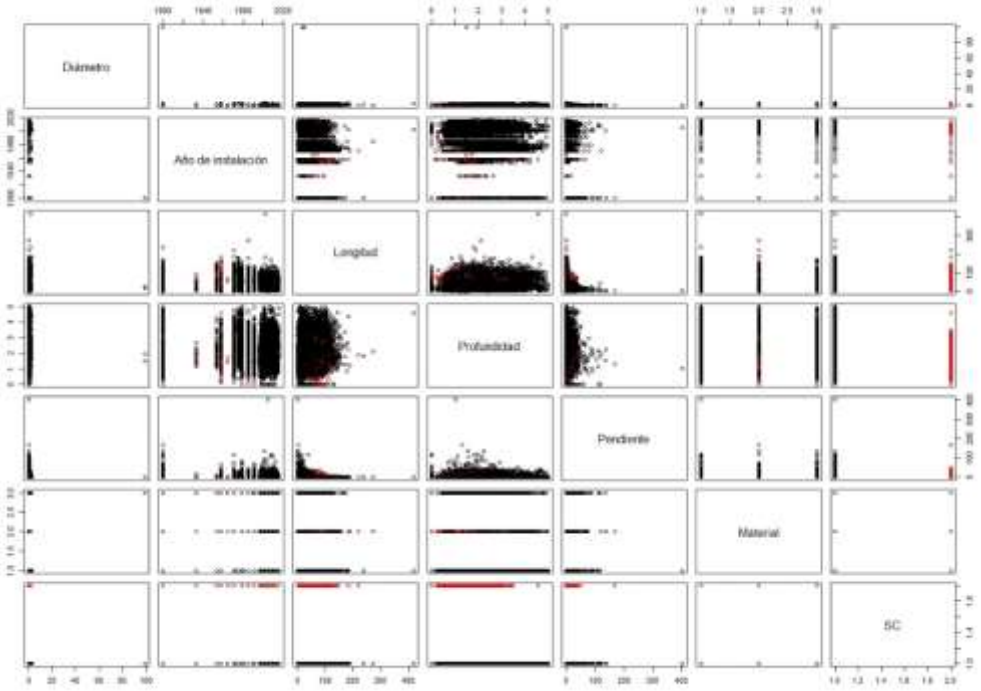


Figura 7-7. Matriz de dispersión de las variables predictoras. El color asignado (rojo o negro) corresponde al valor de la condición estructural.

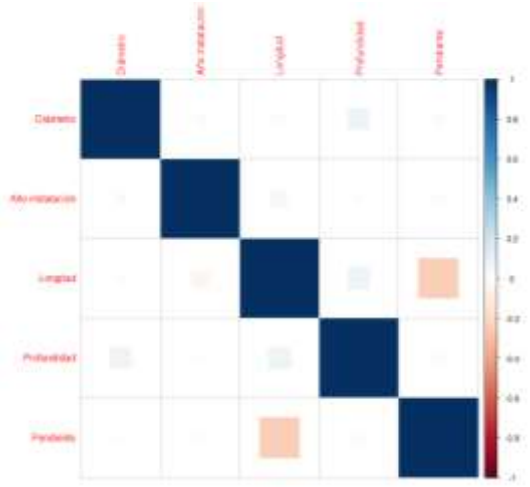


Figura 7-8. Correlaciones entre las variables predictoras.

Lo anterior es de gran importancia pues al implementar algunos modelos se requiere primero considerar si las variables predictoras contienen información redundante y por lo tanto, no aportan información al modelo pero si incrementan su complejidad.

- En los siguientes pasos se realiza la calibración de los modelos de minería considerados. Debido a su capacidad de interpretación se seleccionaron regresión logística y arboles de decisión; mientras que para tener en consideración modelos con una mayor complejidad pero menos interpretabilidad, se seleccionaron los modelos bosques aleatorios y máquinas de soporte vectorial (SVM). En el caso de regresión logística y SVM los datos requieren un preprocesamiento, en el cual se estandaricen los valores de cada característica respecto a su valor medio y desviación, con el propósito de evitar problemas de convergencia debido a la gran diferencia que se puede presentar entre los valores originales de las diferentes características. Adicionalmente, en estos dos modelos se implementó la selección de un número reducido de variables predictoras que garantiza el buen desempeño de los modelos y disminuye la variabilidad de los mismos (debido a la inclusión de más variables que no aportan más información). Este proceso no se realizó en el caso de árboles de decisión y bosques aleatorios debido a que estos algoritmos son poco sensibles a las características de las variables predictoras (no requieren preprocesamiento) y además, la selección de variables relevantes se realiza al limitar la complejidad de los árboles (Harvey & McBean, 2014).

  Los tres principales pasos para la calibración de los modelos consisten en:

  1. Seleccionar las muestras de entrenamiento, prueba y validación (train-test-validate). Valores usuales para la división de entramiento y prueba corresponden a 80% y 20%. A partir del 80% de los datos de entrenamiento se realiza el ajuste de los hiperparámetros de los

modelos mediante validación cruzada ($\mathrm{k-fold\ CV}$), al igual que se estima la capacidad de predicción del modelo de forma más general al considerar el entrenamiento y prueba de los modelos con diferentes conjuntos de datos.

2. Ajustar los hiperparámetros de los modelos (5-Fold Cross-Validation).
3. Construcción del mejor modelo a partir de todos los datos de entrenamiento ($\mathrm{train} = 80\%$) utilizando los hiperparámetros encontrados anteriormente.
4. Evaluar el desempeño de los modelos en los datos de prueba ($\mathrm{test} = 20\%$).

A continuación se presentan los resultados obtenidos para dos conjuntos de datos de diferente tamaño n

7.1.3.1 Regresión logística (LR)

Tabla 7-4. Matriz de confusión Regresión logística. Datos de prueba. Superior (n=4017), Inferior (n=519)

n =	4017			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	72	51
	Buen estado	**0**	16	664
TPR =	$72/(72+51)$ =	58.5%		
TNR =	$664/(16+664)$ =	97.6%		
FPR =	$16/(16+664)$ =	2.4%		
FNR =	$51/(72+51)$ =	41.5%		
Acc =	$(664+72)/(803)$ =	91.7%		

n =	519			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	4	9
	Buen estado	**0**	1	90
TPR =	$4/(4+9)$ =	30.7%		
TNR =	$90/(1+90)$ =	98.9%		
FPR =	$1/(1+90)$ =	1.1%		
FNR =	$9/(4+9)$ =	69.3%		
Acc =	$(90+4)/(104)$ =	87.9%		

7.1.3.2 Árboles de decisión (DT)

Tabla 7-5. Matriz de confusión Árboles de decisión. Datos de prueba. Superior (n=4017), Inferior (n=519)

n =	4017			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	91	32
	Buen estado	**0**	12	668
TPR =	$91/(91+32)$ =	74%		
TNR =	$668/(12+668)$ =	98.2%		
FPR =	$32/(91+32)$ =	26%		
FNR =	$12/(12+668)$ =	1.8%		
Acc =	$(91+668)/(803)$ =	94.5%		

n =	519			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	10	3
	Buen estado	**0**	0	91
TPR =	$10/(10+3)$ =	76.9%		
TNR =	$91/(91+0)$ =	100%		
FPR =	$0/(91+0)$ =	0%		
FNR =	$3/(10+3)$ =	23.1%		
Acc =	$(10+91)/(104)$ =	94.4%		

7.1.3.3 Bosques aleatorios (RF)

Tabla 7-6. Matriz de confusión Bosques aleatorios. Datos de prueba. Superior (n=4017), Inferior (n=519)

n =	4017			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	106	17
	Buen estado	**0**	20	660
TPR =	$106/(106+17)$ =	86.2%		
TNR =	$660/(20+660)$ =	97.1%		
FPR =	$20/(20+660)$ =	2.9%		
FNR =	$17/(106+17)$ =	13.8%		
Acc =	$(106+660)/(803)$ =	95.4%		

n =	519			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	8	5
	Buen estado	**0**	0	91
TPR =	$8/(8+5)$ =	61.5%		
TNR =	$91/(91+0)$ =	100%		
FPR =	$0/(91+0)$ =	0%		
FNR =	$5/(8+5)$ =	38.5%		
Acc =	$(8+91)/(104)$ =	95.2%		

7.1.3.4 Máquinas de soporte vectorial (SVM)

Tabla 7-7 Matriz de confusión SVM. Datos de prueba. Superior (n=4017), Inferior (n=519)

n =	4017			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	104	19
	Buen estado	**0**	34	646
TPR =	$104/(104+19)$ =	84.6%		
TNR =	$646/(34+646)$ =	95%		
FPR =	$34/(34+646)$ =	5%		
FNR =	$19/(104+19)$ =	15.4%		
Acc =	$(104+646)/(803)$ =	93.4%		

n =	519			
Seed =	12		**Predicción**	
			Mal estado	**Buen estado**
			1	**0**
Real	**Mal estado**	**1**	9	4
	Buen estado	**0**	4	87
TPR =	$9/(9+4)$ =	69.2%		
TNR =	$87/(4+87)$ =	95.6%		
FPR =	$4/(4+87)$ =	4.4%		
FNR =	$4/(9+4)$ =	30.8%		
Acc =	$(9+87)/(104)$ =	92.3%		

7.1.3.5 Análisis global de los resultados

En la Tabla 7-8 se presentan las medidas de desempeño calculadas para los diferentes modelos, agrupadas por el tamaño de la muestra con el cual se realizó la calibración.

Tabla 7-8. Resumen medidas de desempeño para los modelos de minería de datos.

	Modelos de Minería de datos							
	n = 4017				**n = 519**			
Medidas de	$\hat{p} = p \pm 1\% = [13.5\% - 15.5\%]$				$\hat{p} = p \pm 3\% = [11.5\% - 17.5\%]$			
desempeño	***LR***	**DT**	**RF**	**SVM**	**LR**	**DT**	**RF**	**SVM**
TPR =	58.5%	74%	86.2%	84.6%	30.7%	76.9%	61.5%	69.2%
TNR =	97.6%	98.2%	97.1%	95%	98.9%	100%	100%	95.6%
FPR =	2.4%	26%	2.9%	5%	1.1%	0%	0%	4.4%
FNR =	41.5%	1.8%	13.8%	15.4%	69.3%	23.1%	38.5%	30.8%
Acc =	91.7%	94.5%	95.4%	93.4%	87.9%	94.4%	95.2%	92.3%
$\hat{p} =$	10.51%	12.7%	14.69%	14.6%	7.89%	11.68%	10.36%	11.74%

Debido a la naturaleza del problema abordado, se considera que las medidas de desempeño más relevantes corresponden a

- TPR (True positive rate): corresponde al número de fallas identificadas correctamente del total de fallas reales.
- FNR (False negative rate): corresponde al número de fallas no identificadas correctamente el total de fallas reales.
- $\hat{p}$: Proporción estimada para todo el conjunto de tuberías en mal estado para todo el conjunto de datos.

Así, de acuerdo a los resultados anteriores no solo es posible identificar un incremento significativo en los porcentajes de tuberías en mal estado estimados correctamente por los modelos (TPR), lo cual implica directamente la disminución de las tasas de error de cada uno de los modelos (FNR); sino también, se observa una estimación mucho más cercana a la real de la proporción de tuberías en mal estado para todo el conjunto de datos. No obstante, al analizar los intervalos esperados para esta variable, se observa que en ambos casos los modelos realizan una estimación cercana a los límites establecidos, y no se presentan valores anómalos que subestimen o sobreestimen de manera significativa la proporción real de tuberías en estado de falla. Lo anterior es de gran importancia, pues permite identificar el verdadero alcance para la generalización de los patrones encontrados en los conjuntos de datos utilizados para la calibración con un nivel de confiabilidad alto.

Así mismo, permite relacionar las medidas de desempeño obtenidas en el proceso de calibración con las medidas de desempeño que podrían observarse al aplicar los modelos a todo el conjunto de

datos. La tendencia anterior se identifica al observar los valores de TPR y $\hat{p}$ para los diferentes modelos, en los cuales se corresponden los valores más altos de TPR con una estimación más acertada de $\hat{p}$, tanto para $n = 519$ como para $n = 4017$.

Por otro lado, respecto a la comparación entre modelos se observa que la capacidad predictora de estos es dependiente de la cantidad de datos disponibles para la calibración y que es posible encontrar que la utilización de un modelo más complejo no siempre resulte en una predicción más acertada de la variable resultante (condición estructural). Lo anterior se vuelve notorio al observar la clasificación de los modelos a partir de su TPR para una menor cantidad de datos $(n = 519)$, siendo DT el modelo con los mejores resultados, seguido de SVM, RF y finalmente LR con una capacidad de predicción inferior al 50%; no obstante, al analizar el orden de estos modelos según su capacidad predictora en un conjunto de datos más grande $(n = 4017)$ se identifica que el mejor modelo corresponde a RF, seguido de SVM, DT y LR. Esto podría ser un indicativo de las limitaciones de modelos más complejos al ser aplicados en conjuntos de datos con un tamaño inferior al requerido para el adecuado aprendizaje de patrones dada la estructura matemática o estadística de cada modelo; y aún más, permite identificar que el uso de modelos menos complejos como arboles de decisión pueden obtener medidas de desempeño similares en escenarios de menor o mayor cantidad de información, al ser debidamente calibrados y ajustados. Por lo cual, la utilización de estos modelos puede ser de gran utilidad en contextos con pocos registros históricos de datos.

No obstante, el análisis realizado anteriormente es dependiente no solo del tamaño de la muestra de entrenamiento sino de las características que pueda tener esta. Entonces, con el propósito de identificar de manera más generalizada el desempeño de estos modelos para la predicción del comportamiento de todo el conjunto de datos, se realizaron 1000 simulaciones de Montecarlo en las cuales se seleccionó la muestra de datos mediante muestreo aleatorio simple (MAS) para cada tamaño de muestra para los modelos LR, DT y RF. En el siguiente capítulo se presenta la metodología utilizada y los resultados obtenidos para cada modelo.

7.1.4 Efecto de la cantidad de información

Con el objetivo de generalizar el efecto de la cantidad de datos para la calibración de los modelos considerando la variabilidad de las características de los datos que pueden presentarse en cada muestra seleccionada aleatoriamente, se realizaron 1000 simulaciones de Montecarlo para cada tamaño de la muestra (n) con la cual se realiza la calibración de los diferentes modelos. Así, para cada 12 tamaños de muestra diferentes (incrementos de 500) se entrenan 1000 modelos a partir de muestras diferentes y se estima en cada uno de ellos el valor de TPR (recall) y la estimación de la proporción de tuberías en mal estado para todo el conjunto de datos de la red de alcantarillado sanitario de la zona 1 de la ciudad de Bogotá.

La Figura 7-9 muestra el diagrama de flujo de la metodología implementada para la realización de este análisis. Vale la pena resaltar que los datos de entrada de este diagrama corresponden a la base de datos sintética generada a partir de la metodología de la Figura 7-1. Este análisis únicamente se llevó a cabo con los modelos regresión logística, arboles de decisión y bosques aleatorios debido a los costos computacionales de ajustar los hiperparámetros de los modelos SVM; estos requieren una gran de tiempo para su calibración adecuada y por lo tanto no se consideró viable la calibración de 1000 modelos diferentes para cada tamaño de muestra.

Así, en total se realiza el entrenamiento (con los datos de entrenamiento), evaluación (con los datos de prueba) y generalización de los resultados (con todos los datos de la zona 1) para 12.000 modelos diferentes, agrupados según el tamaño de la muestra seleccionada para la calibración del modelo. Los resultados obtenidos para cada uno de los modelos respecto a la proporción estimada de tuberías en estado de falla se presentan en la Tabla 7-9 y la Figura 7-10. Igualmente, la Tabla 7-10 y la Figura 7-11 muestran la probabilidad de detección (TPR) en el conjunto de datos de prueba.

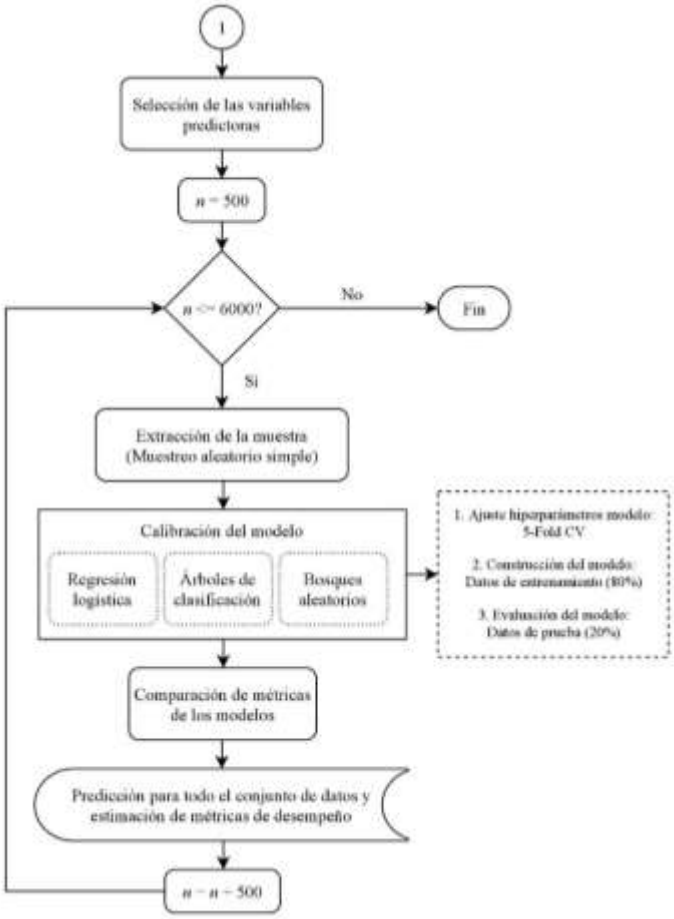


Figura 7-9. Metodología para la estimación de la proporción de tuberías en mal estado para diferentes tamaños de muestra.

Tabla 7-9. Estimación promedio de la proporción de tuberías en mal estado y desviación estándar según n

		Estimación de Proporción de tuberías en mal estado (p)					
		Regresión logística		Árboles de decisión		Bosques aleatorios	
No.	n	Promedio	Desviación estándar	Promedio	Desviación estándar	Promedio	Desviación estándar
1	500	11.76%	1.558%	11.66%	2.991%	12.82%	1.355%
2	1000	11.54%	1.182%	11.36%	2.304%	13.76%	0.809%
3	1500	11.45%	0.929%	11.52%	2.019%	14.19%	0.636%
4	2000	11.42%	0.780%	11.44%	1.938%	14.52%	0.520%
5	2500	11.41%	0.718%	11.58%	1.855%	14.69%	0.436%
6	3000	11.41%	0.664%	11.81%	1.738%	14.80%	0.387%
7	3500	11.36%	0.610%	11.81%	1.739%	14.89%	0.322%
8	4000	11.35%	0.537%	11.90%	1.712%	14.97%	0.293%
9	4500	11.33%	0.521%	12.07%	1.671%	15.02%	0.272%
10	5000	11.32%	0.475%	12.17%	1.612%	15.05%	0.253%
11	5500	11.35%	0.442%	12.24%	1.572%	15.06%	0.234%
12	6000	11.35%	0.448%	12.35%	1.522%	15.11%	0.218%

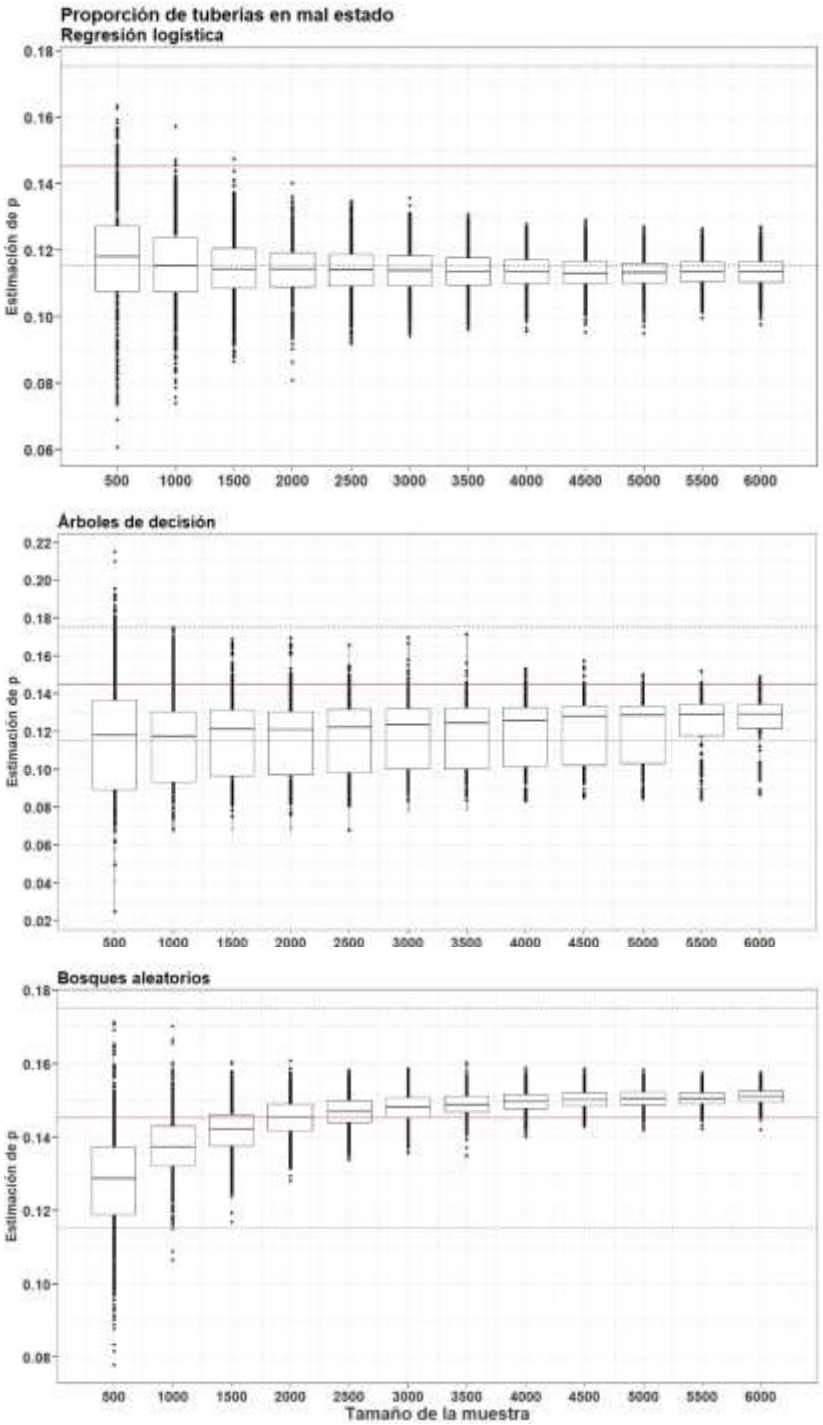


Figura 7-10. Estimación de la proporción de tuberías en mal estado para todo el conjunto de datos – Diferentes modelos de minería de datos.

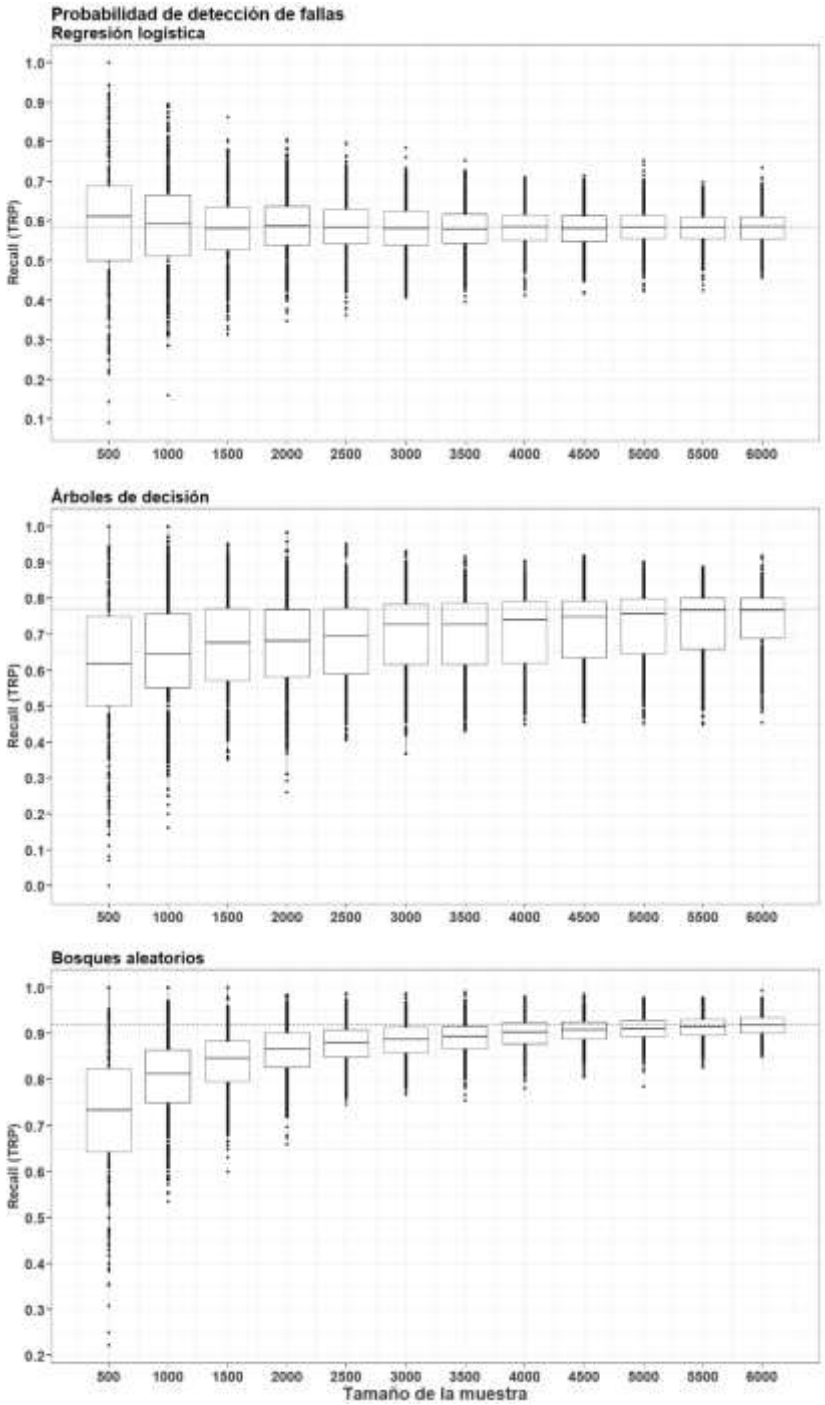


Figura 7-11. Probabilidad de detección de fallas en los datos de prueba – Diferentes modelos de minería de datos.

Tabla 7-10. Probabilidad de detección promedio y desviación estándar según n.

		TPR (Probabilidad de detección de fallas) – Datos de prueba					
		Regresión logística		Árboles de decisión		Bosques aleatorios	
No.	**n**	**Promedio**	**Desviación estándar**	**Promedio**	**Desviación estándar**	**Promedio**	**Desviación estándar**
1	500	59.54%	14.59%	61.25%	17.50%	72.77%	13.41%
2	1000	58.77%	10.88%	64.91%	14.28%	80.69%	8.16%
3	1500	58.03%	8.33%	67.02%	12.47%	83.85%	6.43%
4	2000	58.61%	7.39%	67.23%	12.27%	86.28%	5.31%
5	2500	58.52%	6.72%	68.41%	10.85%	87.83%	4.27%
6	3000	58.20%	6.05%	70.17%	10.65%	88.47%	3.87%
7	3500	57.91%	5.59%	70.34%	10.35%	89.13%	3.57%
8	4000	58.32%	4.97%	71.05%	10.38%	90.01%	3.23%
9	4500	58.04%	4.82%	71.84%	9.83%	90.52%	2.93%
10	5000	58.29%	4.64%	72.66%	9.60%	91.04%	2.73%
11	5500	58.24%	4.08%	73.27%	9.48%	91.34%	2.48%
12	6000	58.40%	4.14%	73.63%	8.98%	91.75%	2.36%

A partir de los resultados obtenidos se identifica que:

- Respecto a la proporción estimada de tuberías en mal estado, se observa que en todos los casos (diferentes modelos) se presenta una disminución significativa de la dispersión de los resultados obtenidos por los modelos a medida en que se incrementa el tamaño de la muestra utilizado para la calibración, lo cual es consistente con lo encontrado por (Ahmadi et al., 2016). No obstante, es posible identificar que los modelos regresión logística y bosques aleatorios presentan resultados para un mismo tamaño de la muestra con menores varianza que los árboles de decisión, al igual que se identifica que los valores mínimos de dispersión alcanzados por los dos primeros modelos son similares $(0.2\% - 0.5\%)$, mientras que la dispersión mínima alcanzada por los modelos arboles de decisión corresponde a 1.522%. Esto se ve reflejado en el mayor rango del eje vertical en que se ubican los registros en la Figura 7-10.
- La precisión para estimar la proporción de tuberías en mal estado en toda la red de alcantarillado de la zona 1 es variable para los tres modelos analizados. En la Figura 7-10 es posible observar las diferencias que se presentan respecto al valor de $\hat{p}$ al cual parecen converger los modelos. En particular, es posible identificar que el modelo de regresión logística no mejora su estimación de p a medida que el tamaño de la muestra incrementa, a pesar de que si disminuye significativamente la dispersión de las estimaciones. Lo contrario ocurre en el caso de árboles de decisión y bosques aleatorios, en los cuales se puede observar como el valor medio de $\hat{p}$ tiende a aproximarse al valor real de todo el conjunto de datos. En el caso de bosques aleatorios este acercamiento al valor real ocurre significativamente más rápido que en el caso de árboles de decisión, pues a partir de un

tamaño de la muestra aproximadamente de 2000 registros, el valor medio estimado por este modelo es muy similar al valor real.

- A pesar de que mediante la aplicación de los tres modelos a todo el conjunto de datos de la red se obtienen valores cercanos a la cantidad real de tuberías que se encuentran en mal estado, se observa que únicamente en el caso de bosques aleatorios las estimaciones realizadas mediante las 1000 simulaciones se encuentran completamente acotadas dentro de los intervalos definidos por un margen de error de 3%. El modelo de regresión logística es el que presenta el peor comportamiento respecto a lo anterior, puesto que, a diferencia del modelo de árboles de decisión, los resultados obtenidos con las simulaciones no presentan una tendencia a acercarse al valor real de p.
- También es posible identificar que la capacidad de predicción de los modelos puede alcanzar valores adecuados al utilizar un conjunto de datos provenientes de un muestreo aleatorio de los registros totales de la red, en los casos en los que los datos presenten un comportamiento consistente, como el generado por el criterio de falla establecido para la generación de los datos sintéticos. En particular, se podría considerar la relevancia de lo anterior en las buenas predicciones que pueden obtenerse cuando los registros de la condición estructural se encuentran agrupados por diferentes mecanismos de falla, como han mencionado diversos autores.
- Lo anterior se identifica a partir de la convergencia de los modelos árboles de decisión y bosques aleatorios, en los cuales la tendencia de los resultados es consistente con lo esperado, pues incrementar los datos utilizados para la calibración de los modelos permite que estos tengan una mayor cantidad de información para identificar y aprender los patrones que se presentan en los datos. En el caso del modelo de regresión logística se obtiene un resultado que no es consistente con lo esperado y podría deberse al algoritmo del mismo, el cual no se selecciona las variables realmente importantes para la predicción por si mismo, a diferencia de los modelos basados en estructuras de árboles. En estos últimos, limitar la complejidad de los árboles permite eliminar niveles que pueden contener variables que aportan poco en el proceso de decisión.
- El comportamiento de la probabildad de detección de fallas (TPR) para los diferentes modelos es muy similar al observado en la estimación de la proporción de tuberías en mal estado en toda la red, con lo cual se podría interpretar que un mejor desempeño en los datos de prueba resulta en una incremento en la capacidad de generalizar los patrones encontrados en todo el conjunto de datos (en el caso de los modelos basados en estructuras de árboles). Vale la pena recordar que debido a que TPR representa el porcentaje de tuberías que son correctamente identificadas en mal estado, $1 - \mathrm{TPR}$ corresponde a la tasa de falla de estos modelos, con lo cual valores de TPR de 60% implican la no identificación de tuberías en mal estado en un 40% de los casos.

8 ANÁLISIS DE LA VIABILIDAD DE MODELOS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE FALLAS EN REDES DE ALCANTARILLADO.

8.1 Ventajas y retos del uso de modelos de Minería de Datos

Teniendo en cuenta los anteriores modelos presentados y las investigaciones realizadas en los últimos años, en las cuales se ha evaluado la aplicabilidad de estos en una gran cantidad de conjuntos de datos diferentes, se pueden identificar una serie de beneficios y retos que surgen al llevar a cabo el proceso de explotación de patrones a partir de bases de datos reales y la capacidad de ser implementados y utilizados en la gestión real de activos en redes de alcantarillado.

En primer lugar es importante recordar que una de las razones por las cuales se ha considerado la utilidad de algoritmos de minería de datos a diferentes problemas y en particular a problemas de predicción de fallas en redes de alcantarillado es debido a la oportunidad que surge al contar con métodos de inspección y monitoreo cada vez más desarrollados que permiten almacenar grandes conjuntos de datos, de los cuales es posible explotar más información no identificable mediante métodos más sencillos. Así mismo, estos métodos permiten superar dificultades a las que se enfrentan otros métodos en los cuales se requiere un entendimiento completo de un proceso complejo (métodos determinísticos) u otros, como los métodos estadísticos, en los cuales las soluciones se encuentran limitadas por la creatividad del modelador para plantear hipótesis. Entonces, el uso directo de los datos históricos de un fenómeno, permite desarrollar modelos que están basados en el comportamiento esperado del sistema pero que no requieren la definición de modelos con estructuras matemáticas o de probabilidad muy complejas.

Por otro lado, el uso adecuado de estos métodos también permite afrontar los retos a los que se enfrentan las empresas prestadoras del servicio para realizar una gestión eficiente de sus activos teniendo en cuenta: (1) la complejidad de los sistemas que deben ser administrados y, (2) las limitaciones de recursos (financieros y humanos) que se tienen para conocer el estado de los sistemas de drenaje y garantizar la prestación de un servicio de calidad.

En relación al primer punto anterior, es importante considerar la complejidad del problema al cual se enfrentan las empresas prestadoras del servicio y la cantidad limitada de recursos (financieros y humanos) que se disponen para estos procesos. En muchos casos el proceso de destinación de recursos tiene en consideración muchos otros factores adicionales a la exclusiva rehabilitación del sistema, como la inversión de capital en mejorar infraestructura visible (New England Interstate Water Pollution Control Commission, 2003), la rehabilitación simultanea de otro tipo de infraestructura como vías y carreteras (Van Riel, Langeveld, Herder, & Clemens, 2014a), e incluso la opinión y el instinto de los administradores del sistema encargados de estas actividades en las empresas prestadoras del servicio (Van Riel et al., 2014b). Lo anterior implica que el proceso de

toma de decisiones para la rehabilitación de sistemas de alcantarillado se encuentra inmerso en un contexto sociotécnico complejo, lo cual garantiza que se trate de un problema intrínsecamente complicado.

En cuanto al segundo punto, se puede resaltar que para llevar a cabo un mantenimiento eficaz en términos de costos y desempeño del sistema, se requiere tener conocimiento de cuáles son los componentes del sistema, su ubicación y estado a lo largo de la vida útil del proyecto (New England Interstate Water Pollution Control Commission, 2003). Por lo cual, el registro y manejo de la información de estos componentes resulta de gran importancia para llevar a cabo la gestión de activos de las redes de alcantarillado, cualquiera que sea el enfoque implementado por las empresas prestadoras del servicio. De manera que, la información que se registra en la inspección de las redes de alcantarillado permite obtener un material del cual es posible explotar más información que permita conocer el estado de las redes y predecir el comportamiento de los sistemas, al igual que garantizar la inversión de recursos de manera más eficiente.

Sin embargo, la aplicación de estas técnicas en conjuntos de datos reales obtenidos mediante agencias prestadoras del servicio puede verse limitada por una serie de factores que influyen tanto en la capacidad de predicción de estos modelos como en la actitud o percepción que tienen los operadores de estos sistemas para confiar en los resultados encontrados. A continuación se presentan algunas de las principales limitaciones que se presentan para la inclusión y uso apropiado de técnicas de minería de datos para el modelamiento predictivo de fallas en redes de alcantarillado.

8.1.1 Intuición en el proceso de toma de decisiones

El proceso de toma de decisiones para la rehabilitación de redes de alcantarillado conlleva tener un entendimiento de los posibles riesgos y consecuencias de la priorización de la inspección y/o la rehabilitación de los diferentes componentes del sistema (EPA, 2009). Sin embargo, este problema se encuentra inmerso en un contexto sociotécnico complejo en el cual se debe integrar información desde múltiples fuentes, pero a la vez se debe considerar datos limitados o incompletos; por lo cual el uso de la intuición por parte de los operadores de los sistemas en muchos casos es favorecida para tomar estas decisiones (Van Riel et al., 2014b). Luego, la implementación de métodos de mantenimiento preventivo como los métodos de minería de datos se ve limitada debido a la renuencia de estos operadores a modificar el proceso mediante el cual se realiza la gestión de activos en las empresas prestadoras del servicio.

De esta manera, es necesario evaluar la efectividad que tienen los procesos de rehabilitación guiados por la experiencia y el reconocimiento de patrones en contraste con los resultados que se obtienen al tomar las decisiones de priorización con base en los resultados de modelos que han considerado información técnica y operacional. A pesar de esto, pocas investigaciones se han desarrollado para estudiar la efectividad del uso de la intuición, entre las cuales se puede resaltar el

trabajo de Van Riel et al. (2014a). En este, los autores analizan los requisitos que deben satisfacerse para que una decisión basada en la intuición pueda considerarse competente (skilled) y no arbitraria, encontrando que para esto las decisiones deben tomarse con suficiente regularidad y debe existir oportunidad de aprender de estas decisiones. Encuentran que en el problema de priorización de activos a rehabilitar en redes de alcantarillado no es posible satisfacer estos requerimientos debido a que se evidencia que: (1) la consideración de múltiples factores acarrea la generación de discrepancias en las decisiones tomadas bajo las mismas circunstancias, (2) No todos los factores considerados por los operadores para la toma de decisiones tienen los mismos objetivos ni son medidos con los mismos criterios de desempeño y (3) las principales fuentes de información para la priorización de activos son la edad de las tuberías y las inspecciones de los pozos (Van Riel et al., 2014b), las cuales pueden ser difíciles de interpretar sin técnicas más avanzadas de análisis de datos.

Así, se puede identificar la necesidad de llevar a cabo estudios adicionales en que se investigue la viabilidad de los procesos de toma de decisiones actuales en diferentes contextos y la comparación de estos con resultados basados en los factores que influyen en el proceso de deterioro en redes de alcantarillado.

8.1.2 Preprocesamiento/Tratamiento de datos

Como se menciona en el capítulo 5, el preprocesamiento de datos es una de las etapas que se debe llevar a cabo para la generación de conocimiento y puede influir significativamente en los patrones o relacionadas encontradas al utilizar técnicas de minería de datos. Es posible reconocer dos principales formas de tratamiento de datos correspondientes a: (1) la limpieza de los datos para la eliminación de datos erróneos y/o valores atípicos y (2) la selección de factores relevantes para la modelación del proceso de deterioro.

En cuanto al primer proceso, Zhang & Yang, (2003) resaltan que la selección de la información relevante y los rangos apropiados de las variables resulta en una disminución significativa del tamaño de las bases de datos lo cual incrementa la eficiencia del proceso de minería de datos, al igual que identifican que el uso de datos de buena calidad permite encontrar patrones con alta confiabilidad. Por otro lado, la selección de las variables relevantes para su inclusión en el modelo puede constituir un proceso muy importante debido al reconocimiento de información redundante en las covariables estudiadas y el efecto negativo que puede generarse al construir modelos con un mayor número de variables, pues el uso de modelos más complejos no garantiza necesariamente resultados más robustos y confiables. Más aún, Mashford et al. (2011) analizan en su trabajo, la construcción de cuatro modelos SVM incrementando el número de variables incluidas en la modelación desde la consideración de atributos intrínsecos de las redes de alcantarillado hasta el uso de toda la información disponible. Estos autores concluyen que la inclusión de más variables en los modelos no resulta en una mejora del desempeño de los mismos; lo cual se puede atribuir a la

"maldición de la dimensionalidad", que indica que a medida que se incrementa la dimensión de entrada (parámetros influyentes), el número de datos requeridos para mantener el nivel de predicción aumenta exponencialmente (Mashford et al., 2011).

8.1.3 Problema de desequilibrio de clases en los datos

El problema de desequilibrio de clases en los datos, como se ha mencionado brevemente en los capítulos anteriores, consiste en un problema que surge cuando se están enfrentado tareas de clasificación, y este sucede cuando la distribución de la variable de respuesta $\hat{y}$ es sesgada y una o varias clases son muy diferentes al compararlas con otras (Carvalho, 2015). Múltiples investigaciones han demostrado que, en general, este desbalance de clases se presenta en la clasificación de las tuberías de acuerdo a su condición debido a que las redes tienden a tener una mayor cantidad de tuberías en condición buena o aceptable y muy pocas tuberías cercanas a la condición de colapso.

Así, al enfrentar el problema de clasificación mediante algoritmos de minería de datos en los conjuntos de datos de tuberías de redes de alcantarillado, es posible que estos no garanticen los resultados esperados debido al sesgo que se genera durante su entrenamiento y validación, si no se consideran las medidas de desempeño apropiadas y/o se implementan estrategias para trabajar con grupos de datos con desbalance de clases (Carvalho, 2015).

Algunas de las medidas que evitan la evaluación sesgada del desempeño de los modelos se presentan en el capítulo 5.2, y estas han sido utilizadas con mayor frecuencia en los estudios realizados en los últimos años (Ahmadi et al., 2015; Bailey et al., 2015; Caradot, Riechel, et al., 2018; Carvalho, 2015; Harvey & McBean, 2014; Harvey et al., 2015; Laakso, Kokkonen, et al., 2018; Mashford et al., 2011; Rokstad & Ugarelli, 2015) con el propósito de garantizar mayor robustez en los modelos en incrementar la confiabilidad por parte de las empresas prestadoras del servicio en los mismos. No obstante, también es posible implementar otras técnicas de aprendizaje automático para modificar el conjunto de datos con el cual se realiza el entrenamiento de los modelos, entre los cuales los métodos de muestreo son los más comunes y estos logran que las clases de los datos tengan la misma frecuencia mediante la eliminación o repetición de datos de una o varias clases (Carvalho, 2015).

8.1.4 Incertidumbre de los métodos de inspección en redes de alcantarillado

Los diferentes métodos de inspección de redes permiten obtener una categorización de la condición estructural u operacional en el que estas se encuentran. Sin embargo, estos resultados se obtienen a partir de una serie de procesos que puede incluir y propagar la incertidumbre desde la codificación de defectos por parte de las normativas hasta la parcialidad de las observaciones debido al componente humano. Luego, el estudio de esta incertidumbre en el problema de predicción de la condición de las tuberías resulta de gran importancia puesto a que los datos resultantes de la

inspección corresponden a la información de entrada requerida para el entrenamiento de los modelos (Caradot et al., 2013). De acuerdo con Dirksen et al. (2013), al analizar los resultados de inspecciones de tuberías realizadas mediante CCTV, la incertidumbre en los resultados obtenidos se ocasionan en cada una de las tres etapas del proceso de inspección: (1) reconocimiento del defecto, (2) descripción de los defectos y (3) interpretación de los reportes de la inspección. En este caso, los autores encuentran que, al identificar los defectos comúnmente encontrados en tuberías bajo la aplicación de diferentes códigos de inspección de tuberías, la tasa de falsos negativos (FP) en la clasificación de la mayoría de defectos puede llegar a alcanzar valores de hasta 75%, mientras que la tasa de falsos positivos generalmente se mantiene baja. Así, de acuerdo a lo anterior, existe una tendencia a subestimar los defectos encontrados en las tuberías inspeccionadas, lo cual puede resultar en consecuencias graves en caso de darse el colapso de las tuberías.

Por otro lado, Caradot, Rouault, Clemens, & Cherqui (2017) analizaron la propagación de la incertidumbre en los resultados de inspecciones de CCTV al evaluar la condición resultante de inspecciones dobles de la misma tubería en un caso de estudio de Alemania, encontrando que hay menos incertidumbre al evaluar la condición de tuberías con pocos defectos o sin defectos graves, pues los inspectores son más propensos a cometer errores cuando hay muchos defectos presentes. Finalmente, Ahmadi et al. (2015) analizan el uso de datos con incertidumbre e incompletos para la planeación de actividades de rehabilitación, encontrando que al incrementar la incertidumbre de la variable edad en un modelo de regresión logística se obtienen resultados más informativos que al no incluir esta información en el modelo. Asimismo, concluyen que al asignar rangos de una variable a una instancia en lugar de valores continuos e incluir la información de otras variables, es poco probable encontrar valores de incertidumbre tan altos como los analizados en este caso.

8.1.5 Capacidad de predicción, interpretación y validación de los modelos

La validación de los resultados de los modelos de minería de datos aplicados al problema de predicción de fallas en redes de alcantarillado es un aspecto muy importante a considerar cuando se analiza la percepción de los operadores de las empresas prestadoras del servicio respecto a la aplicación y uso de estas metodologías, puesto que la aceptación a los cambios que se requieren en muchos casos se ve limitada debido a la poca confianza que tienen los operadores para obtener resultados útiles a partir de estas técnicas. Sin embargo, pocas investigaciones llevan a cabo en sus trabajos un proceso de validación, en el cual se evalué el desempeño de los modelos en un conjunto de datos diferente al de entrenamiento.

La realización de este proceso es necesaria puesto a que la demostración del funcionamiento de estas técnicas puede motivar a las agencias prestadoras del servicio a invertir en procesos que permita mejorar el desempeño de las mismas (Caradot et al., 2017), y también porque a partir de estos resultados es posible interpretar los valores de las medidas de desempeño que se pueden

considerar como satisfactorios para la implementación de estos modelos en otras circunstancias (Mashford et al., 2011).

Por otro lado, la interpretabilidad de los resultados de los modelos también es un factor determinístico para la aplicación exitosa de la modelación predictiva en sistemas de alcantarillado, puesto que a pesar que se han encontrado medidas de desempeño mejores en la utilización de modelos más complejos en algunos casos de estudio, la capacidad de interpretar los factores que son relevantes y los patrones que estos tienen con el estado estructural de los activos en sistemas de alcantarillado permite a los tomadores de decisiones identificar más fácilmente las zonas o sectores de las redes que pueden requerir un mantenimiento proactivo. Y en este sentido, también es posible identificar los requerimientos de información que puede no estar siendo registrada mediante las inspecciones de los sistemas; más aún, al mejorar los beneficios que tienen los datos recopilados se impulsa la recolección y utilización de estos datos en procesos de planificación, lo cual puede resultar en el incremento de la calidad de la información recolectada (Rokstad & Ugarelli, 2016; Rokstad et al., 2015).

8.1.6 Calidad y cantidad de información para la calibración

La cantidad y la calidad de información disponible acerca de los activos de redes de alcantarillado puede corresponder a una limitación para la transición hacia una gestión proactiva de las redes mediante la modelación predictiva como también puede constituir un gran incentivo para su implementación. Lo anterior, es relativo al contexto en que se estudie la viabilidad de aplicar estas metodologías teniendo en cuenta factores como el tipo de mantenimiento histórico realizado a las redes, la confiabilidad en las técnicas de inspección y evaluación de la condición de las tuberías, y la capacidad de las empresas prestadoras del servicio de adoptar el uso eficiente de estas metodologías y garantizar su sostenimiento en el tiempo a partir de una continua recolección de información (Ahmadi et al., 2014b; Rokstad & Ugarelli, 2016; Van der Steen et al., 2014; Van Riel et al., 2014b; Van Riel, Van Bueren, Langeveld, Herder, & Clemens, 2015). En este sentido, es esperado que empresas en las cuales se ha llevado a cabo históricamente un mantenimiento de sus activos principalmente de forma reactiva se enfrenten a un reto mucho más grande para la gestión proactiva de sus sistemas, frente a otras ciudades en las cuales se han implementado diversas medidas de mantenimiento preventivas y proactivas a lo largo de la vida útil de sus sistemas. En particular, lo anterior tiene un gran impacto en la aceptación o confiabilidad que se tiene en técnicas de modelamiento predictivo cuando se trata de la cantidad de datos disponible.

No obstante, a pesar de que la cantidad de información registrada puede no representar un inconveniente para la implementación de modelos de deterioro en ciudades con altas tasas de inspección, la calidad de los datos de inspecciones y también la calidad de los datos con los cuales se calibran los modelos es un punto clave para todas las empresas prestadoras de servicio que buscan hacer un uso eficiente de su información registrada, al igual que mejorar sus procesos de

recolección de información con base en los datos relevantes para el entendimiento del comportamiento de sus sistemas.

En este sentido, diversas investigaciones como las mencionadas en el capítulo 5.4 han analizado las limitaciones que pueden presentarse para el uso costo-efectivo de la información registrada en realizar una gestión proactiva de sus redes que evite los costos económicos, sociales y de salud pública que se generan cuando se presenta una falla en este tipo de infraestructuras. Entre estas limitaciones se destacan la agregación de la información en las normativas/códigos de inspección de las redes, la representatividad de los conjuntos de datos utilizados para la calibración de los modelos y la ausencia del registro de variables que pueden ser explicativas para un mecanismo de falla en particular.

En este caso, se observa una gran necesidad de realizar futuras investigaciones que permitan cuantificar los beneficios y costos obtenidos al recopilar información en suficiente cantidad y adecuada calidad para incentivar los procesos de gestión proactiva de redes de alcantarillado.

9 CONCLUSIONES

Las redes de agua potable y alcantarillado de las ciudades constituyen una gran parte de la infraestructura que permite garantizar estándares básicos de calidad de vida y la protección de la salud pública de los habitantes de estas ciudades. Es por esto, que en los últimos años se han realizado esfuerzos importantes hacia garantizar una alta cobertura de las ciudades y las zonas rurales que permita el abastecimiento y disposición adecuada del agua potable y las aguas residuales y pluviales. Sin embargo, garantizar una prestación del servicio de forma continua y eficiente también depende de otros factores adicionales a la cobertura de las redes que son inherentes a las características de los sistemas y su deterioro a lo largo de su vida útil. En particular, una buena operación y mantenimiento de estos sistemas es vital para evitar el deterioro acelerado y de gran impacto en los activos bajo tierra de las redes.

Teniendo en cuenta lo anterior, en esta investigación se buscó en primer lugar diagnosticar las diferentes técnicas de mantenimiento utilizadas en la actualidad por las empresas prestadoras del servicio e identificar los beneficios y costos que se generan al implementar los diferentes tipos de gestión de los activos de las redes de alcantarillado. Entre los resultados encontrados, se destaca la necesidad actual de las empresas prestadoras de servicios de realizar una transición de las técnicas de mantenimiento correctivo hacia una gestión proactiva en la cual se lleven a cabo en mayor parte actividades preventivas y de ser posible predictivas que eviten la ocurrencia de eventos de falla o perdida de operación en el sistema que pueden resultar en costos muy elevados tanto a nivel económico como social. Así, se identificaron una serie de enfoques principalmente de mantenimiento predictivo, con los cuales se ha buscado enfrentar los retos que implica el mantenimiento de activos que se encuentran en un estado sub-crítico y pueden ser propensos a fallar pero no han sido identificados por medio de indicadores de servicio (cuando fallan).

Este tipo de metodologías busca estimar aproximadamente el comportamiento esperado de los activos en las redes de alcantarillado considerando registros de información disponibles o en algunos casos la opinión y ponderación de expertos frente a las condiciones que pueden resultar en el deterioro acelerado de algunos activos. Diferentes metodologías fueron identificadas desde herramientas para el apoyo a la decisión, la modelación hidráulica, modelos estadísticos y modelos de aprendizaje automático (machine learning); encontrando que se ha considerado más beneficioso la aplicación de estos últimos modelos al problema de deterioro de activos en las redes de alcantarillado debido a limitaciones de otras aproximaciones como la subjetividad de los resultados, las fuertes suposiciones de los modelos estadísticos, y los altos requerimientos de tiempo y recursos que pueden implicar enfoques como la modelación hidráulica para la determinación de posibles fallas en redes de alcantarillado.

En particular, fue posible cuantificar a partir de la revisión bibliográfica realizada, los diferentes problemas asociados a la toma de decisiones respecto al mantenimiento de estos sistemas con base en la opinión de expertos y/o directores de las áreas encargadas de la gestión de las redes en las empresas; pues diversos estudios han establecido la dificultad de entender el comportamiento de este tipo de sistemas que, además, se ven inmersos en muchos casos en procesos de toma decisiones con fundamentos políticos, que dificultan la gestión de los mismos con base en las necesidades reales generadas por el envejecimiento de los sistemas. Además, fue posible establecer que los procesos físicos de deterioro a los cuales se ven sometidos estos sistemas tienen una complejidad muy alta y que invertir esfuerzos en la determinación de estos procesos físicos puede llevar a modelos con una complejidad excesiva que impida su utilización eficiente o a modelos muy simples que no son capaces de estimar acertadamente estos patrones de deterioro.

Así mismo, se identificó que los esfuerzos para la gestión proactiva de redes de alcantarillado ha buscado integrar metodologías que permitan cuantificar los mecanismos de falla de las tuberías y las variables relevantes en estos procesos de una manera objetiva, de forma que el proceso de toma de decisiones se encuentre soportado en las necesidades reales de rehabilitación de estos sistemas. Por lo anterior, se han considerado apropiadas algunas técnicas de aprendizaje automático, con las cuales se realiza el entrenamiento de un modelo matemático (maquina) a partir de registros históricos para la identificación de patrones y/o variables explicativas de los procesos de falla. Más aún, se logró enmarcar el problema de asignar una clasificación estructural a las tuberías de alcantarillado en el contexto de una de las aplicaciones de la Minería de datos correspondiente a la clasificación de observaciones en diferentes categorías. Con lo anterior, se estableció que el problema de la estimación de la clase estructural de tuberías en redes de alcantarillado puede realizarse tomando esta clase estructural como la variable a predecir y diferentes factores físicos/ambientales/de construcción como variables predictoras en un modelo de clasificación.

Entre las diferentes metodologías revisadas se resaltó el uso de modelos como regresión lineal, regresión logística, arboles de decisión, bosques aleatorios, máquinas de soporte vectorial, regresión polinómica evolutiva y redes neuronales. Mediante la aplicación de estos modelos a diferentes casos de estudio en los últimos 20 años se han logrado estimar en zonas con diferentes características topológicas algunos de los patrones de deterioro que pueden identificarse a partir de los registros de inspecciones realizadas previamente a las redes. Sin embargo, también se han identificado una serie de obstáculos relacionados con los tratamientos de los datos, la subjetividad de los procesos de inspección, la calidad y cantidad de información y la cultura empresarial, para la implementación efectiva de estas metodologías.

Frente a lo anterior, se encuentra que la aplicación del modelamiento predictivo como herramienta para el mantenimiento proactivo de las redes requiere un enfoque sistemático, en el cual se consideren todos los pasos necesarios para: preprocesamiento de la información, determinar las variables influyentes en los procesos de deterioro, la división de los datos (entrenamiento y prueba),

implementación y evaluación de los modelos, y la correcta interpretación de los resultados. Así mismo, es importante considerar que en muchos casos, los resultados de estos modelos se encuentran sujetos a las condiciones operacionales de los sistemas, por lo cual la generalización de modelos aplicados en otras circunstancias puede llevar a resultados erróneos. La extrapolación de resultados de estos modelos en estas ciudades requiere no solo de un análisis riguroso de las condiciones de referencia y las condiciones del caso al cual se desea aplicar, sino también de un diagnóstico de la información disponible en las nuevas condiciones. Incluso en el caso en el cual se cuente con condiciones similares, lo anterior puede representar una gran dificultad para la aplicación de modelos que han sido obtenidos con información representativa y de calidad como una primera aproximación en ciudades con pocos registros históricos o bajas tasas de inspección, debido a la gran diversidad de variables predictoras utilizadas. Más aún, la incertidumbre de la clasificación se encuentra inherentemente asociada al hecho de que este proceso se lleva a cabo por humanos, cuya percepción no garantiza una completa objetividad. Por último, respecto a la aplicación de estos modelos, es necesario que las empresas prestadoras del servicio se aseguren del uso de las medidas de desempeño adecuadas, pues el éxito al aplicar estas metodologías depende en gran parte de su correcta interpretación y calibración.

Una de las principales dificultades identificadas transversalmente en los estudios que han buscado aplicar estas técnicas corresponde a la poca disponibilidad de información completa, de calidad, confiable y representativa. Lo anterior, siendo una restricción con mayor importancia en países y/o ciudades en los cuales se reportan bajas tasas de inspección de sus redes o cuentan con registros poco confiables de las mismas. A nivel nacional, se logró identificar que las condiciones anteriores de poca disponibilidad de información completa y de calidad son comunes en las empresas prestadoras de servicio debido a una tradición histórica de mantenimiento correctivo de sus redes, en particular se analizó el caso de la ciudad de Bogotá.

Teniendo en cuenta lo anterior y la literatura actual en la cual se reporta la necesidad de establecer claramente los beneficios y costos de implementar modelos de deterioro para lograr su implementación y uso adecuado por parte de las empresas de acueducto y alcantarillado, se llevó a cabo un análisis de las ventajas y retos que pueden presentarse al considerar este tipo de técnicas. Más aún, dado el contexto de la ciudad de Bogotá se planteó un ejercicio numérico en una base de datos sintética de la red de alcantarillado de la zona 1 de la ciudad, en el cual se estudió el efecto de la cantidad de datos disponible para la calibración de diferentes modelos de minería de datos en la capacidad de estos modelos de generalizar los comportamientos para toda la red. Se encontró que entre los beneficios de incrementar las tasas de inspección de sus redes se podría realizar una estimación con mucha precisión y poca dispersión mediante modelos de minería de datos que no requieren un tratamiento exhaustivo de los datos y cuya calibración a partir de un conjunto limitado de datos permite indagar e interpretar los patrones de deterioro que se presentan en las tuberías de las redes de alcantarillado obteniendo medidas de desempeño consideradas como aceptables.

Finalmente, es muy importante resaltar las necesidades de futuras investigaciones que se encontraron a partir de este estudio, entre las cuales se encuentran: la representatividad de las muestras de datos disponibles para la calibración de los modelos, la necesidad de cuantificar la relación beneficio/costo de incrementar tasas de inspección para la modelación predictiva que incentiven a las empresas, la relevancia de las variables encontradas como determinantes en los procesos de deterioro cuando se cuenta con información limitada y los efectos de las normativas o códigos de inspección en la capacidad predictiva de diferentes modelos, a nivel nacional y local. Así mismo, se deben tener en cuenta las limitaciones del ejercicio numérico aplicado, considerando su aplicación en una base de datos sintética generada a partir de un criterio de falla propuesto. El conjunto de datos resultante podría no incluir todos los patrones existentes en las bases de datos reales, al igual que considerar fallas únicamente en un conjunto de tuberías que pueden presentar un estado muy crítico. Los resultados deben analizarse con cautela pues la realización de este ejercicio se planteó como un primer paso hacia la futura identificación de las necesidades de cantidad de datos de estos modelos para estimar el comportamiento general de todas las tuberías en las redes de alcantarillado.

10 REFERENCIAS

Ahmadi, M., Cherqui, F., Aubin, J., & Le Gauffre, P. (2016). Sewer asset management: impact of sample size and its characteristics on the calibration outcomes of a decision-making multivariate model. *Urban Water Journal*, *13*(1), 41-56. https://doi.org/10.1080/1573062X.2015.1011668

Ahmadi, M., Cherqui, F., De Massiac, J., & Le Gauffre, P. (2014a). From sewer inspection programmes to rehabilitation needs: Research and results related to data quality and availability with the support of numerical experiment. *European Journal of Environmental and Civil Engineering*, *18*(10), 1145-1156. https://doi.org/10.1080/19648189.2014.893212

Ahmadi, M., Cherqui, F., De Massiac, J., & Le Gauffre, P. (2014b). Influence of available data on sewer inspection program efficiency. *Urban Water Journal*, *11*(8), 641-656. https://doi.org/10.1080/1573062X.2013.831910

Ahmadi, M., Cherqui, F., De Massiac, J., & Le Gauffre, P. (2015). Benefits of using basic, imprecise or uncertain data for elaborating sewer inspection programmes. *Structure and Infrastructure Engineering*, *11*(3), 376-388. https://doi.org/10.1080/15732479.2014.887122

Ana, E., & Bauwens, W. (2007). Sewer network asset management decision-support tools: a review. En *International Symposium on New Directions in Urban Water Management* (pp. 1-8). Recuperado de http://www2.gtz.de/Dokumente/oe44/ecosan/en-sewer-network-decision-making-tool-2007.pdf

Ana, E., & Bauwens, W. (2010). Modeling the structural deterioration of urban drainage pipes: The state-of-the-art in statistical methods. *Urban Water Journal*, *7*(1), 47-59. https://doi.org/10.1080/15730620903447597

Ana, E., Bauwens, W., Pessemier, M., Thoeye, C., Smolders, S., Boonen, I., & De Gueldre, G. (2009). An investigation of the factors influencing sewer structural deterioration. *Urban Water Journal*, *6*(4), 303-312. https://doi.org/10.1080/15730620902810902

Angarita, H., Vargas, D., & Torres, A. (2017). Identifying explanatory variables of structural state for optimum asset management of urban drainage networks : a pilot study for the city of Bogota Identificación de factores de riesgo para la gestión patrimonial óptima. *Ingeniería e Investigación*, *37*(2), 6-16. https://doi.org/10.15446/ing.investig.v37n2.57752

Ariaratnam, S., El-Assaly, A., & Yang, Y. (2001). Assessment of infrastructure inspection needs using logistic models. *Journal of Infrastructure Systems*, *24*(7), 261-281.

Arthur, S., & Crow, H. (2007). Prioritising sewerage maintenance using serviceability criteria. *Proceedings of the Institution of Civil Engineers - Water Management*, *160*(3), 189-194. https://doi.org/10.1680/wama.2007.160.3.189

Arthur, S., Crow, H., & Pedezert, L. (2008). Understanding blockage formation in combined sewer

networks. *Proceedings of the Institution of Civil Engineers - Water Management*, *161*(4), 215-221. https://doi.org/10.1680/wama.2008.161.4.215

Arthur, S., Crow, H., Pedezert, L., & Karikas, N. (2009). The holistic prioritisation of proactive sewer maintenance. *Water Science and Technology*, *59*(7), 1385-1396. https://doi.org/10.2166/wst.2009.134

Baik, H., Jeong, H., & Abraham, D. (2006). Estimating transition probabilities in Markov Chain-Based deterioration models for management of Wastewater Systems. *Journal of Water Resources Planning and Management*, *132*, 15-24. https://doi.org/10.1061/(ASCE)0733-9496(2006)132:1(15)

Bailey, J., Keedwell, E., Djordjevic, S., Kapelan, Z., Burton, C., & Harris, E. (2015). Predictive risk modelling of real-world wastewater network incidents. *Procedia Engineering*, *119*(1), 1288-1298. https://doi.org/10.1016/j.proeng.2015.08.949

Berardi, L., Giustolisi, O., Savic, D., & Kapelan, Z. (2009). An effective multi-objective approach to prioritisation of sewer pipe inspection. *Water Science and Technology*, *60*(4), 841-850. https://doi.org/10.2166/wst.2009.432

Butler, D., Davies, J. (2011). *Urban Drainage, 3rd Edition*.

Caradot, N., Hernandez, N., Sonnenberg, H., Torres, A., & Rouault, P. (2018). From CCTV Data to Strategic Planning: Deterioration Modelling for Large Sewer Networks in Germany and Colombia. En *International Conference on Hydroinformatics* (Vol. 3, pp. 351-355). https://doi.org/10.29007/nbx2

Caradot, N., Kley, G., Kropp, I., & Schmidt, T. (2013). *Review of Available Technologies and Methodologies for Sewer Condition Evaluation (SEMA)*.

Caradot, N., Riechel, M., Fesneau, M., Hernandez, N., Torres, A., Sonnenberg, H., … Rouault, P. (2018). Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in Berlin, Germany. *Journal of Hydroinformatics*, *20*(5), 1131-1147. https://doi.org/10.2166/hydro.2018.217

Caradot, N., Rouault, P., Clemens, F., & Cherqui, F. (2017). Evaluation of uncertainties in sewer condition assessment. *Structure and Infrastructure Engineering*, *14*(2), 264-273. https://doi.org/10.1080/15732479.2017.1356858

Caradot, N., Sonnenberg, H., Kropp, I., Ringe, A., Denhez, S., Hartmann, A., & Rouault, P. (2017). The relevance of sewer deterioration modelling to support asset management strategies. *Urban Water Journal*, *14*(10), 1007-1015. https://doi.org/10.1080/1573062X.2017.1325497

Caradot, N., Sonnenberg, H., Kropp, I., Schmidt, T., Ringe, A., Denhez, S., … Rouault, P. (2014). Sewer deterioration modelling for asset management strategies. *Water Asset Management International*, *10*(3), 1-24.

Carvalho, G. (2015). *Data Mining techniques to predict sewer condition*.

Carvalho, G., Amado, C., Brito, R., Coelho, S., & Leitão, J. (2018). Analysing the importance of variables for sewer failure prediction. *Urban Water Journal*, *15*(4), 338-345. https://doi.org/10.1080/1573062X.2018.1459748

Cochran, W. G. (1977). *SamplingTechniques*. *Sampling Techniques*. Recuperado de https://scholar.google.com.tr/scholar?q=sampling+techniques&btnG=&hl=en&as_sdt=0,5#0

Damvergis, C. (2014). Sewer systems: Failures and rehabilitation. *Water Utility Journal*, *8*, 17-24. Recuperado de http://www.ewra.net/wuj/pdf/WUJ_2014_08_02.pdf

Daumé, H. (2017a). Decision trees. En *A course in machine learning* (pp. 1-18).

Daumé, H. (2017b). Linear model. En *A course in machine learning* (pp. 1-17).

Davies, J., Clarke, B., Whiter, J., & Cunningham, R. (2001). Factors influencing the structural deterioration and collapse of rigid sewer pipes. *Urban Water*, *3*(1-2), 73-89. https://doi.org/10.1016/S1462-0758(01)00017-6

Deng, N., Tian, Y., & Zhang, C. (2013). *Support vector machines: optimization based theory, algorithms and extensions*.

Dirksen, J., Clemens, F., Korving, H., Cherqui, F., Le Gauffre, P., Ertl, T., … Snaterse, C. (2013). The consistency of visual sewer inspection data. *Structure and Infrastructure Engineering*, *9*(3), 214-228. https://doi.org/10.1080/15732479.2010.541265

Duncan, H., & Arthur, S. (2005). The development of a methodology to manage the proactive maintenance of sewerage assets within the context of serviceability. En *10th International Conference on Urban Drainage* (pp. 21-26).

El Espectador. (2018, julio 26). Lo que falta en suministro de agua y alcantarillado en Colombia. Recuperado de https://www.elespectador.com/economia/lo-que-falta-en-suministro-de-agua-y-alcantarillado-en-colombia-articulo-802501

El Tiempo. (2017). En el 2020 Bogotá tendra 300.000 habitantes más. Recuperado de https://www.eltiempo.com/bogota/poblacion-por-edades-de-bogota-2017-109238

Empresa de Acueducto de Bogotá E.S.P. NS-061 Aspectos técnicos para la rehabilitación de redes y estructuras de alcantarillado (2001).

Empresa de Acueducto de Bogotá E.S.P. (2006). *PLAN MAESTRO ACUEDUCTO Y ALCANTARILLADO*.

Empresa de Acueducto de Bogotá E.S.P. NS-058 Aspectos técnicos para inspección de redes y estructuras de alcantarillado (2010).

Empresa de Acueducto de Bogotá E.S.P. (2016). *Informe de gestión*.

Empresa de Acueducto de Bogotá E.S.P. (2017). *Informe de Gestión*.

Empresa de Acueducto de Bogotá E.S.P. (2018). *Informe de gestión*.

Empresa de Acueducto de Bogotá E.S.P. (2019a). Gestión Empresarial. Recuperado de https://www.acueducto.com.co/wps/portal/EAB/aempsecsecundaria/bempresagestionempresarial/!ut/p/z1/tVNdT-MwEPwtPOSx9SahbXJvKVfBIRVE-bjGL8hx3cRHY7u208C_Z6tWSNyVRgguSuR41jszu9YSSuaEKraRJfNSK7bCfU6Hj2n6MwsjiK7OL9IYsodZ9Cu7OwMYRuSBUEINlwuSpwCj0xDgNBYDxkYsGaYhE

Empresa de Acueducto de Bogotá E.S.P. (2019b). Zonas. Recuperado de https://www.acueducto.com.co/wps/portal/EAB/aempsecsecundaria/empresazonas/!ut/p/z1/tZNtT8IwEMc_zV5CbxsPw3dDiUhEjcrD9sZ0XQdV1pa2A-XTewgxGlFjiMuWbde7__-X65WkZEpSSVdiRp1Qki7wP0lbD53OWewHEFyd9zshxOPb4CK-PwVoBWRMUpJqJnKSsLbfLHKARsibWd4qonbeoAXLMr8ZsTb1t5lMOu3

EPA. (2009). *White Paper on Condition Assessment of Wastewater Collection Systems*. https://doi.org/10.1080/02560054.2014.886658

EPA. (2017). Eliminating Sanitary Sewer Overflows in New England. Recuperado de https://www3.epa.gov/region1/sso/

Géron, A. (2017). *Hands-on machine learning with scikit-learn & tensorflow*. *O'Reilly Media*. O'Reilly Media. https://doi.org/10.3389/fninf.2014.00014

Giustolisi, O., & Savic, D. (2006). A symbolic data-driven technique based on evolutionaty polynomial regression. *Journal of Hydroinformatics*, *8*(3), 207-222. https://doi.org/10.2166/hydro.2006.020

Giustolisi, O., Savic, D., & Laucelli, D. (2004). Data Mining for Management and Rehabilitation of Water Systems : The Evolutionary Polynomial Regression Approach. *Wasserbauliche Mitteilungen*, (27), 285-296.

Grus, J. (2015). *Data science from scratch: first principles with phyton*. Recuperado de http://math.ecnu.edu.cn/~lfzhou/seminar/[Joel_Grus]_Data_Science_from_Scratch_First_Princ.pdf

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in Data Mining. *International Journal of Computer Applications*, *163*(8), 15-19. Recuperado de https://pdfs.semanticscholar.org/fd39/e1fa85e5b3fd2b0d000230f6f8bc9dc694ae.pdf

Harvey, R., & McBean, E. (2014). Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure. *Journal of Hydroinformatics*, *16*(6), 1265-1279. https://doi.org/10.2166/hydro.2014.007

Harvey, R., Wheeler, A., & Mcbean, E. (2015). A Data Mining Tool for Planning Sanitary Sewer

Condition Inspection. En *Conflict Resolution in Water Resources and Environmental Management* (pp. 1-20). https://doi.org/10.1007/978-3-319-14215-9

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning Data mining, inference and prediction*. *Math. Intell.* https://doi.org/111

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. *Springer* (Vol. 7). https://doi.org/10.1007/978-1-4614-7138-7

Jung, I., Garrett Jr., J., Soibelman, L., & Lipkin, K. (2012). Application of classification models and spatial clustering analysis to a sewage collection system of a mid-sized city. *Computing in Civil Engineering*, 537-544. https://doi.org/10.1061/9780784412343.0068

Khan, Z., Zayed, T., & Moselhi, O. (2010). Structural Condition Assessment of Sewer Pipelines. *Journal of Performance of Constructed Facilities*, *24*(2), 170-179. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000081

Kleidorfer, M., Möderl, M., Tscheikner-Gratl, F., Hammerer, M., Kinzel, H., & Rauch, W. (2013). Integrated planning of rehabilitation strategies for sewers. *Water Science and Technology*, *68*(1), 176-183. https://doi.org/10.2166/wst.2013.223

Kohavi, R., & Provost, F. (1998). Machine learning, 271-274. https://doi.org/10.1023/A

Krenker, A., Bester, J., & Kos, A. (2011). Introduction to the Artificial Neural Networks. En *Artificial Neural Networks - Methodological Advances and Biomedical Applications* (pp. 4-18). https://doi.org/10.5772/15751

Laakso, T., Ahopelto, S., Lampola, T., Kokkonen, T., & Vahala, R. (2018). Estimating water and wastewater pipe failure consequences and the most detrimental failure modes. *Water Science and Technology: Water Supply*. https://doi.org/10.2166/ws.2017.164

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2018). Sewer condition prediction and analysis of explanatory factors. *Water (Switzerland)*, *10*(9), 1-17. https://doi.org/10.3390/w10091239

Lab41. (2019). The 10 algorithms machine learning engineers need to know. Recuperado de https://gab41.lab41.org/the-10-algorithms-machine-learning-engineers-need-to-know-f4bb63f5b2fa

Liu, Y., & Zheng, Y. (2005). One-against-all multi-class SVM classification using reliability measures. En *International Joint Conference on Neural Networks* (pp. 1-4). https://doi.org/10.1109/IJCNN.2005.1555963

Lohr, S. L. (2010). *Sampling: Design and Analysis*. https://doi.org/10.1017/CBO9781107415324.004

López Kleine, L., Hernandez, N., & Torres, A. (2016). Physical characteristics of pipes as indicators of structural state for decision-making considerations in sewer asset management. *Ingeniería e Investigación*, *36*(3), 15-21. https://doi.org/10.15446/ing.investig.v36n3.56616

Mashford, J., Marlow, D., Tran, D., & May, R. (2011). Prediction of Sewer Condition Grade Using Support Vector Machines. *Journal of Computing in Civil Engineering*, *25*(4), 283-290. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000089

Mcdonald, S., & Zhao, J. (2001). Condition assessment and rehabilitation of large sewers. En *International Conference on Underground Infrastructure Research* (pp. 361-369).

Micevski, T., Kuczera, G., & Coombes, P. (2002). Markov Model for Storm Water Pipe Deterioration. *Journal of Infrastructure Systems*, *8*(2), 49-56. https://doi.org/10.1061/(ASCE)1076-0342(2002)8:2(49)

Morgan, J., Dougherty, R., Hilchie, A., & Carey, B. (2003). Sample Size and Modeling Accuracy with Decision Tree Based Data Mining Tools. *Academy of Information and Management Science Journal*, *6*(2), 71-99. https://doi.org/10.1007/s13398-014-0173-7.2

New England Interstate Water Pollution Control Commission. (2003). *Optimizing operation, maintenance, and rehabilitation of sanitary sewer collection systems*.

Organizacion Panamericana de la Salud. (2005). *Operación y mantenimiento de sistemas de alcantarillado sanitario en el medio rural*. Recuperado de http://www.bvsde.paho.org/tecapro/documentos/sanea/152esp-O&M-alcantar.pdf

Rencher, A., & Schaalje, B. (2007). *Linear models in statistics*. *John Wiley & Sons*. https://doi.org/10.1002/9780470192610

Rokstad, M., & Ugarelli, R. (2015). Evaluating the role of deterioration models for condition assessment of sewers. *Journal of Hydroinformatics*, *17*(5), 789-804. https://doi.org/10.2166/hydro.2015.122

Rokstad, M., & Ugarelli, R. (2016). Improving the benefits of sewer condition deterioration modelling through information content analysis. *Water Science and Technology*, *74*(10), 2270-2279. https://doi.org/10.2166/wst.2016.419

Rokstad, M., Ugarelli, R., & Sægrov, S. (2015). Improving data collection strategies and infrastructure asset management tool utilisation through cost benefit considerations. *Urban Water Journal*, *13*(7), 710-726. https://doi.org/10.1080/1573062X.2015.1024692

Salman, B., & Salem, O. (2012). Modeling Failure of Wastewater Collection Lines Using Various Section-Level Regression Models. *Journal of Infrastructure Systems*, *18*(2), 146-154. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000075

Savic, D., Giustolisi, O., & Laucelli, D. (2009). Asset deterioration analysis using multi-utility data and multi-objective data mining. *Journal of Hydroinformatics*, *11*(3), 211-225. https://doi.org/10.2166/hydro.2009.019

Savic, D., Giustolisi, O., & Shepherd, W. (2006). Modelling sewer failure by evolutionary computing. *Journal of Water Management*, (WM2), 111-118.

Scheidegger, A., & Maurer, M. (2012). Identifying biases in deterioration models using synthetic sewer data. *Water Science and Technology*, *66*(11), 2363-2369. https://doi.org/10.2166/wst.2012.471

Stanić, N., Langeveld, J., & Clemens, F. (2014). HAZard and OPerability (HAZOP) analysis for identification of information requirements for sewer asset management. *Structure and Infrastructure Engineering*, *10*(11), 1345-1356. https://doi.org/10.1080/15732479.2013.807845

Stein, D., & Stein, R. (2004). *Rehabilitation and maintenance of drains and sewers*. *Unitracc*. Recuperado de http://www.unitracc.com/know-how/fachbuecher/rehabilitation-and-maintenance-of-drains-and-sewers/damage-its-causes-and-its-consequences/cracks-pipe-breaks-collapse-en

Syachrani, S., Seok, J., & Chung, C. (2013). Decision Tree–Based Deterioration Model for Buried Wastewater Pipelines. *Journal of Performance of Constructed Facilities*, *19*, 633-645. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000349.

Torres, M., Rodríguez, J., & Leitão, J. (2017). Geostatistical analysis to identify characteristics involved in sewer pipes and urban tree interactions. *Urban Forestry and Urban Greening*, *25*(April), 36-42. https://doi.org/10.1016/j.ufug.2017.04.013

Udhayakumarapandian, D., & Chandrasekaran, R. (2016). Data Size versus Accuracy: Performance by different Data Mining Tools. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *5*(3), 577-580. Recuperado de https://github.com/Dans-labs/recommender-systems/blob/.../d

Ugarelli, R., Kristensen, S., Røstum, J., Sægrov, S., & Di Federico, V. (2009). Statistical analysis and definition of blockages-prediction formulae for the wastewater network of Oslo by evolutionary computing. *Water Science and Technology*, *59*(8), 1457-1470. https://doi.org/10.2166/wst.2009.152

Ugarelli, R., Selseth, I., Le Gat, Y., Rostum, J., & Krogh, A. (2013). Wastewater pipes in Oslo: From condition monitoring to rehabilitation planning. *Water Practice and Technology*, *8*(3-4), 487-494. https://doi.org/10.2166/wpt.2013.051

UIAF. (2014). *Técnicas de minería de datos para la detección y prevención del lavado de activos y la financiación del terrorismo (LA/FT)*.

Van der Steen, A., Dirksen, J., & Clemens, F. (2014). Visual sewer inspection: detail of coding system versus data quality? *Structure and Infrastructure Engineering*, *10*(11), 1385-1393. https://doi.org/10.1080/15732479.2013.816974

Van Riel, W., Langeveld, J., Herder, P., & Clemens, F. (2014a). Integrating road and sewer works : risk attitude and costs. En *World Congress on Engineering Asset Management* (pp. 1-12). Pretoria, South Africa.

Van Riel, W., Langeveld, J., Herder, P., & Clemens, F. (2014b). Intuition and information in decision-making for sewer asset management. *Urban Water Journal*. Taylor & Francis. https://doi.org/10.1080/1573062X.2014.904903

Van Riel, W., Van Bueren, E., Langeveld, J., Herder, P., & Clemens, F. (2015). Decision-making for sewer asset management: Theory and practice. *Urban Water Journal*, *13*(1), 57-68. https://doi.org/10.1080/1573062X.2015.1011667

Wirahadikusumah, R., Abraham, D., & Iseley, T. (2001). Challenging issues in modeling deterioration of combined sewers. *Journal of Infrastructure Systems*, *7*(1), 77-84.

Witten, I., Frank, E., Hall, M., & Pal, C. (2017). *Data Mining: Practical machine learning tools and techniques*. Recuperado de https://bit.ly/2RizoxG

WRc. (2001). *Sewer rehabilitation manual*.

Wright, L., Heany, J., & Dent, S. (2006). Prioritizing sanitary sewers for rehabilitation using least-cost classifiers. *Journal of Infrastructure Systems*, *12*(3), 174-183. https://doi.org/10.1016/j.tree.2006.02.003

Zaher, K. (s. f.). Design of sewer systems. Cairo: Cairo University. Recuperado de https://scholar.cu.edu.eg/khaledzaher/files/2__design_of_sewers_v2_0.pdf

Zhang, C., & Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence*, *17*, 375-381. https://doi.org/10.1080/713827180