Study of the spatiotemporal correlation between sediment-related blockage events in the sewer system in Bogotá (Colombia)

Eliana Soriano Pulido, Carlos Valencia Arboleda and Juan Pablo Rodríguez Sánchez

ABSTRACT

The planning and scheduling of maintenance operations of large conventional sewer systems generate a complex decision-making environment due to the difficulty in the collection and analysis of the spatiotemporal information about the operational and structural condition of their components (e.g. pipes, gully pots and manholes). As such, water utilities generally carry out these operations following a corrective approach. This paper studies the impact of the spatiotemporal correlation between these failure events using Log-Gaussian Cox Process (LGCP) models. In addition, the association of failure events to physical and environmental covariates was assessed. The proposed methods were applied to analyze sediment-related blockages in the sewer system of an operative zone in Bogotá (Colombia). The results of this research allowed the identification of significant covariates that were further used to model spatiotemporal clusters with high sediment-related failure risk in sewer systems. The LGCP model proved to be more accurate in comparison to those models that build upon a fundamental assumption that a failure is equally likely to occur at any time regardless of the state of the system and the system’s history of failures (i.e. a homogeneous Poisson process model).

Key words | intensity of failures, Log-Gaussian Cox Process, sediment-related blockages, sewer systems, spatiotemporal correlation

INTRODUCTION

Urban drainage systems provide fundamental services to the health and well-being of populations, as well as to the protection of the environment (Santos et al. 2017). However, it is well known that wastewater and urban runoff transport a variety of solid particles that can create obstructions in the sewer system, thus deteriorating its operation. Sediment-related blockages may result in the release of raw sewage into the environment due to a reduced hydraulic capacity that increases the probability of local surcharge and/or combined sewer overflow events (Xie et al. 2017). The malfunctioning of the sewer system and subsequent flooding constitute a risk to human health as a result of exposure to potential water-borne pathogens (ten Veldhuis et al. 2010; Collender et al. 2016). Furthermore, the release of untreated wastewater (e.g. via sewer overflows) into water bodies such as rivers and lakes negatively alters the water quality and health of fluvial and lacustrine ecosystems, affecting fish and other types of aquatic life (Munawar et al. 1993; Rodríguez-Jeangros et al. 2018). In general, the consequences of these contaminated water flow releases range from mild disruption for a small number of residents to events that may seriously damage the environment (Xie et al. 2017).

Urban drainage services account for a significant portion of the built infrastructures, which require large investments for their construction, protection, and maintenance (Santos et al. 2017). In different studies, it has been stated that blockage events account for a significant proportion of reported failures in sewer systems (Jin & Mukherjee 2010; Xie et al. 2017). These blockages are...
normally detected when a component of the system, or the system as a whole, stops performing as expected (Santos et al. 2017). Although many sewer system failures triggering mechanisms have been identified, sediment-related sewer blockages often appear at random (Rodríguez et al. 2012). Thus, it is necessary to improve our understanding of the processes involved in the system’s deterioration to support its maintenance and guarantee adequate levels of performance (Torres et al. 2017).

Urban water management generates a particularly complex decision-making environment that is subject to engineering, social, environmental, and economic constraints (Makropoulos et al. 2003). The latter is particularly marked when planning sewer maintenance operations: as a consequence of the subterraneous nature of the system and due to time and cost constraints, it is normally unfeasible to collect comprehensive information about the state of its components (Fenner 2000; Yang & Su 2007). Although the amount of data relevant to sewer asset management and its variety is expected to increase in the near future (e.g. new low-cost sensors, growing use of the Internet of Things, and the potential of social media platforms to report different types of problems) (Sirkia et al. 2017), the lack of knowledge of the current state of the system is the reason why most water utilities carry out maintenance operations with a corrective approach (Abraham et al. 1998).

In the case of corrective maintenance, the system is exposed to overflows and flooding until the maintenance operation is effectively performed (Fontecha et al. 2016), thus generating cost overruns with respect to preventive maintenance. In particular, such cost overruns include, but are not limited to: costs for service disruptions, adverse publicity costs, and costs associated with public health and safety problems (Korving et al. 2006; Rodríguez et al. 2012; Hickcox 2013). Thus, the prediction of blockages plays an important role in the management of the infrastructure of urban water systems. Deterioration models of sewer systems can be classified into two main groups: (a) physically based, and (b) statistical models. Physically based models are developed based on the understanding of the physical mechanisms that govern sewer system deterioration. However, scarcity of data about local conditions leads to great uncertainty about the suitability of these models (Fenner et al. 2007). Consequently, most deterioration models are of the statistical kind based on observations of failure events (Rodríguez et al. 2012).

Much of the research on sewer blockages has either focused on predicting mean blockage rates and on examining how these failure rates change over time (Savic et al. 2006; Ugarelli et al. 2009; Jin & Mukherjee 2010; Rodríguez et al. 2012; Post et al. 2016; Santos et al. 2017; Xie et al. 2017), or on representing the spatial distribution of sewer failures and/or flood hazards (Caradot et al. 2011; Cherqui et al. 2015; Post et al. 2017). However, blockages come in clusters, perhaps due to weather conditions, or maybe simply because releasing sediment from one part of the system leads to a knock-on blockage in another structural unit (Rodríguez et al. 2012). Despite the fact that sediment-related blockages in sewer systems occur in a spatiotemporal context, to the best knowledge of the authors of this paper, there is no previous research on modeling the intensity of these failures and relative risk as continuous phenomena in both time and space, considering that there may be a statistical residual component of spatiotemporal variation in these response surfaces that will only be captured by including in the model one or more latent stochastic process (Taylor et al. 2015).

This paper proposes a modeling approach to assess the impact of the spatiotemporal correlation between sediment-related blockages considering those factors that are associated with spatiotemporal clustering of failures, which may be physical properties of the sewer system or characteristics of the location where the failure event occurs. The blockage data is modeled as spatiotemporal point process by means of a spatiotemporal Log-Gaussian Cox Process (LGCP). The principal advantage is that LGCP provide a flexible and relatively tractable class of empirical models for describing spatially and temporally correlated phenomena (Diggle et al. 2015).

MATERIAL AND METHODS

Log-Gaussian Cox Process

The LGCP is a class of statistical models for spatial and spatiotemporal point processes. Contrary to other point process models, the LGCP does not rely on a random contagious mechanism to create clusters and spatial correlations. Instead, it uses a stochastic intensity function to capture the aggregation of events driven by environmental conditions. The resulting process is extremely flexible as it enables the presence of both fixed and random effects in the space-time behavior (Taylor et al. 2015). It can be stated as a generalization of the Poisson process, but allows more realistic correlations between non-overlapping spatial/temporal regions and the incorporation of covariates effects.
To define the LGCP process, let \( N(A, T) \) be the number of events located inside a region \( A \subseteq S \) and in the interval time \( T^0 = [t_1, t_2] \subseteq T \), where \( S \) and \( T \) are the region and horizon of study, respectively. Then, conditional on the intensity function \( R(s, t) \), this random variable follows a standard Poisson process, that is,

\[
N(A, T^0) | R(s, t) = r(s, t) \sim \text{Poisson} \left[ \int_{A,T} r(s, t) dsdt \right] \tag{1}
\]

where the intensity \( R(s, t) \) is considered a positive stochastic function on \( S \times T \) (Cox process), and \( r(s, t) \) is a particular realization of this function. The LGCP models \( R(s, t) \) as the logarithm of a Gaussian process which, allowing the incorporation of spatiotemporal covariates, results in an intensity function that can be expressed as (Møller et al. 1998):

\[
R(s, t) = \lambda(s, t) \exp \{Z(s, t)\beta + Y(s, t)\} \tag{2}
\]

where \( \lambda(s, t) \) is a known offset that represents some measure of exposure, \( Z(s, t) \) are the covariates (explicitly measured in space and/or time); and \( Y(s, t) \) is the latent Gaussian process on \( S \times T \). Note that \( Y(s, t) \) is the component that makes the intensity a non-deterministic function. In the stationary case, the process \( Y \) is a second-order process with constant mean and parameterized spatiotemporal covariance function, which is assumed to be separable; that is,

\[
\text{cov}[Y(s1, t1), Y(s2, t2)] = \sigma^2 \exp(-\|s2 - s1\|/\phi - |t2 - t1|/\theta) \tag{3}
\]

The parameter \( \sigma \) represents the variance and \( \varnothing \) is the scale parameter of spatial dependence, which determines the strength of the spatial correlation. \( \theta \) is the scale of temporal dependence, which determines the temporal correlation in \( Y \). For convenience, we use the standard approach of parameterizing \( E[Y] = -\sigma^2/2 \) (Diggle et al. 2013; Taylor et al. 2013, 2015), that makes \( E[exp(Y)] = 1 \). This parameterization allows for \( \exp(Y) \) to be interpreted as covariate-adjusted relative risk.

The parameters in the model are \( \beta \) (the vector of covariate effects), and \( \eta = \{\log(\sigma), \log(\varnothing), \log(\theta)\} \), the parameters of the spatiotemporal Gaussian process \( Y \) transformed onto log-scale (Taylor et al. 2015).

For computational estimation, a discrete representation of space and time is necessary. The discretization of the space is performed considering squared cells and the period of reference is divided into intervals of the same size. Implicitly, it is assumed that the intensity is constant or presents little variation within each cell (voxel). In this study, the authors consider spatial cells according to the strength of the spatial correlation, whereas time intervals coincide with the months in order to match the measurement of some temporal covariates.

**Bayesian inference**

Given the observed data of the point process \( X = (x_1, \ldots, x_n) \), where \( x_i = (s_i, t_i) \), and the measurements of covariates \( Z(s, t) \); the inferential problem is to estimate the unknown parameters \( \beta \) and \( \eta \). There are three main methodologies for inference of LGCP in the literature (see e.g. Diggle et al. (2013)): moment based estimation, maximum likelihood, and Bayesian inference. The first one is considered a simplistic approach rarely used in applications. Both maximum likelihood and Bayesian estimation require the joint probability distribution of the data points, which considering the double stochastic nature of the model, is not easy to optimize. That is, the price of having a flexible probabilistic generation of the point process implies that the likelihood is derived as:

\[
\ell(\beta, \eta|X) = P(X|\beta, \eta) = \int_R P(X, R|\beta, \eta) dR \tag{4}
\]

where the integral on the right is made over the random intensity function that is, by definition, an infinite dimensional probability measure. This difficulty results in the use of numerical methods that project the continuous process on a discrete representation of time and space.

In this paper, Bayesian estimation is used given the complications of the maximum likelihood direct approach. Additionally, using a computational approximation of the posterior parametric distributions has several advantages, such as being able to incorporate the predictive probability model of new data points and forecasting the process during the periods that follow the observational window. In the Bayesian approach, \( Y, \eta \) and \( \beta \) are considered random variables, and therefore, assigning prior probability densities to \( \eta \) and \( \beta \) allows to make inferential statements using the posterior (also the predictive) distribution, expressed as:

\[
P(\beta, \eta|X) \propto \ell(\beta, \eta|X) g(\beta, \eta) \tag{5}
\]

where \( g(\beta, \eta) \) is the prior distribution and the symbol \( \alpha \) means proportionality up to a constant.
The complexity of the posterior distribution requires the use of numerical integration by computational methods. A sampling mechanism of the posterior through Markov chain Monte Carlo (MCMC) is used, in particular, selecting the Metropolis-adjusted Langevin algorithm (MALA), which mimics a Langevin diffusion process, and has been recommended by several authors for this type of spatiotemporal models (see, for example, Roberts & Tweedie (1996) and Brix & Diggle (2001, 2003)). Details about the computational implementation are presented in Appendix A (available with the online version of this paper).

**Predictive analysis**

The Bayesian estimation procedure with a double stochastic process allows the modeler to obtain a posteriori distributions of the whole intensity function given the observed data points. Therefore, it is possible to sample the behavior of \( Y(s, t) \) given the data \( X \), and from it construct posteriori distributions of different target parameters of interest. In particular, the predictive functionality is very useful for simulating a new set of data points, affected conditionally by parameters \( \beta \) and \( \eta \), which take into account the effects of covariates and spatiotemporal correlations and also the effect of the previous set of points. One of the main advantages of this method is the smooth transition from estimation to predictive analysis, on which the distribution of new \( m \) data points \( X^* = (x^*_1, \ldots, x^*_m) \) can be represented as:

\[
P(X|\beta, \eta, X) \approx P(X^*|\beta, \eta)P(\beta, \eta|X) = P(X^*|\beta, \eta)g(\beta, \eta)
\] (6)

This distribution allows a prediction (probabilistically) of where and when new cases will appear. The number of cases can also be derived directly from the conditional inhomogeneous Poisson (predictive) distribution. Forecasting can also be performed for future periods as a regular temporal process where, conditional to the observations until time \( T \), it is possible to use the temporal correlation structure to predict future events. The prediction and forecasting features of the model will be used to perform validation with real data.

**Case study**

Bogotá is the capital of Colombia and the country’s largest city, with a population of approximately eight million inhabitants. The water utility of the city, Empresa de Acueducto y Alcantarillado de Bogotá (EAB), serves over 1.7 million residential users, approximately 400,000 commercial users, and provides services to 11 neighboring municipalities. In order to provide integral and efficient service to citizens, the water utility divided the city into five operative zones. Each zone is responsible for the operation and maintenance of water supply and sewage networks in a specific area of the city. For a detailed description of Bogotá’s drainage system see Rodríguez et al. (2008). In this research, the proposed statistical analyses methods are evaluated in the case study of sediment-related blockages in the sewer system of Zone 2, which drains an area of approximately 100 km² by means of both combined and separate sewer systems (see Figure 1).

**Customer complaints database**

The record of sewer failures was obtained from a platform managed by Bogotá’s water utility that collects and handles customer complaints. The customer complaints database available for testing the methods proposed in this research corresponds to a period of seven years and four months (from January 2004 to April 2011). For each reported failure, the following information is available: geographic coordinates where the failure was indicated, report date and the system component in which the failure was observed (i.e. pipe, gully pot, or manhole). The database contains 32,097 reports of sediment-related failures that occurred in the case study area during the period of analysis. For a detailed description of the customer complaints database see Rodríguez et al. (2012).

![Figure 1](image-url) | Case study area in the north of Bogotá (Colombia) (Zone 2 is highlighted) and an example of the spatial and temporal distribution of customer complaints.
Covariates

Covariates are the factors that are suspected to be associated with the occurrence of sediment-related failures. They include sewer system properties as well as environmental factors. Twenty-one covariates were available and considered for this study, of which 20 were of spatial type and the remaining were of temporal type (i.e. monthly precipitation). Models can also include other temporal explanatory variables, such as soil wetness, water consumption rates, and pipe structural condition, to simulate the time-dependence of hydraulic deterioration. However (both in general and in this case study), such data is not available. Covariates are described in Table 1. Note that categorical covariates (i.e. pipe material and land use) were transformed into continuous covariates.

Covariates data was handled and analyzed using ArcGIS 10.2.2 (ESRI 2014) and R (R Core Team 2015). The effect of pipe material on sediment-related blockages was evaluated by comparing the two most common materials used in the sewer network (i.e. concrete and reinforced concrete for the stormwater collection system and concrete and stoneware for the sanitary and combined system). The impact of land use on sediment-related failures was also evaluated by comparing the two most common uses in the case study area (i.e. residential and commercial). Monthly precipitation data was provided by Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Because there was available information from only three rainfall stations within the case study area, this feature worked as a temporal covariate.

RESULTS AND DISCUSSION

Model implementation

The dataset that was used in the construction of the model comprises reported failures from January 2004 to February 2011. According to the procedure for Bayesian Modeling of LGCP proposed by Taylor et al. (2015), approximate estimates of the parameters $\eta$ were computed using the minimum contrast method, which resulted in initial values for the MCMC routine of $\sigma = \log(\sqrt{1.7})$, $\phi = \log(283)$ and

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stormwater pipe length</td>
<td>Stormwater pipe length in the cell</td>
</tr>
<tr>
<td>Sanitary and combined pipe length</td>
<td>Sanitary and combined pipe length in the cell</td>
</tr>
<tr>
<td>Stormwater pipe diameter</td>
<td>Mean stormwater pipe diameter in the cell</td>
</tr>
<tr>
<td>Sanitary and combined pipe diameter</td>
<td>Sanitary and combined pipe diameter in the cell</td>
</tr>
<tr>
<td>Stormwater pipe depth</td>
<td>Mean stormwater pipe depth in the cell</td>
</tr>
<tr>
<td>Sanitary and combined pipe depth</td>
<td>Mean sanitary and combined pipe depth in the cell</td>
</tr>
<tr>
<td>Stormwater pipe slope</td>
<td>Mean stormwater pipe slope in the cell</td>
</tr>
<tr>
<td>Sanitary and combined pipe slope</td>
<td>Mean sanitary and combined pipe slope in the cell</td>
</tr>
<tr>
<td>Stormwater pipe age</td>
<td>Mean stormwater pipe age in the cell</td>
</tr>
<tr>
<td>Sanitary and combined pipe age</td>
<td>Mean sanitary and combined pipe age in the cell</td>
</tr>
<tr>
<td>Stormwater concrete pipe</td>
<td>Proportion of stormwater pipe length of the cell made from concrete</td>
</tr>
<tr>
<td>Stormwater reinforced concrete pipe</td>
<td>Proportion of stormwater pipe length of the cell made from reinforced concrete</td>
</tr>
<tr>
<td>Sanitary and combined concrete pipe</td>
<td>Proportion of sanitary and combined pipe length of the cell made from concrete</td>
</tr>
<tr>
<td>Sanitary and combined stoneware pipe</td>
<td>Proportion of sanitary and combined pipe length of the cell made from stoneware</td>
</tr>
<tr>
<td>Number of trees</td>
<td>Number of trees in a 1 km buffer centered in the cell centroid</td>
</tr>
<tr>
<td>Number of gully pots</td>
<td>Number of gully pots in a 1 km buffer centered in the cell centroid</td>
</tr>
<tr>
<td>Number of manholes</td>
<td>Number of manholes in a 1 km buffer centered in the cell centroid</td>
</tr>
<tr>
<td>Commercial – services</td>
<td>Proportion of the cell area assigned to commercial-services use</td>
</tr>
<tr>
<td>Residential</td>
<td>Proportion of the cell area assigned to residential use</td>
</tr>
<tr>
<td>Population</td>
<td>Population density in the cell</td>
</tr>
<tr>
<td>Monthly precipitation</td>
<td>Average monthly precipitation (in mm)</td>
</tr>
</tbody>
</table>
\( \theta = \log(0.5) \). The initial values for \( \beta \) were not defined; by default, LGCP initialized the MCMC using the estimates obtained from an over dispersed Poisson generalized linear model fitted to the cell counts against the covariates, ignoring spatial correlation. Considering that \( \phi = 283 \) m was the approximate distance beyond which the failure events were no longer spatially correlated, it was necessary to select a cell width that captured these dependence properties and also generated an efficient computational grid size. Thus, this value was set as equal to 400 m. Under the assumption that the estimated temporal correlation parameter was approximately 0.55, the temporal correlation in the \( Y \) process was expected to drop to 0.04 after six months (\( \exp(-0.55 \times 6) = 0.04 \)). Therefore, an observation window width of six months was used, comprising February 2011 and the preceding five months of data.

Multivariate Gaussian prior densities for the parameters \( \beta \) and multivariate Gaussian prior densities on the log-scale for the parameters \( \eta \) were defined as follows: \( \beta \sim N(0, 10^{6}) \) (it was a diffuse prior), \( \log \sigma \sim N(0, 1.2) \), \( \log \phi \sim N(0, 700), 0.05 \) and \( \log \theta \sim N(0, 0.2), 0.1 \). Additionally, it was specified that the spatial dependence properties of the latent field \( Y \) should follow an exponential covariance function.

An offset was not defined in this study. By default, LGCP package assigns the value of 1 to \( \lambda(s,t) \) if the cell of the computational grid falls within the case study area, and 0 otherwise. The extension parameter of the computational grid in the \( x \) and \( y \) directions was set as equal to 2. The target acceptance probability of the MCMC algorithm was set to 0.574, which would be approximately optimal for a Gaussian target explored by MALA proposal as the dimension of the problem tends to infinity (Taylor et al. 2013, 2015).

The MALA algorithm was run for 2,000,000 iterations, with an initial burn-in of 200,000 iterations, followed by a further 1,800,000 iterations, of which every 3,600th sample was retained so as to give a sample size of 500. For the base model (i.e., model with all covariates), the total computation time was around 4.47 days on a 2.4 Ghz Intel Xeon E5-2695 v2 PC with 64 Gb RAM. For the final model (i.e., model with significant covariates), it was approximately 3.772 days on the same PC.

Convergence and mixing diagnostics of the final model are presented in Appendix B (available with the online version of this paper). These analyses indicate that the chain appeared to have reached stationarity with low autocorrelation. Figure 2 compares the prior (represented by the continuous curve) and posterior (represented by the histogram) distributions of parameters \( \eta \) and \( \beta \) of the spatiotemporal LGCP model. The plots suggest that the covariate effects \( \beta \) are well identified by the data, and with respect to \( \eta \), it is observed that whilst there was information in the data on parameter \( \sigma \), the parameters \( \phi \) and \( \theta \) were not so well identified.

**Significant covariates**

The model was first run with all 21 covariates. The decision on which covariates were to play a role in the analysis was based on their statistical significance and the redundant information (multicollinearity) among groups of variables. Of the 21 covariates considered, four were selected for the final model. These were: population density, monthly precipitation, sanitary and combined sewer pipe length, and stormwater pipe diameter. Thus, the final model was run with these four covariates. Table 2 summarizes the parameters estimates from this model (i.e. median, 0.025 and 0.975 quantiles from the posterior density that correspond to the credible intervals).

These results show that the estimated number of sediment-related failures was higher in places with higher population density and, therefore, in places with higher domestic wastewater discharges. In addition, the estimated intensity was higher in those months in which the precipitation levels were greater, probably due to increased runoff sediment loads into the sewer system. Finally, an increase in the sanitary and combined pipe length and a decrease in the stormwater pipe diameter led to an increase in expected number of blockages which are both expected results.

**Intensity and relative risk**

Figure 3 presents the resulting maps of the expected number of sediment-related blockage events for the observation window comprising September 2010 to February 2011. The intensity of failures shows a clear spatial and temporal variation. The highest values are concentrated mainly in the center and northwest of the case study area, coinciding with the most populated zones.

Figure 4 shows the predicted covariate-adjusted relative risk surfaces derived from the model for all six months of the observation window. That is, after removing the effect of the significant variables, there is a relative risk estimation that explains the spatial and temporal expected intensity. These relative intensities may be the effect of different variables, or a product of the natural correlation of the process. These plots reveal some small clusters of raised risk, mainly in the center, northwest, and east of the studied area, which were probably formed as a consequence of the spatiotemporal correlation.
In this study, Poisson intensity was forecasted for March 2011 and April 2011. The resulting maps are shown in Figure 5. For each month, we measured the spatial forecasting performance of the size of the error using the mean square error (MSE), which was calculated on a monthly basis as the average of the sum of the squares of errors:

\[ \text{MSE}_t = \frac{\sum_{i \in I} (Y_{it} - F_{it})^2}{n} \]  

(7)

where \( I \) is the set of cells, \( Y_{it} \) is the actual number of sediment-related failures in cell \( i \) occurred during month \( t \), \( F_{it} \) is the forecasted number of failures in cell \( i \) during month \( t \), and \( n \) is the number of cells.

Given that the actual total number of failures occurred was not high enough (i.e. 396 for March 2011 and 437 for April 2011), a large number of cells of 400 × 400 m² were not subject to any blockage events, which could affect the results of the performance measure. As a consequence, bigger cells of 800 × 800 m² comprising four cells of the original grid were created, each of which was assigned values of \( Y_{it} \) and \( F_{it} \) equivalent to the sum of the values obtained by the corresponding four cells. Under this scheme, the total number of cells was equal to 1024.

The forecasted total number of failures was of 317 for March 2011 and 419 for April 2011. The corresponding
MSE values were 0.891 and 0.837, respectively. For comparison, on the same grid and with the same dataset, a homogeneous Poisson process was also fitted independently in each cell. This baseline model produced MSE values of 0.894 and 0.863 for March and April, respectively, proving that the LGCP can give more
In addition to the MSE measure, the area under the Receiver Operating Characteristic (ROC) curve (AUC) measure was also employed, considering the variable related to:

Figure 4 | Plots of the covariate-adjusted relative risk. From top left to bottom right: September 2010, October 2010, November 2010, December 2010, January 2011 and February 2011, respectively.

accurate predictions given its ability to model the complete stochastic process and explain the covariables effect.
to the probability that the number of failures in a cell of 800 × 800 m² is greater than or equal to 1. The AUC values were 0.938 for March 2011 and 0.940 for April 2011 (Figure 6). The Poisson process, per cell, gives almost the same AUC performance.

Finally, confusion matrices were generated for both months, assuming a threshold value of 0.5 (see Table 3). These matrices present the contrast of the predicted classes (failure or no-failure) versus the real classes using both testing datasets (March and April of 2011). On the diagonal, the numbers of cells with correct prediction are presented. Sensitivity values (percentage of cells with failures well predicted) were 0.729 for March 2011 and 0.732 for April 2011. Specificity values (percentage of cells with no-failures well predicted) were 0.942 and 0.938 for both months, respectively. The lower sensitivity of the model is due to the fact that most of 1024 cells of 800 × 800 m² did not present any failure events during March and April 2011.

Results of the predictive performance, by MSE, AUC and metrics of the confusion matrix, indicate that the proposed model has a very good capacity to anticipate the spatial distribution of failures in the sewer system. At the scale we defined the grid, this indicates that it is possible to plan ahead maintenance operations or rehabilitation strategies. It is important to take into account that the forecasting of several months in the future could be much less precise. However, the model can be estimated periodically in order to have better predictions.

Figure 5 | Plots of the forecasted Poisson intensity of failures for two months: March 2011 (left) and April 2011 (right).

Figure 6 | ROC curves of the forecasted Poisson intensity of failures for two months: March 2011 (left) and April 2011 (right).
CONCLUSIONS

It is known that sediment-related blockages in sewer systems can be considered a spatiotemporal failure process (e.g. Rodríguez et al. 2012). However, there is a lack of studies assessing appropriate methods to model such processes simultaneously in time and space. In this work, we applied a LGCP model to fill this knowledge gap. The LGCP model uses a stochastic intensity function to capture the aggregation of events driven by diverse physical and environmental conditions. Population density, monthly precipitation, sanitary and combined sewer pipes length and stormwater pipes diameter were identified as significant physical and environmental covariates in our case study. The LGCP model led to more accurate predictions than the homogenous Poisson model, given its ability to model the complete stochastic process and explain the covariates effect. Due to data availability restrictions, it was not possible to include other potential temporal explanatory variables (e.g. soil wetness, water consumption rates, and pipe structural condition, among others); thus, further work is required using case studies with more comprehensive datasets.

REFERENCES


Post, J., Langeveld, J. & Clemens, F. 2015 Analysing spatial patterns in lateral house connection blockages to support...


First received 6 February 2019; accepted in revised form 7 May 2019. Available online 15 May 2019