Multivariate probability distribution for sewer system vulnerability assessment under data-limited conditions

G. Del Giudice, R. Padulano and D. Siciliano

ABSTRACT

The lack of geometrical and hydraulic information about sewer networks often excludes the adoption of in-deep modeling tools to obtain prioritization strategies for funds management. The present paper describes a novel statistical procedure for defining the prioritization scheme for preventive maintenance strategies based on a small sample of failure data collected by the Sewer Office of the Municipality of Naples (IT). Novelty issues involve, among others, considering sewer parameters as continuous statistical variables and accounting for their interdependences. After a statistical analysis of maintenance interventions, the most important available factors affecting the process are selected and their mutual correlations identified. Then, after a Box-Cox transformation of the original variables, a methodology is provided for the evaluation of a vulnerability map of the sewer network by adopting a joint multivariate normal distribution with different parameter sets. The goodness-of-fit is eventually tested for each distribution by means of a multivariate plotting position. The developed methodology is expected to assist municipal engineers in identifying critical sewers, prioritizing sewer inspections in order to fulfill rehabilitation requirements.

Key words | asset management, geographical information system, maintenance, multivariate distribution model, pipe failure, prioritization strategies, sewer system, statistical analysis

G. Del Giudice
R. Padulano (corresponding author)
D. Siciliano
DICEA,
Università degli Studi di Napoli Federico II,
80125 Napoli,
Italy
E-mail: *roberta.padulano@unina.it*

NOTATION

AD	asset database
AIC	Akaike index
a	pipe age (-)
B	pipe width (m)
d	cover depth (m)
de	equivalent diameter (m)
f	theoretical pdf (-)
H	pipe height (m)
ID	incident database
L	log-likelihood function
l	pipe length (m)
M	model label
MSE	mean square error
m^2	Mahalanobis distance
N	number of failure events (-)
p	number of variables (-)
SSE	square sum of errors
sh	shape factor (-)
sl	slope (-)
x	statistical variable
α	Chi-squared probability
λ	Box-Cox transformation parameter
μ	mean
$\boldsymbol{\Sigma}$	variance-covariance matrix
σ	standard deviation

The following subscripts and superscripts are used in this paper:

T	transformed variable;
i	statistical variable, ranging between 1 and p
j	empirical observation, ranging between 1 and N
k	empirical observation, ranging between 1 and N

doi: 10.2166/wst.2015.546

INTRODUCTION

A sewer network typically reproduces the street planimetry, resulting in a complex and widespread infrastructure which needs specific maintenance planning and scheduling. Because of its extreme diffusion in urban areas (Ariaratnam *et al.* 2001), failures in a sewer network can cause considerable damage, such as traffic disruption, sinkholes, back-ups, spills and flooding, and corruption of the nearest water bodies. All considered, the in-charge authorities must develop managing strategies which can guarantee an efficient service to population.

The field of asset management has overcome a paradigm shift from an *a posteriori* approach, responding to failures with rehabilitation and replacement projects, to an *a priori* approach, predicting failures before they occur and mitigating the risk through risk assessment and preventive maintenance strategies (Allbee & Byrneb 2009). A successful asset management program should provide prioritization strategies for funds management by adopting predictive tools to anticipate sewers failure and to assess the risks associated with such failures (Fenner 2000; WRC 2001).

A large number of papers deal with the quantification of the deteriorating process of infrastructures, especially involving wastewater; such deterioration models must provide a predictive tool to assess the failure probability of sewers at any given time. However, sewer failures result from a complex process that is not only time-dependent but it is affected by other parameters whose influence on failure probability must be carefully determined (Davies *et al.* 2001a; Baur & Herz 2002; Hahn *et al.* 2002). Such factors can be roughly separated in two groups relating to structural and hydraulic deterioration, respectively (Davies *et al.* 2001a; Wirahadikusumah *et al.* 2001; Del Giudice & Farina 2007). The structural deterioration involves the weakening of pipe structural integrity resulting in an eventual collapse, whereas hydraulic deterioration refers to the reduced ability of the sewer to transport sewage resulting in surcharges, spills, or flooding. Various sewer deterioration models have been used in the literature to assess the condition of sanitary and storm sewers; examples are statistical, deterministic and artificial intelligence models (Kleiner & Rajani 2001; Savic *et al.* 2006; Tran *et al.* 2006; Berardi *et al.* 2008; Yamijala *et al.* 2009; Khan *et al.* 2010). Most of them imply records of pipe failures over a number of years for the prediction of pipe deterioration due to the aging process and consequent failure rates; however, time-dependent data are difficult to achieve, so that they often prove unavailable for sewer networks (Egger *et al.* 2013). Furthermore, the wastewater infrastructure building process usually spans over the centuries with limited or no data available; another problem is the absence of a standardized method for the description of sewer conditions apart from those involving expensive CCTV inspections, regulated by international norms such as EN 13508-2 (CEN 2003), so that in many countries there is a general lack of systematic information about sewer networks (Fenner 2000). In turn, the lack of data contributes to the lack of available modeling tools to predict failure patterns to assess the risks associated with the physical damage and the consequent disruption of service.

Because of the above-mentioned problems, failure models are often requested which can account for easily collectable physical data concerning the asset and generic information about historical records of pipe failure events. The present paper provides an analysis of Naples combined sewer network (Italy): for this system both an asset database and a failure database are available for the development of a model aiming to locate the sewer branches prone to failure by means of a statistical analysis of failure records. The proposed model considers sewer information in the failure dataset as a set of statistical variables. As a novelty element, the proposed analysis is possible when a correlation between variables is present, which means the hypothesis of stochastic independence is violated. Moreover, the parameters will be treated as continuous variables, whereas several statistical models can be found in the literature that require a division in classes, so that sewer parameters must be considered as discrete variables (Lei & Saegrov 1998; Caruso *et al.* 2002; Savic *et al.* 2006; Wright *et al.* 2006).

DATABASE OVERVIEW AND PREPROCESSING

Asset database

The asset database (AD), made up of about 50,000 records each corresponding to a sewer segment, shows an overview of the combined sewer system of Naples; for each record, information about pipe shape, width B in [m], height H in [m] and length l in [m] is available; also, invert and surface level above a fixed datum for both the ends of each pipe are provided. No other information was inferable from the available data: thus, the proposed analysis only relies on basic physical information, inevitably neglecting parameters, such as pipe material or ground conditions, which have a key role in the pipe deteriorating process (Davies *et al.* 2001a). However, the database needed a preprocessing operation in order to eliminate records where some information was missing or unreliable (Savic *et al.* 2006); this reduced the number of records to about 40,000. For the remaining pipes, some additional parameters were computed, namely an equivalent diameter *de*, in [m], and a shape parameter *sh*, defined as follows:

$$de = \sqrt{\frac{4BH}{\pi}}; \qquad sh = \frac{B}{H} \tag{1}$$

in order to numerically account for the different possible cross-section shapes, $sh = 1$ implying a circular or square cross-section. Also, slope *sl*, in [m/m], was computed as the ratio of elevation difference between the ends of the pipe to the pipe length; cover depth *d*, in [m], was computed as a difference between surface and invert level. Eventually, each record contains pipe id, equivalent diameter *de* in [m], cover depth *d* in [m], slope *sl* in [m/m] and shape factor *sh* (dimensionless); these parameters can be considered as the most common physical quantities related to blockage events. For instance, shape and dimensions have an influence on the transport capacity of the sewer mains. Slope also deeply influences water velocity: high slope implies high velocity and a subsequent abrasion of the sewer walls, whereas mild slope implies low velocity, with possible sedimentation of solid material and obstructions.

In order to obtain a more complete analysis, an additional sewer factor should be considered, namely sewer age, which is an important parameter in the deterioration process of a conduit (Davies *et al.* 2001a). Age is presumably related to the construction material used for sewers over time, shifting from stone to plastic; it also implies changes in the shape of sewer sections. No information about age is reported in the AD, so that a rough approximation is held by giving each sewer an age referred to the construction period of the corresponding urban area (Davies *et al.* 2001b; Ahmadi *et al.* 2014): in the city of Naples three different urban expansion periods can be found, corresponding to years <1900, 1900–1950 and >1950, respectively. For the sake of mathematical analysis, a polytomous age variable *a* was considered, equal to 1 when construction period was before 1900, equal to 2 when construction period was between 1900 and 1950, and equal to 3 when construction period was after 1950.

Histograms in Figure 1 show the frequency distribution of the above-mentioned sewer factors, namely, for each

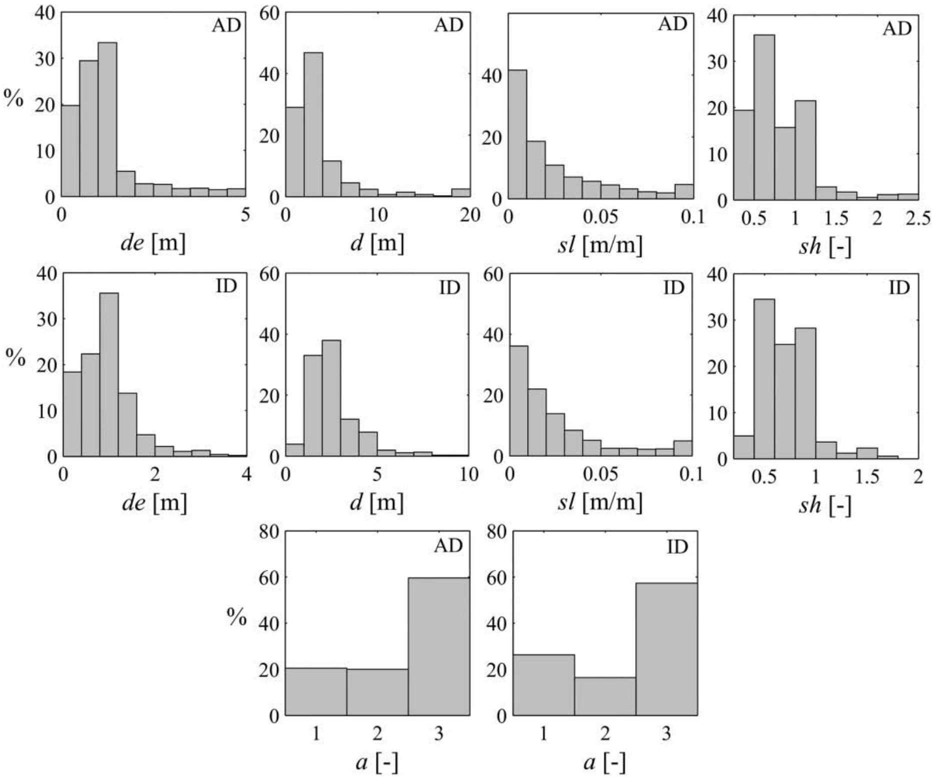

Figure 1 | Distribution of sewer parameters within AD and incident database (ID).

class, the total length of sewer trunks belonging to that class divided by the entire network length (about 1,150 km) and expressed as a percentage. For each parameter, the range of variation was sectioned in a different number of classes to give a realistic representation of frequencies.

Frequencies in Figure 1 show a deep asymmetry with respect to all the considered sewer parameters: about 90% of the whole network extension is characterized by diameters ranging from the minimum to 2 m, cover depths up to 6 m, slopes up to 0.02 m/m, and shape factors up to 1.5. As concerns age, 60% of sewers belong to the third class (>1950).

Incident database

Information concerning failure events was provided by the Sewer Office of the Municipality of Naples in the shape of an incident database (ID) covering years 2002–2011; the database contains, for each record (total 914 records), the event id, date and geographic point location of the failure; then, each point was associated with the corresponding pipe and its physical characteristics by means of a geographical information system (GIS) tool. Database only refers to ordinary maintenance operations due to blockage events: they will be referred to as 'failure events'.

Figure 1 shows the distribution of physical parameters within the ID. Histograms show that the highest failure frequency, specifically the ratio of the number of failure events for each class to the total number of failures, occurs for small diameters, small slopes, small depths and for high rectangular cross-sections, showing a deep asymmetry in the frequency distribution of events for each of the considered parameters. It should be noted that the pipes provided with these features also have the largest number of occurrences in the AD. Similarly, almost 60% of failure records belong to the most recent pipe class, but this is merely because class 3 is also the more diffusely spread in the city.

METHODOLOGY

Statistical distribution of failure frequency

It is possible to conceive the incident dataset as a failure statistical sample; for each parameter a theoretical pdf $f(x_i)$ can be conceived that best suits the observed one, with $i = 1,\dots p$ being p the number of considered parameters. However, this procedure is not enough to characterize failure frequency as there could be interactions among the parameters. Thus, each $f(x_i)$ can be considered as a marginal pdf for a multivariate distribution referring to the vector of variables $\bar{x} = [x_1, \dots x_p]$; if variables are independent, the multivariate pdf $f(\bar{x})$ can be computed as the product of marginal pdfs. If a correlation exists, the joint pdf specifies in a more complex expression which must take the variance-covariance matrix into account. To understand about possible correlations between pairs of variables it is more convenient to analyze the matrix of Pearson indices which gives dimensionless correlation estimates.

In order to obtain information concerning sewer system failure probability, a multinormal distribution function can be tested, since such a model can account for dependences among variables by means of the variance-covariance matrix. If original variables have a different marginal pdf each, it is possible that a normal model does not fit with their multivariate distribution. However, if single variables have normal marginal pdfs, the possibility that their multivariate pdf is normal increases. To facilitate this, the Box-Cox transformation (Box & Cox 1964) can be applied to each sewer parameter: this allows for representing the marginal pdfs as normal probability distributions, and the normality of each parameter distribution can be tested by means of a Q-Q plot. Further, the bivariate normal distributions for each possible pair of parameters are assumed and tested by plotting contour lines of the bivariate normal density functions. Finally, a joint p-variate normal distribution is adopted; a normality test can be conducted by means of a Chi-squared plot (Johnson & Wichern 2007).

Marginal probability distributions

The first step of the proposed methodology consists of the evaluation of the marginal pdfs of each physical parameter; this is done by applying the Box-Cox transformation (Box & Cox 1964). According to Box and Cox, every random variable can be interpreted with a normal distribution if a suitable transformation is applied to the original variable by means of a transformation coefficient λ that maximizes the log-likelihood function $L(\lambda)$ (Box & Cox 1964). Once λ has been computed, the normally distributed variables are

$$\begin{aligned} x_T &= \frac{x^\lambda - 1}{\lambda} \qquad \text{if} \quad \lambda \neq 0 \\ x_T &= \ln(x) \qquad \text{if} \quad \lambda = 0 \end{aligned} \tag{2}$$

The normality of transformed data can be tested for each variable by means of a Q-Q plot which compares the theoretical and the observed quantiles, involving the mean and standard deviation of transformed data; normality is confirmed if the fitting line of data resembles the 1:1 line.

Joint bivariate and *p*-variate normal distributions

Once the marginal pdfs are known for each of the considered parameters, the normal multivariate density distribution is tested for estimating the failure probability of the sewer network

$$f(x_{Tj1}, x_{Tj2}, \ldots, x_{Tjp}) = f(\bar{x}_{Tj}) = \frac{1}{(2\pi)^{p/2} \cdot |\Sigma_T|^{1/2}} \cdot \exp\left[-\frac{(\bar{x}_{Tj} - \bar{\mu}_T)^T {\Sigma_T}^{-1} (\bar{x}_{Tj} - \bar{\mu}_T)}{2}\right] \quad (3)$$

where $\bar{x}_{Tj}$ is the vector of transformed variables referring to observation j (with j varying between 1 and N), $\bar{\mu}_T$ is the mean vector of transformed variables, and Σ_T is the variance-covariance matrix referring to $\bar{x}_T$. For a univariate distribution, the argument of the exponential measures the square distance between each data and the mean value; this can be generalized for N multivariate observations on p variables obtaining the square Mahalanobis distance, or generalized distance, m^2 (Härdle & Simar 2007)

$$m^2(x_{Tj1}, x_{Tj2}, \ldots, x_{Tjp}) = m^2(\bar{x}_{Tj}) = m^2_{Tj} = (\bar{x}_{Tj} - \bar{\mu}_T)^T {\Sigma_T}^{-1} (\bar{x}_{Tj} - \bar{\mu}_T) \quad (4)$$

with all the symbols previously specified; if $m^2_{Tj} = 0$, which implies the vector of observations coincides with the mean vector, the p-variate normal density has its maximum value. For a multivariate normal distribution, m^2_{Tj} is a random variable with a Chi-squared distribution with p degrees of freedom (Johnson & Wichern 2007). Substituting Equation (4) into Equation (3), normal multivariate distribution can be expressed as

$$f(\bar{x}_{Tj}) = \frac{1}{(2\pi)^{p/2} \cdot |\Sigma_T|^{1/2}} \cdot \exp\left[-\frac{m^2_{Tj}}{2}\right] \quad (5)$$

Once the marginal distributions for each parameter are confirmed normal, an intermediate step consists of testing the normality of the bivariate distributions corresponding to each possible pair of transformed variables in the case study (Johnson & Wichern 2007). The test consists of the observation of data scatter plots; for each pair, normality is proved if about 50% of the data are included within the ellipse corresponding to the 50th percentile of a Chi-squared distribution with 2 degrees of freedom, and, simultaneously, about 95% of the data are included in the ellipse corresponding to the 95th percentile of the same distribution (Johnson & Wichern 2007). Note that for a Chi-squared distribution $m^2{}_{Tj}(\alpha = 0.50) = 1.39$ and $m^2{}_{Tj}(\alpha = 0.05) = 5.99$.

The final step of the procedure consists of verifying the hypothesis of a multivariate normal distribution accounting for all the p variables altogether; the test can be performed by evaluating the multivariate generalized distances by means of Equation (4) with a variance-covariance matrix including the whole set of available parameters; then, the computed distances are compared with the theoretical quantiles of a Chi-squared distribution with p degrees of freedom by means of a so-called Chi-squared plot (Johnson & Wichern 2007). The test is passed if data have a fitting line that resembles the 1:1 line; also, 50% data must have $m^2{}_{Tj} \le m^2{}_{Tj}(\alpha = 0.5)$ and 95% data must have $m^2{}_{Tj} \le m^2{}_{Tj}(\alpha = 0.05)$. Note that for a Chi-squared function with five degrees of freedom $m^2{}_{Tj}(\alpha = 0.5) = 4.35$ and $m^2{}_{Tj}(\alpha = 0.05) = 11.07$.

Adaptation to data and goodness-of-fit measures

Different multivariate distributions can be obtained by varying the set of involved variables. This can be done in order to investigate about the amount of information obtained by adding explanatory variables to the model, since this cannot be inferred by simply evaluating the goodness-of-fit of multivariate Chi-squared plots.

Given a number M of possible models, each differing for number of variables p or specific variables in the dataset, a comparison can be done between the computed joint cumulative distribution function (cdf) Φ_m given by model m (with $m = 1 \ldots M$) and an empirical joint cumulative frequency F_m. In this paper the adopted multivariate plotting position formula is the one proposed by Gringorten (1963)

$$F(\bar{x}_{Tk}) = \Pr\{\bar{x}_T \le \bar{x}_{Tk}\} = \Pr\{x_{T1j} \le x_{T1k} \text{ and } x_{T2j} \le x_{T2k} \text{ and } \ldots \text{ and } x_{Tpj} \le x_{Tpk}\} = \frac{\text{No of }\{x_{T1j} \le x_{T1k} \text{ and } x_{T2j} \le x_{T2k} \text{ and } \ldots \text{ and } x_{Tpj} \le x_{Tpk}\} - 0.44}{N + 0.12} \quad (6)$$

In Equation (6) N is the sample size, p is the number of variables in the model and $j, k = 1 \ldots N$. If the model perfectly predicts failure probability, scatter plot of pairs (F, Φ)

resembles the 1:1 line. However, given the complexity of sewer failure process, a satisfying adaptation can be assessed if the interpolating line is close to the 1:1 line, with residuals having zero mean and a suitably small standard deviation. An objective criterion to establish which is the best model among the tested ones is the Akaike criterion (Akaike 1974) *AIC*

$$AIC_m = N \cdot \log(MSE_m) + 2 \cdot p_m = N \cdot \log\left(\frac{SSE_m}{N - p_m}\right) + 2 \cdot p_m \tag{7}$$

$$SSE_m = \sum_{j=1}^{N} (F_{mj} - \Phi_{mj})^2 \tag{8}$$

where *MSE* is the mean square error, derived from the square sum of errors (*SSE*), and p_m is the number of variables in the model *m*. The model with best adaptation to data is the one having minimum *AIC*.

CASE STUDY AND DISCUSSION

To prove the robustness of the proposed methodology, calibration of the multivariate model was done by using a split sample technique. The original ID was randomly divided into a calibration sample ($N = 460$) and a validation sample ($N = 454$). The former was used to compute transformation coefficients λ, and all the normality tests were performed as discussed in previous sections. The latter was treated by using the same λ values, along with the mean vector and the variance-covariance matrix of the calibration sample, to perform all the previously described normality tests. Table 1 provides the mean vector and the Pearson matrix of the whole incident dataIDset. A deep correlation between diameter and cover depth can be found: this was expected since larger pipes require deeper excavation operations in order to be laid. Small dependences can also be found between slope, age and shape factor.

The Box-Cox transformation was applied to the five variables of the calibration sample, namely *de*, *d*, *sl*, *sh* and *a*, obtaining λ values shown in Table 1. Figure 2 shows the log-likelihood functions for λ ranging between −3 and 3. It can be noted that peak log-likelihoods are obtained for λ values very close to 0; this is not entirely true for age, but in this case the likelihood function is so flat that its peak *L* is very close to $L(\lambda = 0)$. As a results, marginal pdfs were considered lognormal for all the variables of interest.

Figure 3 shows Q-Q plots for transformed equivalent diameter de_T, transformed cover depth d_T, transformed slope sl_T and transformed shape factor sh_T; the Q-Q plot for transformed age is not shown because its graphical meaning is not significant, given that variable can assume only three values. For all variables normality is confirmed.

As concerns bivariate distributions, Table 2 shows percentages of data laying within 50th and 95th quantile ellipses. Values confirm the assumption of normal bivariate distributions for each pair of variables, except for the pair *de-a*. For this pair, the first percentage is considerably lower than 50%, whereas the second percentage is higher than 95%; this was not unexpected, since the concept of age district is a rough way of accounting for pipe age.

Once the bivariate normal distributions are confirmed, four multivariate models were assumed: the first (M1) considers all the available sewer parameters in the dataset, namely equivalent diameter, cover depth, slope, shape factor and age ($p = 5$); the second (M2) neglects age information assuming that the incident dataset consists of ordinary maintenance operations, which do not strictly depend on pipe deterioration conditions ($p = 4$); the third (M3) and fourth (M4) also neglect cover depth and diameter respectively, based on the fact that *de* and *d* are deeply correlated as shown in Table 1 ($p = 3$). Figure 4 shows Chi-squared plots for all the multivariate distributions; the

Table 1 | ID mean values, Box-Cox λ values and Pearson matrix (%)

Variables mean and λ values			**Pearson matrix**					
	μ	**λ**		**de**	**d**	**sl**	**sh**	**a**
de	0.957 (m)	0.171	*de*	100	45.8	−2.8	1.3	−9.2
d	2.566 (m)	−0.027	*d*	45.8	100	−1.4	−19.1	6.4
sl	0.027 (m/m)	0.112	*sl*	−2.8	−1.4	100	−23.2	−26.5
sh	0.743 (–)	−0.134	*sh*	1.3	−19.1	−23.2	100	27.5
a	2.311 (–)	1.838	*a*	−9.2	6.4	−26.5	27.5	100

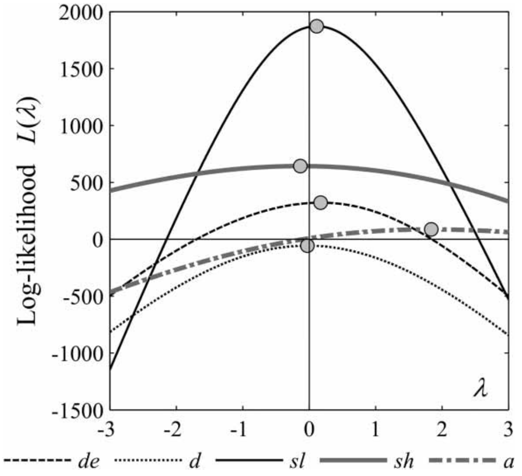

Figure 2 | Log-likelihood function for sewer parameters.

Chi-squared plots thoroughly confirm the hypothesis of multinormal joint distribution for all models, as the 1:1 line perfectly fits data, so that none of the four considered models can be rejected. Thus, for all of them a comparison between computed and observed cdf was evaluated by means of Equations (6)–(8).

Figure 5 shows this comparison for all four models, for both calibration and validation samples. The first important consideration is that failure prevision is falsified when age is considered within the parameters set, being the computed probability systematically smaller than the observed one (Figure 5(a)). This could either happen because age is not an explanatory variable for blockage events, which do not physically depend on pipes deteriorating process, or because the assumption of age district is too rough to represent reality; this also reflects on bivariate distributions in Table 2 involving age, whose adaptation to the normal model is the worst. Once age is removed, the remaining models have similar adaptation to data, being experimental points aligned around the 1:1 line with good fit. This is especially true for M2 and M3, but M3 provides a wider variation of residuals, as shown in both Figure 5 and Table 3. M4 provides the smallest residual standard deviation, but, as for M1 but in a slighter amount, probability prevision is distorted. Table 3 also shows *AIC* values for the models. The above-mentioned considerations about both prevision estimate and standard deviation of residuals are well summarized in *AIC*, since the smallest *AIC* matches with M2. Note that values in Table 3 are computed for the whole ID.

A second consideration is that removing information from the dataset, namely moving from left to right in Figure 6, results in an increasing number of pipes belonging to the critical class: this underlines the effect of adding variables as a filter-like behavior, which could facilitate allocating economic resources when a small budget is available. If model M1 is selected instead of M2, Figures 6(a)

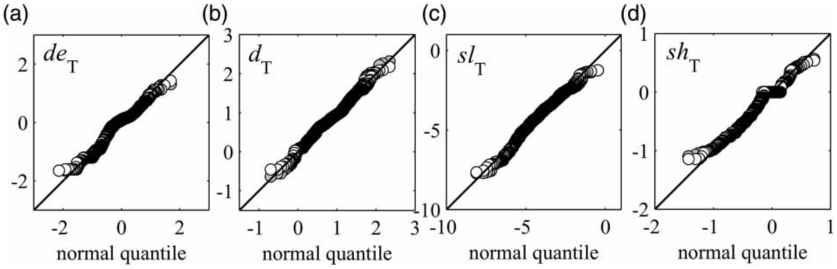

Figure 3 | Q-Q plot of sewer parameters with transformed variables for calibration (gray circles) and validation (white circles) samples.

Table 2 | Percentages of data laying within 50th percentile (above main diagonal) and 95th percentile (below main diagonal) for calibration and validation samples

Calibration sample (%)						**Validation sample (%)**					
	de	*d*	*sl*	*sh*	*a*		*de*	*d*	*sl*	*sh*	*a*
de	–	53.5	50.9	51.7	43.0	*de*	–	49.8	50.4	53.3	44.9
d	92.4	–	54.3	52.2	49.8	*d*	93.6	–	50.0	50.9	47.4
sl	93.9	94.3	–	48.3	52.6	*sl*	94.1	93.8	–	47.8	50.0
sh	94.6	95.2	96.1	–	53.0	*sh*	94.9	95.8	96.5	–	52.4
a	98.3	95.2	96.1	98.0	–	*a*	98.9	96.9	96.5	97.6	–

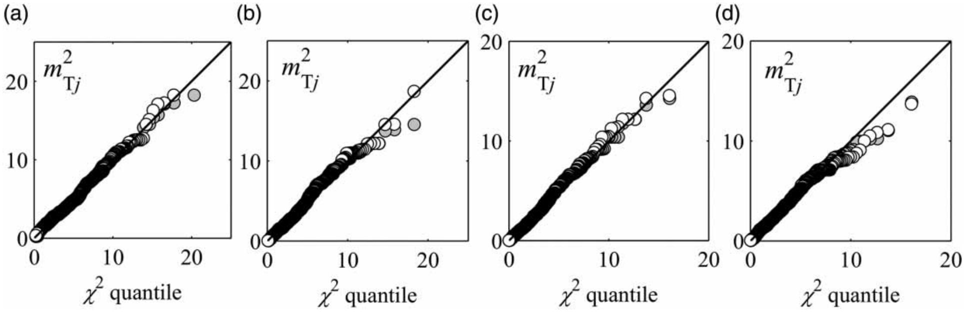

Figure 4 | Chi-squared plot with transformed variables for calibration (gray circles) and validation (white circles) sample for models M1 (a), M2 (b), M3 (c), and M4 (d).

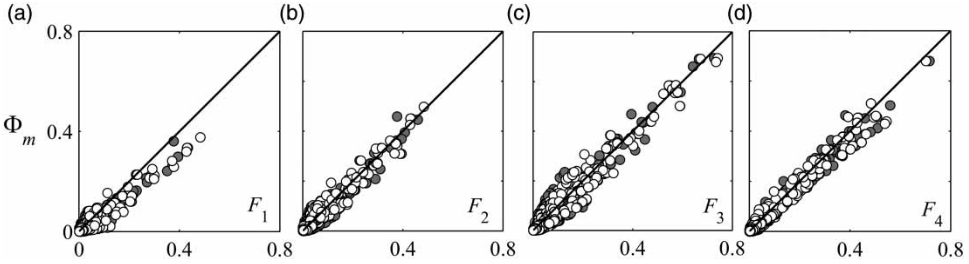

Figure 5 | Comparison between theoretical and observed cdf for calibration (gray circles) and validation (white circles) samples for models M1 (a), M2 (b), M3 (c), and M4 (d).

Table 3 | Goodness-of-fit measures

Model	*p*	$\mu_{\text{residuals}}$	$\sigma_{\text{residuals}}$	*SSE* × 10^2	*AIC*
M1	5	0.0126	0.0241	67.67	−2849
M2	4	0.0030	0.0202	38.19	−3079
M3	3	0.0035	0.0309	88.02	−2750
M4	3	0.0078	0.0210	45.60	−3011

and 6(b) show that the number of critical pipes is lower for M1 than M2; however, an in-deep fitting analysis as the one provided in Figure 5 shows that the advantage of having a highly selective model is thwarted by its unreliability when describing the sample: in other words, M1 considers non-critical pipes which actually are.

Figure 7 shows goodness of fit and failure frequency when a model M5 is considered that assumes variables *de*, *d*, *sl* and *sh* as uncorrelated. Figure 7(a) particularly shows how M5 is completely unable to describe failure frequency within the sample, especially for low probability values.

Figure 8 shows the results of model M2 applied to the Naples sewer system, with λ values, mean vector and variance-covariance matrix computed for the calibration sample. The use of a GIS software enables the evaluation of the most critical sewer pipes as those corresponding to the maximum values of $f(\bar{x}_T)$; density values were divided into five classes according to Jenks optimization method (Jenks 1967), the red one corresponding to the maximum f values. The map can be adopted to obtain a prioritization ranking of the sewer network, to locate the trunks which are particularly prone to failure, where prompt interventions are needed: a greater selectivity of critical sewer trunks can be obtained by increasing the number of classes.

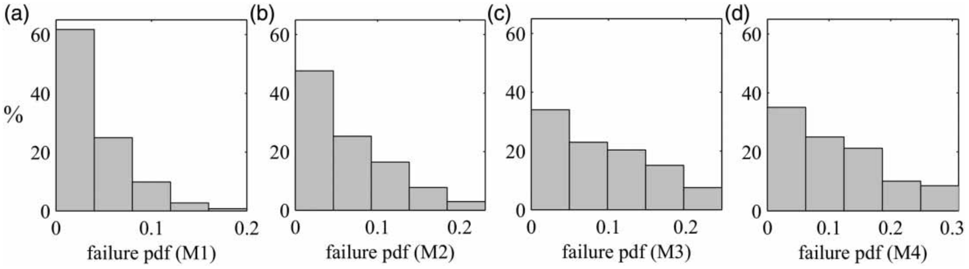

Figure 6 | ID failure frequency computed with models M1 (a), M2 (b), M3 (c), and M4 (d).

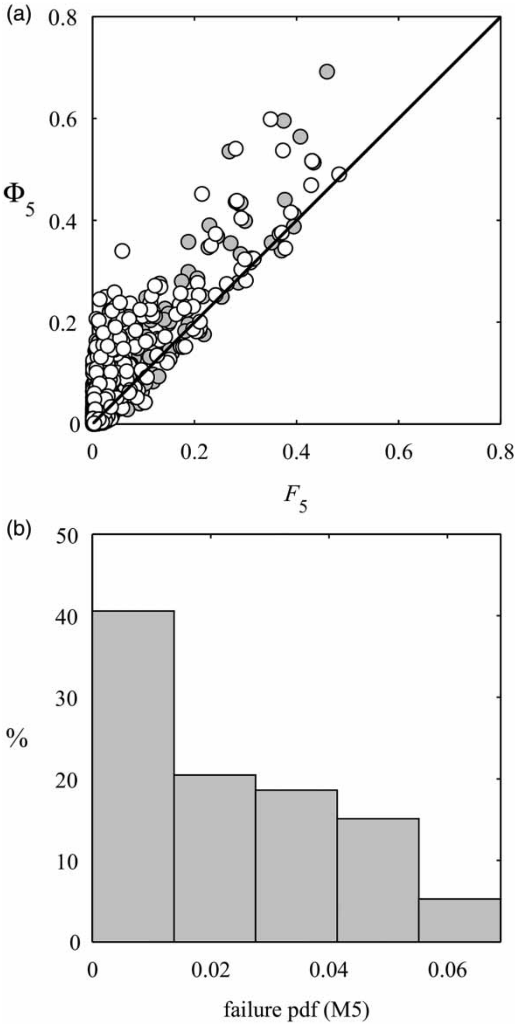

Figure 7 | Comparison between theoretical and observed cdf (a) and ID failure frequency (b) computed with model M5.

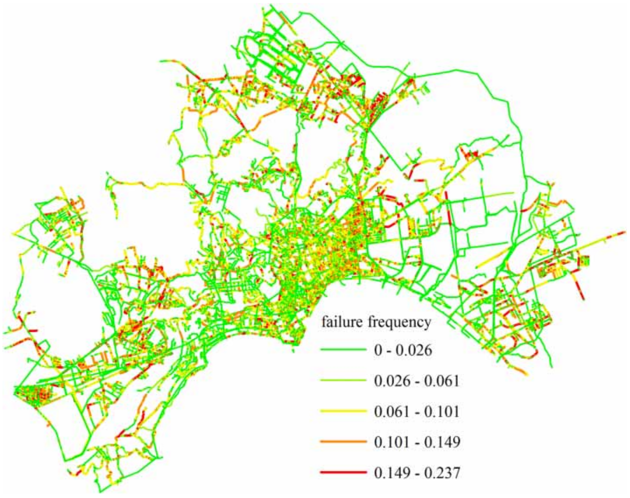

Figure 8 | Spatial distribution of failure probability within the city of Naples computed with model M2.

CONCLUSIONS

Basing on an AD and failure events records of the sewer system in the city of Naples (Italy), only containing basic physical information about sewer pipes, a statistical model is provided that allows for the estimation of failure probability and the location of critical sewers. The model can be successfully coupled to an intervention strategy aimed at optimizing the allocation of economic resources for the management of a sewer network. In the case study, marginal cdfs are lognormal, whereas the best fitting occurs when the joint failure model neglects age information.

Within the statistical framework, compared to literature regression models, the proposed procedure does not need a division of data in classes, so that computing is straightforward and each variable acts in the model with its own value; also, dependences among parameters can be taken into account by means of the variance-covariance matrix. Moreover, the model is flexible since it can be applied with any number p of variables, just requiring an increasing number of marginal and bivariate normal tests. Conversely, a considerable drawback is that the model is stationary in the sense that it does not account for variations in time; also, the ID covers a short number of years. Consequently, failure probability estimates should be considered as a short-term provision. Another marginal drawback is that the model does not automatically update sewer conditions by taking into account historical records about interventions: this fail is negligible for ordinary maintenance operations, but could invalidate failure probability estimates for extraordinary repairs, often entailing entirely substituting pipes. In this case, the replaced pipes should be removed from the database as a preliminary operation.

REFERENCES

Ahmadi, M., Cherqui, F., De Massiac, J. C. & Le Gauffre, P. 2014 Influence of available data on sewer inspection program efficiency. *Urban Water Journal* **11** (8), 641–656.

Akaike, H. 1974 A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** (6), 716–723.

Allbee, S. & Byrneb, R. 2009 Report 1: A global vision for driving infrastructure asset management improvement. In: *Strategic Asset Management of Water Supply and Wastewater Infrastructure: Invited Papers from the IWA Leading Edge Conference on Strategic Asset Management (LESAM)*, IWA Publishing, Lisbon, October 2007.

Ariaratnam, S. T., El-Assaly, A. & Yang, Y. 2001 Assessment of infrastructure inspection needs using logistic models. *Journal of Infrastructure Systems* **7** (4), 160–165.

Baur, R. & Herz, R. 2002 Selective inspection planning with ageing forecast for sewer types. *Water Science and Technology* **46** (6), 389–396.

Berardi, L., Giustolisi, O., Kapelan, Z. & Savic, D. 2008 Pipe deterioration models for water distribution systems. *Journal of Hydroinformatics* **10** (2), 113–126.

Box, G. E. P. & Cox, D. R. 1964 An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* **26** (2), 211–252.

Caruso, G., Salese, M. R., Del Giudice, G. & Rasulo, G. 2002 Operational method for the analysis of urban drainage systems. In: *Proceedings International Conference on Sewer Operation and Maintenance – SOM2002*. Bradford, UK, *November 26–28*.

CEN 2003 *Investigation and Assessment of Drains and Sewer Systems Outside Buildings. Part 2: Visual Inspection Coding System*. European Standard, European Committee for Standardization CEN, Brussels, Belgium.

Davies, J. P., Clarke, B. A., Whiter, J. T. & Cunningham, R. J. 2001a Factors influencing the structural deterioration and collapse of rigid sewer pipes. *Urban Water* **3** (1), 73–89.

Davies, J. P., Clarke, B. A., Whiter, J. T., Cunningham, R. J. & Leidi, A. 2001b The structural condition of rigid sewer pipes: a statistical investigation. *Urban Water* **3** (4), 277–286.

Del Giudice, G. & Farina, L. 2007 Fuzzy logic for prioritizing sewer line maintenance. In: *Proceedings of the XXXII IAHR Congress*, Venice, Italy, July 1–6.

Egger, C., Scheidegger, A., Reichert, P. & Maurer, M. 2013 Sewer deterioration modeling with condition data lacking historical records. *Water Research* **47** (17), 6762–6779.

Fenner, R. A. 2000 Approaches to sewer maintenance: a review. *Urban Water* **2** (4), 343–356.

Gringorten, I. I. 1963 A plotting rule of extreme probability paper. *Journal of Geophysical Research* **68** (3), 813–814.

Hahn, M. A., Palmer, R. N., Merrill, M. S. & Lukas, A. B. 2002 Expert system for prioritizing the inspection of sewers: knowledge base formulation and evaluation. *Journal of Water Resources Planning and Management* **128** (2), 121–129.

Härdle, W. K. & Simar, L. 2007 *Applied Multivariate Statistical Analysis*. 2nd edn, Springer, Berlin.

Jenks, G. F. 1967 The data model concept in statistical mapping. *International Yearbook of Cartography* **7**, 186–190.

Johnson, R. A. & Wichern, D. W. 2007 *Applied Multivariate Statistical Analysis*. 6th edn, Prentice Hall, Upper Saddle Rive, New Jersey, USA.

Khan, Z., Zayed, T. & Moselhi, O. 2010 Structural condition assessment of sewer pipelines. *Journal of Performance of Constructed Facilities* **24** (2), 170–179.

Kleiner, Y. & Rajani, B. 2001 Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water* **3** (3), 131–150.

Lei, J. & Saegrov, S. 1998 Statistical approach for describing failures and lifetimes of water mains. *Water Science and Technology* **38** (6), 209–217.

Savic, D., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S. & Saul, A. 2006 Modelling sewer failure by evolutionary computing. In: *Proceedings of the ICE – Water Management* **159** (2), 111–118.

Tran, D. H., Ng, A. W. M., Perera, B. J. C., Burn, S. & Davis, P. 2006 Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. *Urban Water Journal* **3** (3), 175–184.

Water Research Centre (WRC) 2001 *Sewerage Rehabilitation Manual*. 4th edn. Water Research Centre/Water Authorities Association, Swindon, UK.

Wirahadikusumah, R., Abraham, D. & Iseley, T. 2001 Challenging issues in modeling deterioration of combined sewers' *Journal of Infrastructure Systems* **7** (2), 77–84.

Wright, L. T., Heaney, J. P. & Dent, S. 2006 Prioritizing sanitary sewers for rehabilitation using least-cost classifiers. *Journal of Infrastructures Systems* **12** (3), 174–183.

Yamijala, S., Guikema, S. D. & Brumbelow, K. 2009 Statistical models for the analysis of water distribution system pipe break data. *Reliability Engineering and System Safety* **94**, 282–293.

First received 10 August 2015; accepted in revised form 8 October 2015. Available online 26 October 2015