Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in Berlin, Germany

N. Caradot, M. Riechel, M. Fesneau, N. Hernandez, A. Torres, H. Sonnenberg, E. Eckert, N. Lengemann, J. Waschnewski and P. Rouault

ABSTRACT

Deterioration models can be successfully deployed only if decision-makers trust the modelling outcomes and are aware of model uncertainties. Our study aims to address this issue by developing a set of clearly understandable metrics to assess the performance of sewer deterioration models from an end-user perspective. The developed metrics are used to benchmark the performance of a statistical model, namely, GompitZ based on survival analysis and Markov-chains, and a machine learning model, namely, Random Forest, an ensemble learning method based on decision trees. The models have been trained with the extensive CCTV dataset of the sewer network of Berlin, Germany (115,258 inspections). At network level, both models give satisfactory outcomes with deviations between predicted and inspected condition distributions below 5%. At pipe level, the statistical model does not perform better than a simple random model, which attributes randomly a condition class to each inspected pipe, whereas the machine learning model provides satisfying performance. 66.7% of the pipes inspected in bad condition have been predicted correctly. The machine learning approach shows a strong potential for supporting operators in the identification of pipes in critical condition for inspection programs whereas the statistical approach is more adapted to support strategic rehabilitation planning.

Key words | asset management, CCTV, machine learning, Random Forest, sewer, survival analysis

N. Caradot (corresponding author)
M. Riechel
M. Fesneau
H. Sonnenberg
P. Rouault
Kompetenzzentrum Wasser Berlin,
Cicerostr. 24, Berlin,
Germany
E-mail: *nicolas.caradot@kompetenz-wasser.de*

N. Hernandez
A. Torres
Pontificia Universidad Javeriana, Faculty of Engineering,
Bogotá, DC,
Colombia

E. Eckert
N. Lengemann
J. Waschnewski
Berliner Wasser Betriebe,
Neue Jüdenstraße, 10179 Berlin,
Germany

INTRODUCTION

Context

Insufficient public and municipal investment represents a major challenge for the sustainable management of sewer networks. More than a decade ago, the American Water Works Association estimated that a new era was dawning: the replacement era in which the country will need to rehabilitate massively the water and sewer networks built by the previous generations (AWWA 2012). In many cities worldwide, the buried infrastructure is nearing the end of its useful life and will soon reach the age of renewal. The need for maintaining and expanding the United States' water and sewer network is estimated to approximately $126 billion in 2020 (ASCE 2011). In Germany, a recent national study highlighted that 20% of the sewer network have severe defects that require short- or mid-term rehabilitation (Berger *et al.* 2015). Over the last years, the annual investment for sewer rehabilitation has been about four billion € whereas capital need is estimated to be more than seven billion €, indicating a capital deficit of at least three billion € (IPK 2014; KfW 2016). Delaying further the investment will result in degrading water and sewer services,

doi: 10.2166/hydro.2018.217

escalating flood risk, increasing environmental impacts and raising expenditures for emergency repairs.

The funding gap and the investments needed to cope with sewer deterioration will necessarily lead to future increases of the water tariff (Oelmann *et al.* 2017). Utilities are challenged to develop efficient rehabilitation strategies in order to keep the same level of service. A promising leverage of utilities is the improvement of technical asset management and, in particular, the use of digital solutions to improve the efficiency of inspection and rehabilitation strategies. Utilities often lack appropriate tools to plan and manage long-term investment needs (Black & Veatch 2013) and rely on reactive strategies, repairing mainly when failures occur.

To tackle this issue, deterioration models have been developed to forecast the evolution of the system according to its current and past condition observed during CCTV inspections. Deterioration models can be used to simulate the condition of non-inspected pipes and to forecast the evolution of the condition of the sewer network under different investment strategies. Model outputs provide key information to operators and municipalities for the scheduling of inspection programs (i.e., the detection of sewers in critical condition) and the planning of rehabilitation budgets (i.e., the comparison of different sewer rehabilitation scenarios and the evaluation of necessary investment rates).

Several modelling approaches have been developed over the last 20 years to support rehabilitation planning (e.g., Babovic *et al.* 2002; Savić *et al.* 2009; Ward & Savić 2012; Scholten *et al.* 2014) as well as inspection and maintenance prioritization (e.g., Baur & Herz 2002; Berardi *et al.* 2009). For a detailed review of modelling approaches, refer to Ana & Bauwens (2010), Rokstad & Ugarelli (2015) or Kley & Caradot (2013). Many studies have intended to evaluate the performance of statistical and machine learning deterioration models (e.g. Tran *et al.* 2006; Chughtai & Zayed 2008; Ana *et al.* 2009; Khan *et al.* 2010; Salman 2010; Harvey & McBean 2014; Sousa *et al.* 2014; Ahmadi *et al.* 2015; Rokstad & Ugarelli 2015). The outcomes of these studies underline the relevance of using deterioration models to support asset management strategies but suffer from two main shortcomings as follows:

- The lack of data for model calibration: models are often calibrated with less than 2,000 pipes which represent only a small part of the networks (except for the studies of Salman (2010) and Rokstad & Ugarelli (2015) that developed deterioration models in the cities of Cincinnati in the USA and Oslo in Norway using 11,373 and 12,003 pipes, respectively).
- The lack of clear metrics adapted to utilities' issues for the assessment of model performance: most metrics are based on statistical tests (e.g., mean square error, goodness-of-fit and coefficient of determination) and do not provide a full understanding of the potential of deterioration models for municipalities and sewer operators. Furthermore, the metrics often assess the overall performance of the models without exploring the single performances for the prediction of each condition class. Since the condition of sewer networks is mostly imbalanced (many pipes in good condition and few pipes in poor condition), this assessment can lead to biased conclusions.

Research objectives

The proper validation of deterioration models is the key to build the confidence of utilities regarding the models' use. Deterioration models can be successfully deployed only if decision-makers trust the modelling outcomes and are aware of the model uncertainties. Our study aims to address this issue with the following objectives:

- Develop a set of metrics adapted to the local utility needs to assess the performance of sewer deterioration models from an end-user perspective. The metrics must be intuitive, self-explanatory and so clearly understandable by the sewer operator. Thus, they should be able to convince the utility about the relevance or uselessness of using deterioration models to support asset management strategies. This point is crucial for facilitating the communication of the outcomes and ensuring the acceptance of the results.
- Apply the set of developed metrics to benchmark the performance of a statistical and a machine learning deterioration model trained with the extensive CCTV dataset of the sewer of Berlin, Germany. In Germany, regional regulations commit sewer operators to inspect their networks entirely every 10 or 15 years. In Berlin, each pipe of the almost 10,000 km sewer network has

been inspected at least once by the end of 2016. The inspection database is a highly valuable knowledge that can be exploited to assess model performance in the special case of full data availability.

The statistical approach selected is the model GompitZ (Le Gat 2008; Werey *et al.* 2012), based on the theories of survival analysis and Markov chains. The machine learning approach selected is Random Forest, an ensemble learning method for classification or regression, already successfully implemented for the prediction of sewer pipes condition (Harvey & McBean 2014; Vitorino *et al.* 2014; Rokstad & Ugarelli 2015). Outcomes of the benchmark analysis will be used to draw conclusions on the strengths and weaknesses of both approaches regarding asset management objectives.

MATERIALS AND METHODS

Data preparation

The study has been performed using the extensive GIS and CCTV database of the city of Berlin in Germany (3.5 million inhabitants).

The sewer network is composed of 235,988 pipes (9,710 km) registered in the GIS database. Most pipes are sanitary sewers (45%), 35% are stormwater sewers and 20% are combined sewers. Clay and concrete are the two dominating materials with proportions of 54% and 25%, respectively. The GIS database contains the main pipes' characteristics (construction year, material, type of effluent, shape, diameter, length, depth, slope, city district) and has been extended with environmental features expected to influence sewer deterioration (tree density, proximity of tramway or subway, groundwater level, type of soil).

The Berlin water company has conducted extensive CCTV inspection programs since the 1980s. Sewer defects observed during inspections are systematically coded in a local coding system similar to the German guideline ATV M 143-2 (1999). Sewer structural condition is evaluated using an internal company classification system with three grades indicating the emergency of rehabilitation.

After data preparation (consistency check, filtering and clean up), 115,258 inspections with a length of 4,825 km over 102,258 pipes were available for the study. The distribution of the main pipe characteristics and environmental features for the inspected sewer network of Berlin is shown in Figure 1.

Of the inspected pipes, 22% are in poor or very poor condition (condition class 3) and require immediate or short-term rehabilitation measures; 24% of the pipes are in a medium condition (condition class 2) and must be rehabilitated in the medium term (time horizon: 10 years) whereas 54% of the pipes are in good or perfect condition (condition class 1). Figure 2 shows correlations between the main sewer characteristics and the sewer structural condition. The condition is clearly correlated with the pipe age; old pipes are in worse condition than new pipes. However, the condition of very old pipes (>100 years old) seems to improve slightly. This phenomenon is known as survival selection bias. Inspection data have a tendency to be biased as the observations are carried out in a restricted time window (for this study from 2001 to 2016). Most old or deteriorated pipes have already been replaced, thus are not fully represented in the sample of inspection data.

The condition is also correlated with the pipe material; sewers made of concrete are in worse condition than clay pipes or PVC pipes. Since stormwater pipes are mainly constructed with concrete, they appear to be in worse condition than sanitary pipes. The width and the depth also seem to play a relevant role: small-diameter and shallow pipes are in worse condition than big pipes and deeply laid pipes, respectively. The condition distribution varies strongly between the districts probably due to cross correlations with other pipes' external characteristics (e.g., type of soil, type of effluent, age of the network, etc.).

Modelling approaches

Random Forest

Random Forest (RF) is an ensemble learning method for classification or regression. It consists in growing hundreds of decision tree classifiers – in our case of type 'CART' (Breiman *et al.* 1984) – and combining them in a single ensemble of models (Breiman 2001). For classification tasks, the goal is to predict a class output (e.g., condition class) from a set of numerical or categorical variables. The

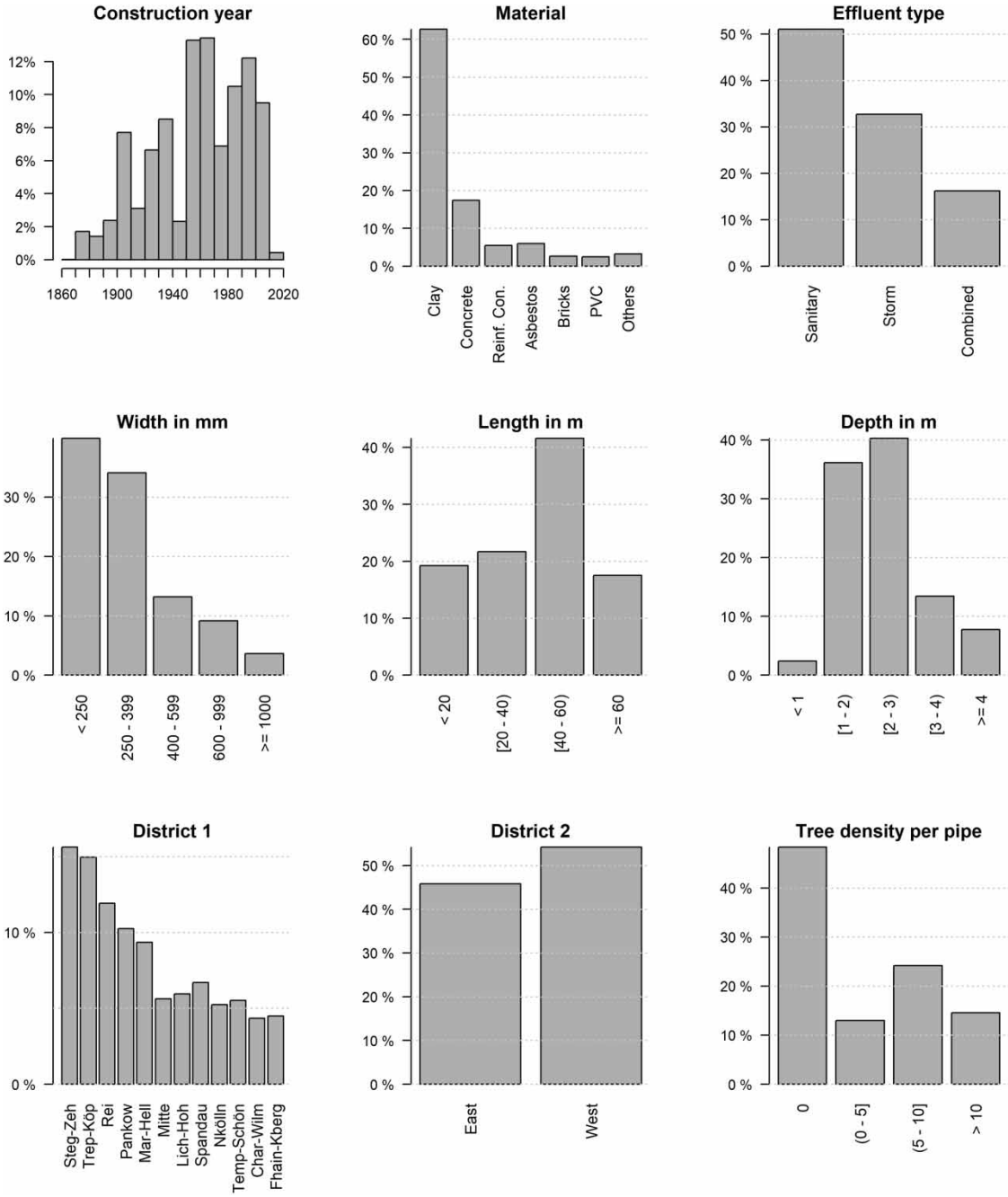

Figure 1 | Distribution of pipes characteristics in the sewer network of Berlin – only inspected pipes are considered.

algorithm builds individual unpruned trees using bootstrap aggregation with the following procedure:

- Sample n instances randomly (with replacement) from the original training dataset.
- Start the construction of the tree from the root with the n instances.
- Search through *mtry* random variables among M variables to find the best binary split into two children nodes. The best split is determined by minimizing the Gini criterion (Breiman 2002). The criterion evaluates the performance of the split to classify the output: the maximum value of the Gini criterion is obtained if the

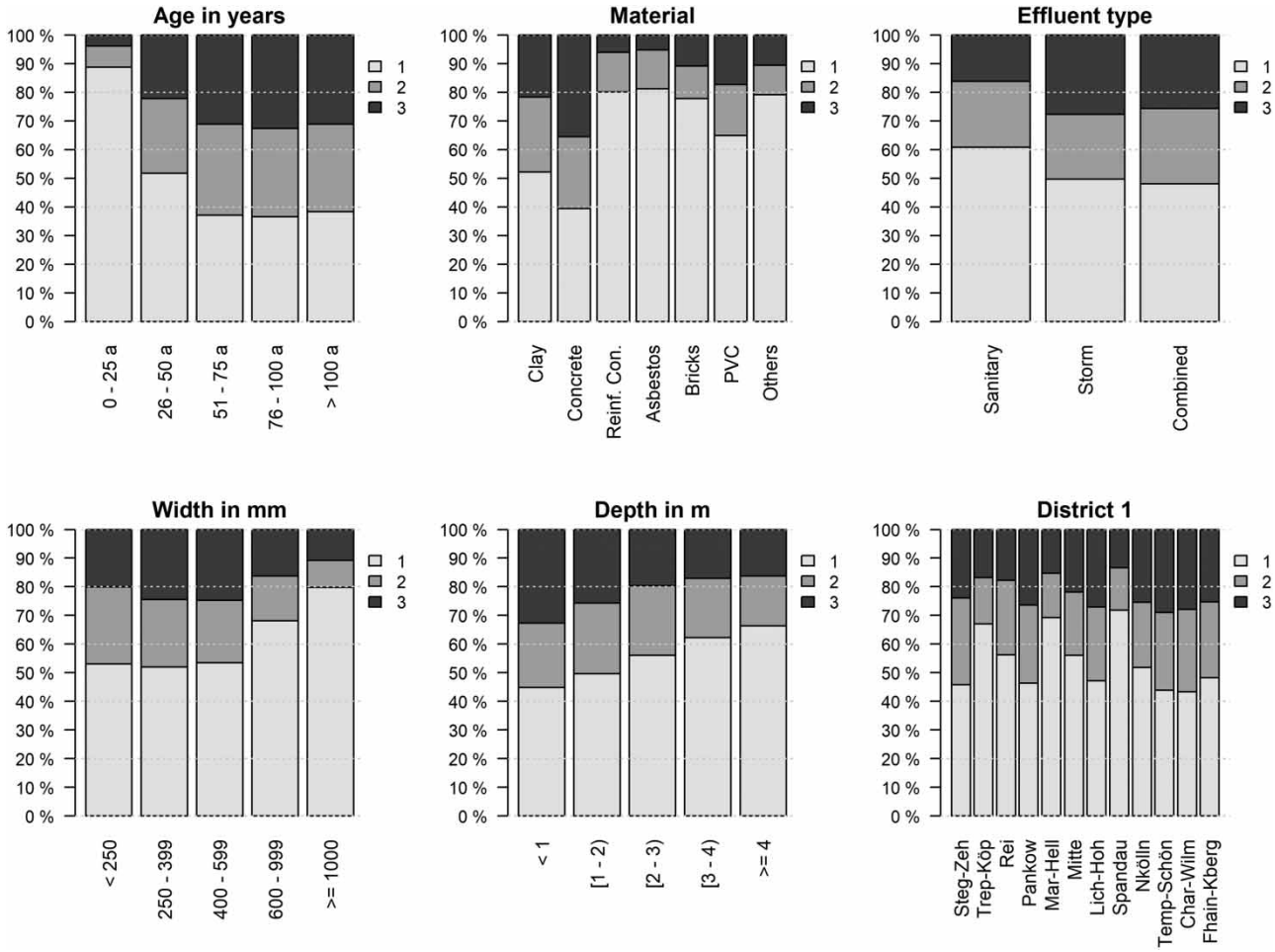

Figure 2 | Correlation between sewer characteristics and condition. Class 1 (light grey), 2 (medium grey) and 3 (dark grey) represent good, medium and bad condition, respectively.

distribution of the output is the same in both nodes (poor classification); the Gini criterion is 0 if the values of the outputs are perfectly separated between the two nodes (excellent classification).

- Repeat the previous step to grow the tree until the number of instances in the children nodes reach the critical size defined by the hyperparameter *nodesize*. The *nodesize* determines the size of the trees.

A total of *ntree* trees are grown with the same procedure. The resulting ensemble of trees composes the RF. For a given set of variables, each tree delivers a class output; the prediction of the RF is the mode of the *ntree* class outputs. Class probabilities can also be estimated as the percentage of each class among the *ntree* class outputs.

RF can also deal with imbalanced data in which one of the output classes constitutes a small minority of the data. In such cases, the interest usually leans towards the correct classification of the minority class (e.g., fraud detection, disease diagnostic, etc.). A classical RF might fail because it will seek to minimize the overall error rate, rather than paying special attention to the minority class. The main approach to tackle the problem of imbalanced data is to incorporate class weights into the RF to penalize the misclassification of the minority class. Class weights are used to weight the Gini criterion and to determine the class output at the terminal node of each tree using a weighted majority vote (Chen & Breiman 2004).

The analysis of the trees structure highlights the relative importance of each variable in the model. The minimal depth is a dimensionless statistic measuring the

productiveness of a variable in a decision tree (Ishwaran *et al.* 2010). The minimal depth of a given variable in a tree is the highest level of the nodes in which the variable has been selected to classify the pipes in the training process. Since the algorithm selects at each node the variable that leads to the best classification, important variables have small minimal depths.

Harvey & McBean (2014) used RF to predict sewer pipes in bad condition in Ontario, Canada. Results were satisfying with false negative rate of 11%, false positive rate to 25% and an area under the ROC curve >0.80 (with 1.0 indicating a perfect model). However, only 1,255 pipes were available for training from which around 10% in poor condition. Rokstad & Ugarelli (2015) used RF with data of the city of Oslo, Norway. They conclude that deterioration model application may be beneficial for prioritizing inspection programs and that the performance is limited by the adequacy of the explanatory variables available. Vitorino *et al.* (2014) also demonstrated an application of a RF implemented in the software platform Baseform.org.

Random forests for this study were developed using the randomForest package in the software R (Liaw & Wiener 2002). Using three output classes, the following hyperparameters need to be set up by the user:

- *mtry* number of variables randomly sampled as candidates at each split
- *nodesize* minimum size of terminal nodes
- *ntree* number of trees in the forest
- *w1*, *w2* priors (i.e., weights) of the classes 1 and 2. *w3* is not defined as a parameter since its value can be calculated from the other priors ($w1 + w2 + w3 = 1$).

Markov-chains and survival analysis

The model GompitZ (Le Gat 2008) is based on the theories of survival analysis and Markov-chains. The goal of the model is to predict a probability class output (probability for a pipe to be in a given condition class) from the pipe age and a set of numerical or categorical variables.

Prior to model calibration, pipes are generally grouped in cohorts, i.e., homogenous groups of sewer pipes sharing similar features, e.g., same material and type of effluent. During the calibration procedure, survival functions are estimated for each cohort. Survival curves have the mathematical form of a Gompertz distribution. They are calibrated during a regression procedure using the maximum likelihood estimation and represent the mean deterioration of pipes over time: they define the proportion of pipes that have survived at a given age. Additionally, the shape of the survival curves can be modulated by further numerical or categorical variables (also called covariates).

The calibrated survival curves are used to calculate the probability for a pipe to be in a given condition at a given age. If the pipe has never been inspected, the class output is estimated directly from the survival curves. For example, for a given pipe and three possible condition classes, a probability vector P is estimated at year T from the survival curves *SC1* and *SC2*:

$$\begin{aligned} P(T) &= (P_1(T),\ P_2(T),\ P_3(T)) \\ &= (SC_1(T),\ SC_2(T) - SC_1(T),\ 1 - SC_2(T)) \end{aligned} \quad (1)$$

If the pipe has been inspected at least once, the condition is known at the year of the inspection and the survival curve cannot be used straightforwardly to estimate the future condition. For example, if the pipe has been inspected in condition 1 at year T:

$$P(T) = (1,\ 0,\ 0) \quad (2)$$

At year T, the pipe is in condition 1 with 100% probability. The probability vector has been initialized at year T of the inspection. In this case, a Markov-chain is used to simulate the future evolution of the pipe condition. The probability vector at year $T+1$ depends on the probability vector at year T and on the transition matrix Q at year $T+1$.

$$P(T+1) = P(T) \times Q(T+1) \quad (3)$$

The transition matrix can be mathematically derived from the slope of the survival curves (Le Gat 2008). The elements of the matrix are time-dependent and indicate the probability for a pipe to stay in a given condition i (probability $q_i(T)$) or to transit to the next condition $i+1$

(probability $1 - q_i(T)$)

$$Q(T) = \begin{vmatrix} q_1(T) & 1-q_1(T) & 0 \\ 0 & q_2(T) & 1-q_2(T) \\ 0 & 0 & q_3(T) \end{vmatrix} \quad (4)$$

The Markov-chain for a prediction of the pipe condition at year n can be written as follows:

$$P(T+n) = P(T) \times Q(T+1) \times Q(T+2) \times \ldots \times Q(T+n) \quad (5)$$

$$P(T+n) = P(T) \times \prod_{1}^{n} Q(T+i) \quad (6)$$

Performance metrics

A set of performance metrics has been defined in consultation with Berlin Water Company (BWB) in order to benchmark and evaluate model performance. The metrics assess model performance at two main levels: the network and the pipe levels. At network level, the metrics indicate to which extent the model is able to predict the condition distribution of the entire network, i.e., the number of pipes in each condition. At pipe level, the metrics verify to which extent the model is able to predict correctly the inspected condition class of each single pipe. Information of both is needed for different purposes: network level metrics show models' relevance for supporting strategic rehabilitation planning; pipe level metrics illustrate the potential for supporting inspection strategies by identifying pipes in critical condition.

Metrics at network level

The performance metrics at network level describe the deviation between the predicted and inspected condition distributions, for the entire network and for different age groups. Six metrics have been defined with the sewer operator.

Deviation of the condition distribution – all pipes: K1, K2, K3 are the absolute deviations between the percentages of sewers predicted and inspected in each condition – K1 for good condition, K2 for medium condition and K3 for bad condition.

Deviation of the condition distribution – only pipe category 51–75 years: K4, K5, K6 are the absolute deviations – for the age category 51–75 years only – between the percentages of sewers predicted and inspected in each condition – K4 for good condition, K5 for medium condition and K6 for bad condition.

K1, K2 and K3 assess the ability of the model to simulate the condition distribution of the entire network whereas K4, K5 and K6 evaluate the ability of the model to consider the deterioration process. The age category of 51–75 years has been selected instead of the oldest category because it corresponds to the depreciation period of concrete and clay pipes. Indeed, older age categories might be biased since many pipes have already been rehabilitated introducing a survival selection bias.

Metrics at pipe level

A model can provide excellent results at network level and nevertheless fail to simulate the right condition of each pipe by simulating the right proportions of each condition but the wrong pipes in each condition. An exhaustive model assessment requires the analysis of the confusion matrix of the outcomes (Table 1). The confusion matrix compares the predicted and observed class of each pipe and counts the number of agreements and disagreements.

Several metrics have been derived from the matrix and validated with the sewer operator.

The True Positive rate: also called sensitivity, indicates the percentage of sewers inspected in condition 'i' that have been correctly predicted in the same condition 'i'.

$$K7 = \frac{\text{number of correct predictions in good condition}}{\text{number of observations in good condition}} = \frac{420}{508} = 83\%$$

Table 1 | Example of confusion matrix with fictive numbers

		Prediction			
		Good	Medium	Bad	Sum observations
Observation	Good	420	56	32	508
	Medium	64	140	25	229
	Bad	36	28	123	187
Sum predictions		520	224	180	

$$K8 = \frac{\text{number of correct predictions in medium condition}}{\text{number of observations in medium condition}} = \frac{140}{229} = 61\%$$

$$K9 = \frac{\text{number of correct predictions in bad condition}}{\text{number of observations in bad condition}} = \frac{123}{187} = 66\%$$

The False Negative rate: also called miss rate, indicates the percentage of sewers inspected in condition 'i' that have been wrongly predicted in a better condition 'j'. False Negative predictions overestimate the inspected condition of the pipes.

$$K10 = \frac{\text{number of pipes observed in medium condition but predicted in good condition}}{\text{number of observations in medium condition}} = \frac{64}{229} = 28\%$$

$$K11 = \frac{\text{number of pipes observed in bad condition but predicted in good condition}}{\text{number of observations in bad condition}} = \frac{36}{187} = 19\%$$

There is no False Negative rate for the good condition since it cannot be overestimated.

The False Positive rate: also called false alarm probability, indicates the percentage of sewers inspected in condition 'i' that have been wrongly predicted in a worse condition 'j'. False Positive predictions underestimate the inspected condition of the pipes.

$$K12 = \frac{\text{number of pipes observed in good condition but predicted in bad condition}}{\text{number of predictions in bad condition}} = \frac{32}{180} = 18\%$$

Summary metrics

The metrics defined above are intuitive and clearly understandable by the sewer operator and will be used to convince the utility about the relevance or uselessness of using deterioration models. In order to simplify the search of the best combination of hyperparameters in the step of model training, metrics at network and pipe levels have also been summarized in one unique single metric each. The summary metrics are defined as the root mean square error of the six indicators on both network and pipe level. The square root is used to give more weight to large errors. K7, K8 and K9 are normalized, i.e., subtracted from 100, to have an optimum value of 0.

$$K_{\text{Network}} = \sqrt{\frac{K1^2+K2^2+K3^2+K4^2+K5^2+K6^2}{6}}$$

$$K_{\text{Pipe}} = \sqrt{\frac{(100-K7)^2+(100-K8)^2+(100-K9)^2+K10^2+K11^2+K12^2}{6}}$$

Model training and testing

After the withdrawal of pipes with missing age, length or depth information or highly underrepresented profiles, material or soil types, the dataset has been separated in two random subsets: training (60%, 58,528 pipes) and test (40%, 39,019 pipes) subsets. The partition 60/40 is commonly used in statistical and machine learning studies, among ratios usually varying between 50/50 and 90/10. In this study, considering the large size of the dataset, the partition size has little influence and does not influence significantly the values of the parameter estimates and performance metrics obtained (results not shown here).

A Chi-squared test of independence (χ^2) has been performed for each pipe feature presented in Figure 1 to compare the training and test subsets. It tests the equality of proportions between the two subsets with the null hypothesis that the distributions of the categorical variables are the same in the two subsets. Reported p-values were higher than the significance level of 0.05 indicating that the null hypothesis cannot be rejected and that the distributions are the same for each categorical variable.

For the Random Forest model, the best combination of hyperparameters have been analysed in two steps in order to reduce the computation time. The idea is to run a first coarse grid search to find the optimal values for the most sensitive parameters, in our case the weighting factors, and a second fine grid search with fixed weight values to identify

the optimal values of the remaining two hyperparameters. With this two-step procedure, the computation time for the parameter search can be reduced compared to a full grid search covering all hyperparameters (Bergstra & Bengio 2012).

- Step 1: random search: a list of 1,000 random hyperparameter combinations has been prepared based on the reasonable range of variation of the four hyperparameters (*w1*, *w2*, *mtry* and *nodesize*). For each combination of hyperparameters, a five-fold cross-validation procedure has been performed on the training dataset using the performance metrics defined in the previous section.
  - (1) The training dataset is divided into five random equal sized subsets.
  - (2) Of the five subsets, a single subset is retained as the validation data to calculate the performance metrics and the remaining four subsets are used to train the model.

  The procedure (1 and 2) is repeated five times with each subset used once as validation data. Finally, the mean of the five sets of performance metrics is calculated. The values of the metrics $K_{Network}$ and K_{Pipe} are plotted against the weight values *w1* and *w2* in order to identify the values that maximize the performance.
- Step 2: grid search: step 1 is repeated with fixed values of weights and varying the values of the remaining hyperparameters *mtry* and *nodesize*. The values of the metrics $K_{Network}$ and K_{Pipe} are plotted against the hyperparameters *nodesize* and *mtry* in order to identify the combination of hyperparameters that maximize the performance. Finally, the best combination of hyperparameters is implemented to train the Random Forest model.

For the GompitZ model, the training consists in identifying the relevant cohorts and variables for the calibration of the survival curves. Similar to Random Forest, a five-fold cross-validation procedure has been performed on the training dataset using the performance metrics defined in the previous section.

- In a first step, a cross-validation has been run for all combinations of cohorts built with one to five categorical variables.
  - If only one variable is used to build the cohorts, pipe groups are composed based on the categorical values of the variable (e.g., for the variable material: clay, concrete, etc.).
  - If several variables are used to build the cohorts, the combination of the variables' values composes the cohorts (e.g., for variables material and sewerage: concrete-sanitary, concrete-storm, clay-sanitary, etc.).
- For the best combination of cohorts, a second cross-validation has been applied by considering additional numerical variables as variables for model calibration.

Finally, the trained models have been tested on the independent test subset to assess model performance. The sewer condition has been predicted at the year of inspection with each sewer and the performance metrics have been evaluated.

Assessment of model performance

The metrics proposed in the section 'Performance metrics' enable the evaluation of model performance. In order to understand the variation's range of the metrics, the model metrics have been compared with the performance of a simple random model and an ideal model. The goal is to assess the meaning of the model metrics within their range of variations from a poor performing model to an ideal model.

The simple random model attributes randomly a condition class to each inspected pipe. The values of the metrics obtained represent a poor performing model. The ideal model has the same accuracy as a sewer inspection; it is based on the assumption that it is impossible to know the sewer condition better than with a CCTV inspection. A methodology to assess CCTV uncertainties has been proposed by Caradot *et al.* (2017). The approach is based on the analysis of repeated inspections of sewer pipes; it considers only repeated inspections that occur within a short time period (<3 or 5 years) in order to neglect sewer deterioration. Thus, variations between the condition classes reflect the uncertainties of the procedure of sewer condition assessment. The methodology has been applied using repeated CCTV inspections of 4,695 pipes in Berlin in

order to assess the True Positive, False Negative and False Positive rates of sewer inspection.

RESULTS AND DISCUSSION

Random Forest

The random search analysis (step 1) has been performed on the training dataset using the four hyperparameters. The tested ranges of the hyperparameters are: *nodesize*: 4–1,808; *mtry*: 1–12; *w1*: 0.2–3; *w2*: 0.2–3. Figure 3 shows the sensitivity of the summary indicators at network and pipe levels depending on the weights factors *w1* and *w2*. The graphs indicate optimal weight values at network level ($w1 = 2$; $w2 = 1$) and at pipe level ($w1 = 1$; $w2 = 0.8$) and suggest the training of one model for each level.

The grid search analysis (step 2) has been performed with fixed weight values to identify the optimal values for *nodesize* and *mtry* at both network and pipe levels. Figure 4 shows the sensitivity of the pipe level metrics to the variation of the two hyperparameters: *nodesize* (minimum size of terminal nodes for each tree) and *mtry* (number of variables randomly sampled as candidates at each split). At pipe level, best results are obtained with high values of *mtry* (11) and values of *nodesize* between 20 and 90. The number of trees grown to build the forest (*ntree*) has no influence on model performance (results not shown here). Table 2 summarizes the best combination of hyperparameters at both network and pipe levels.

Finally, the trained models have been tested on the independent test subset to assess model performance. Figure 5 compares the inspected and predicted condition distributions of the network for the test dataset.

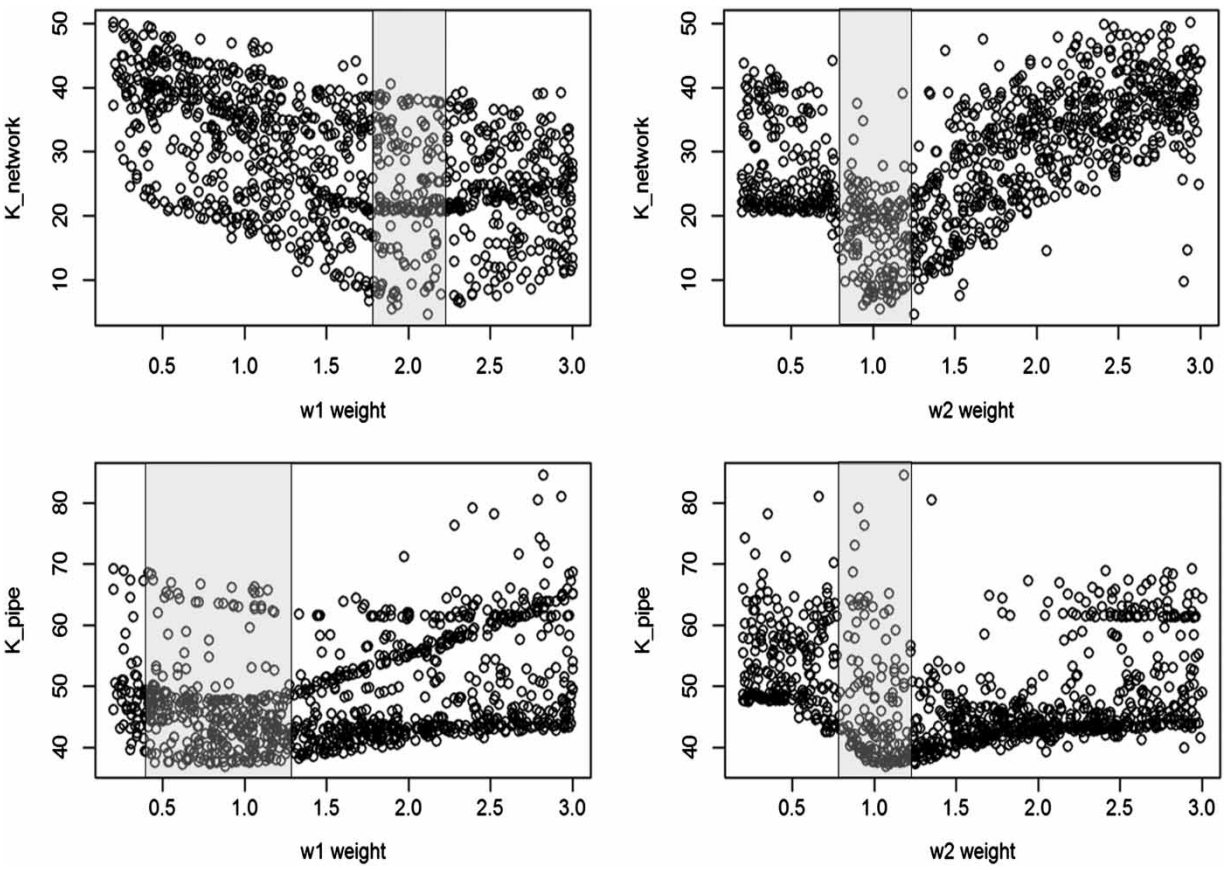

Figure 3 | Influence of the weights factors w1 (left) and w2 (right) over the summary indicators $K_{network}$ (top) and K_{pipe} (bottom); the optimal parameter's window is shaded in grey.

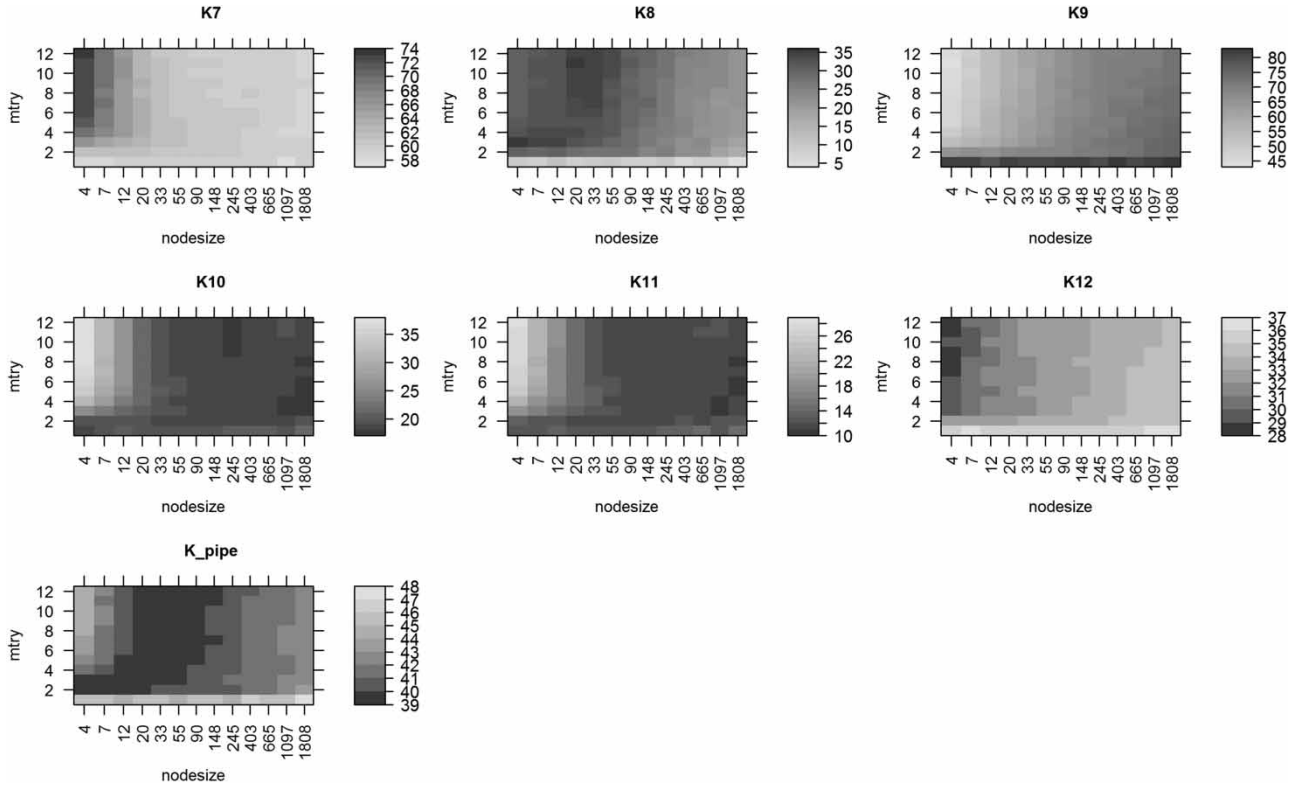

Figure 4 | Sensitivity of the pipe level metrics to the variation of nodesize and mtry.

Table 2 | Best combinations of hyperparameters at both network and pipe levels

Hyperparameter	Best model at network level	Best model at pipe level
ntree	100	100
nodesize	7	55
mtry	10	11
w1	2.0	1.0
w2	1.0	0.8
w3	1.0	1.0

Table 3 summarizes the value of metrics obtained on the test dataset. The metrics have been calculated with the best models at both network and pipe levels. The deviations at network level (K1 to K6) are relatively low, below 5%. At pipe level, 64.0% of the pipes inspected in good condition have been predicted correctly (K7), 40.0% of the pipes inspected in medium condition have been predicted correctly (K8) and 66.7% of the pipes in bad condition have been predicted correctly (K9). 17.1% of the pipes inspected in medium condition and 9.5% of the pipes inspected in bad condition have been falsely predicted in good condition (K10 and K11). 28.3% of the pipes inspected in good condition have been wrongly predicted in bad condition (K12).

Figure 6 shows the distribution of the minimal depth of each variable among the 100 trees of the forest built for the pipe level. The most important variable for classification is the sewer age; in 95 of 100 trees, the age is the variable selected by the algorithm for the first split at the root. The material, the shape and the type of effluent are the following most relevant variables; the material and the type of effluent are strongly correlated since most sanitary and stormwater pipes are built of clay and concrete, respectively. The district shows also a strong influence in the model. The district is strongly correlated with pipes' characteristics (for example, pipes in the city centre are mainly combined sewers made of bricks and built in the 19th century) but also with other potential relevant deterioration factors such as the traffic load, the type of soil or the quality of the construction between the former east and west Berlin. The width, length and

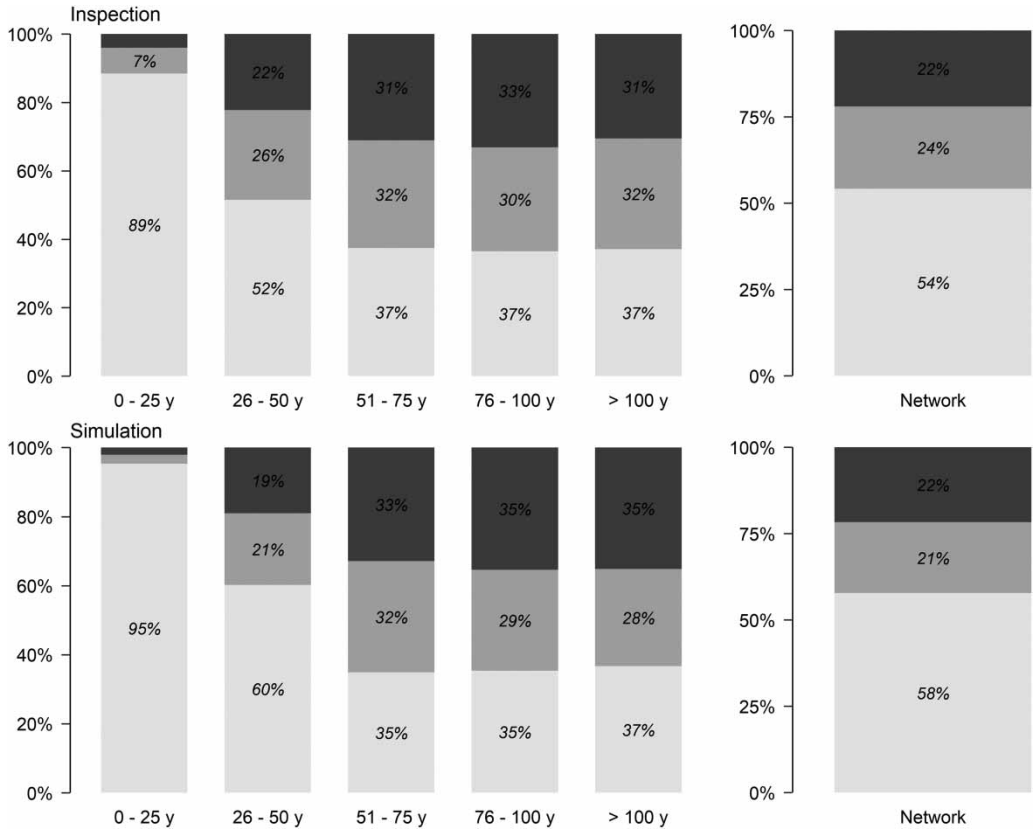

Figure 5 | Inspected and predicted condition distributions with Random Forest for the entire network (right) and for each age group (left). The colours light grey, medium grey and dark grey represent good, medium and bad condition, respectively.

Table 3 | Summary of performance metrics for the two models on the test dataset

				RF	GompitZ
Network level	K1	Deviation of the condition distribution	∈ [0, 100] Goal = minimize	−3.5%	0.8%
	K2			3.4%	−0.1%
	K3			0.1%	−0.7%
	K4	Deviation of the condition distribution		2.0%	0.1%
	K5	– 51–75 years only		−0.1%	0%
	K6			1.9%	−0.1%
Pipe level	K7	True Positive rate	∈ [0, 100] Goal = maximize	64.0%	64.1%
	K8			40.0%	29.0%
	K9			66.7%	32.9%
	K10	False Negative rate	∈ [0, 100] Goal = minimize	17.1%	42.8%
	K11			9.5%	38.0%
	K12	False Positive rate		28.3%	38.5%
Summary	$K_{Network}$	Summary metric for network level	∈ [0, 100] Goal = minimize	2.3	0.5
	K_{Pipe}	Summary metric for pipe level		34.5	51.0

depth are also relevant for describing sewer deterioration but secondary compared to the material and district. Finally, the environmental features show little or no influence (type of soil, tree density, groundwater level).

Markov-chains and survival analysis

The cross-validation procedure has been performed on the training dataset for all combinations of cohorts built with

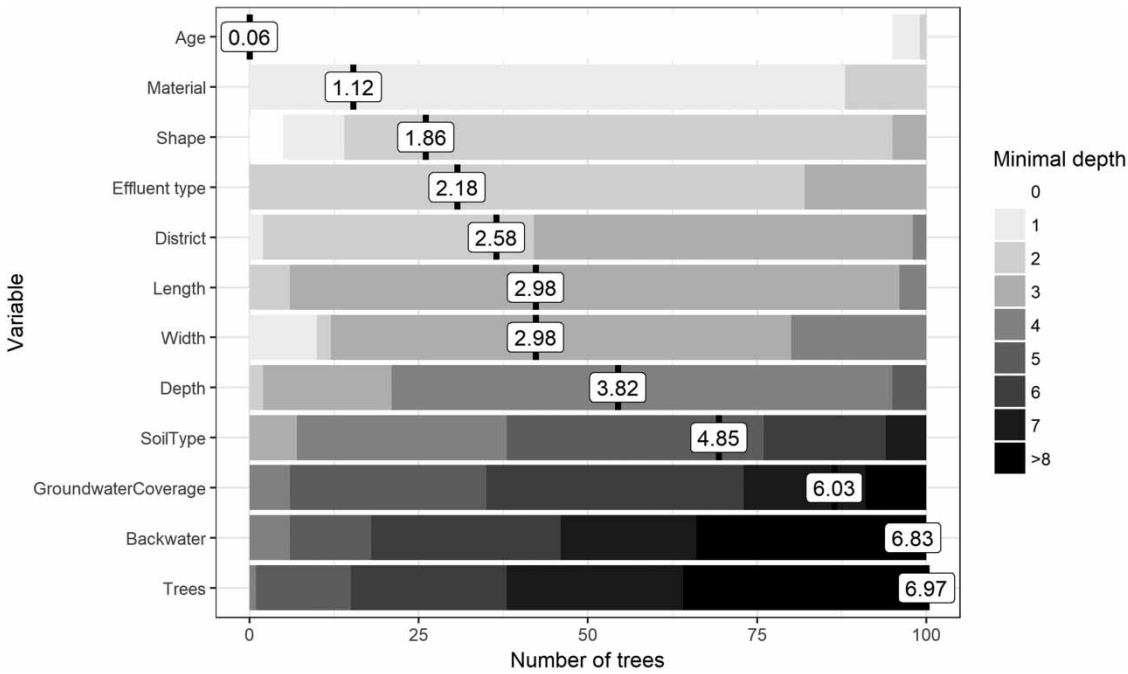

Figure 6 | Distribution of the minimal depth of each variable among the 100 trees. The values in the bar plots indicate the mean minimal depth of each variable.

one to five categorical variables. The best cohort combination is composed of four categorical variables: the material, the district, the shape and the type of effluent. The consideration of additional numerical variables does not improve the model performance. Table 3 summarizes the value of metrics obtained on the test dataset. Figure 7 shows the inspected and predicted condition distributions of the network.

Comparison of model outcomes

Network level

Random Forest and GompitZ give satisfactory outcomes at network level (Table 3). Deviations obtained with Random Forest are below 5%; deviations reached with GompitZ are even lower, below 1%. Both models are able to reproduce accurately the condition distribution of the entire network and for different age groups.

Pipe level

At pipe level, Random Forest performs better than GompitZ (Table 3). In particular, the True Positive rates for pipes in medium and poor condition are 30% and 100% higher, respectively. The False Negative rates and False Positive rate are also minimized with Random Forest. It is interesting to note that the most relevant variables are the same for both models: the material, the district, the shape and the type of effluent.

Figure 8 plots the pipe level metrics of both models and compares model performance to the performances of a random and an ideal model. The following outcomes can be derived for the pipe level:

- GompitZ does not perform better than the random model, expect for the simulation of pipes in good condition (K7).
- The Random Forest performs much better than the random model. The performance is excellent for the simulation of pipes in poor condition (K9) being close to the performance of the ideal model. The False Negative rates (K10 and K11) are also very low, similar to the ideal model. On the other hand, the Random Forest model fails to identify pipes in medium condition (K8) and the False Negative rate is high compared to the ideal model (K12).

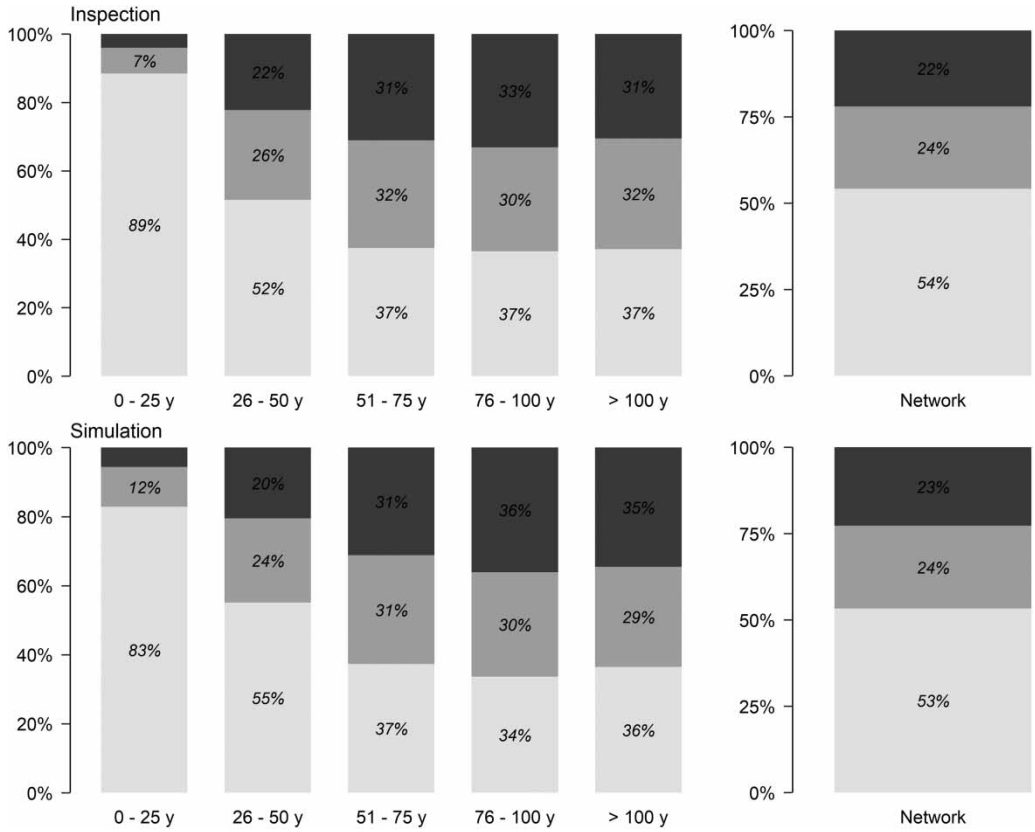

Figure 7 | Inspected and predicted condition distributions with GompitZ for the entire network (right) and for each age group (left). The colours light grey, medium grey and dark grey represent good, medium and bad condition, respectively.

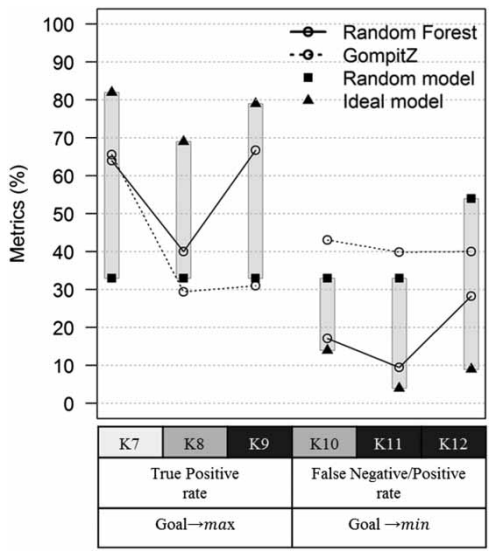

Figure 8 | Comparison of model outcomes with a random and an ideal model.

Simulation example

In order to visualize the outcomes of the Random Forest and GompitZ models, the condition of three single pipes with different characteristics have been simulated with both models (Figure 9). For example, the first pipe shows the deterioration process of a circular sanitary clay pipe situated in the Pankow district of Berlin (Figure 9 – left). The deterioration behaviour is similar with the two models: faster for clay pipes in Pankow, slower for clay pipes in Steglitz and much slower for brick pipes in Mitte. The deterioration is smoother with GompitZ since the survival function follows a Gompertz distribution (Le Gat 2008). The deterioration with Random Forest is much sharper since the model learned from available data without the support of statistical regression. The sharpness of the Random

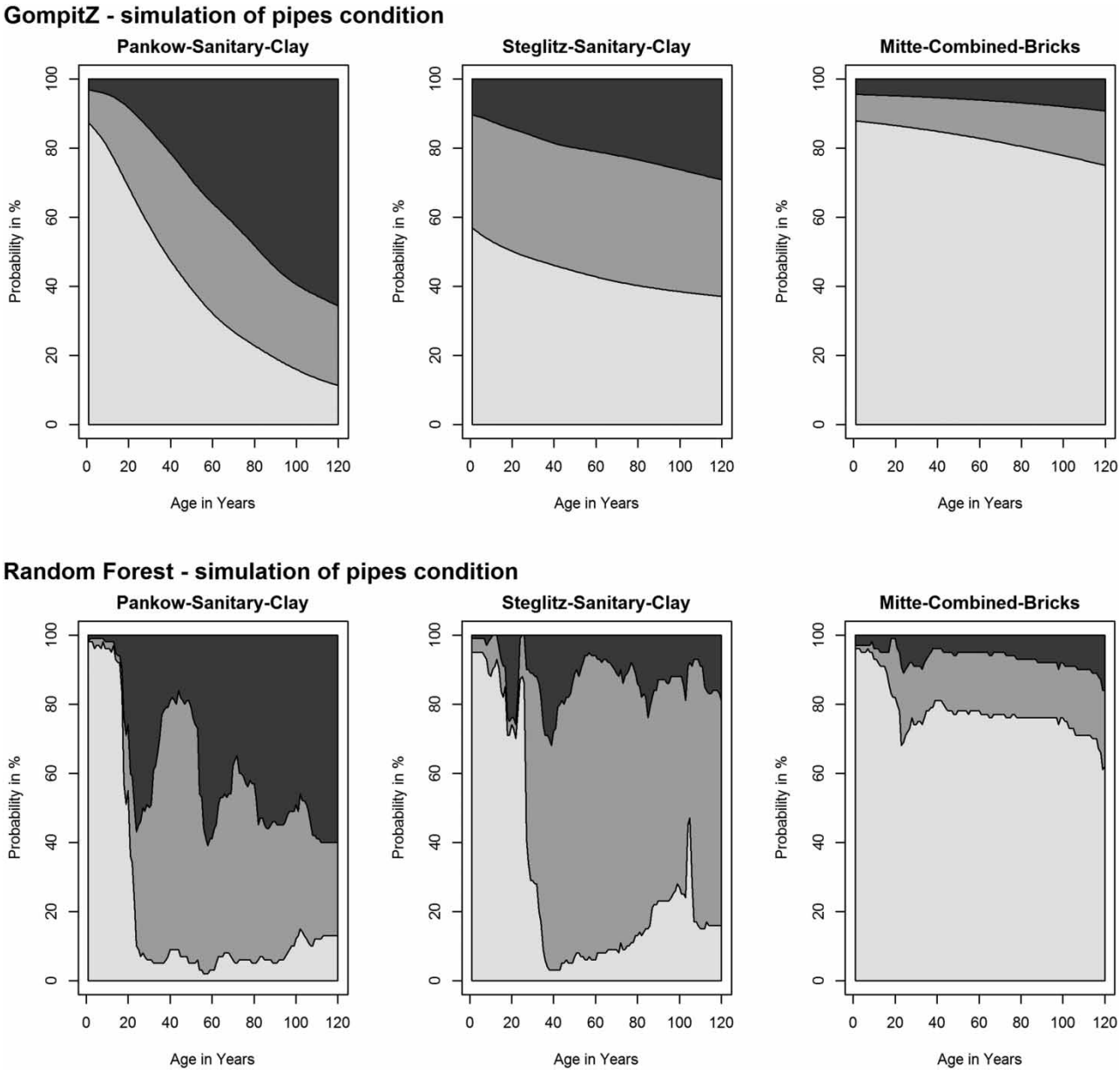

Figure 9 | Example of simulation for three pipes with different characteristics (the title of each graph indicates pipes' characteristics in the following order: district – type of effluent – material). The colours light grey, medium grey and dark grey represent the probability for the pipe to be in good, medium and bad condition, respectively.

Forest prediction is both the strength and the weakness of the model. It allows a more accurate prediction of the condition of specific pipes. However, it might also lead to doubtful predictions such as condition improvement along with pipe age. This leads to the conclusion that the tested machine learning approach shall only be used for ad hoc classification of the sewer pipes but not for long-term forecasts into the future. This characteristic of machine learning models is still to be investigated in order to guarantee the plausibility of future predictions.

CONCLUSION

This study aimed to assess the performance of a statistical and a machine learning deterioration model using the

extensive CCTV and GIS dataset of the city of Berlin, Germany. A set of understandable and intuitive metrics has been defined together with the sewer operator in order to evaluate sewer performance from an end-user perspective. The selected metrics aim at convincing the municipality about the relevance or uselessness of using a given deterioration model to support asset management strategies. The main outcomes can be summarized as follows:

- At network level, both Random Forest and GompitZ give satisfactory outcomes. Deviations between the predicted and inspected condition distributions, for the entire network and for different age groups, are below 5% using Random Forest and even lower (below 1%) using GompitZ. This result underlines the strong potential of both statistical and machine learning models to simulate the condition distribution of the network.
- At pipe level, GompitZ does not perform better than a simple random model, which attributes randomly a condition class to each inspected pipe. GompitZ is not able to simulate the condition of single pipes accurately.
- At pipe level, the Random Forest performs better than GompitZ. Random Forest performance is satisfying for the simulation of pipes in poor condition: 66.7% of the pipes inspected in bad condition have been predicted correctly and only 9.5% of the pipes inspected in bad condition have been falsely predicted in good condition. The True Positive rate of Random Forest for pipes in bad condition (67%) is close to the True Positive rate of a CCTV inspection (79%). The Random Forest model shows a strong potential for supporting sewer operators in the identification of pipes in critical condition for inspection programs.
- The main weakness of the Random Forest model lies in its high False Positive rate: 28.3% of pipes predicted in bad condition are actually in good condition. This aspect of the performance might be improved in further studies by considering additional variables and testing other modelling approaches.
- Another weakness of the Random Forest model is the lack of physical information about pipe deterioration in the model's structure. The model learns and reproduces the patterns observed in the inspection dataset. It can lead to doubtful prediction such as condition improvement along with pipe age. This leads to the conclusion that the tested machine learning approach shall only be used for ad hoc classification of the sewer pipes but not for long-term forecasts into the future. This problem does not occur with GompitZ since the deterioration follows a GompertZ distribution that prevents any condition improvement. This aspect of machine learning should be carefully investigated before deploying such models in practice.

ACKNOWLEDGEMENTS

Part of this research was sponsored by the Berlin Water company (Berliner Wasser Betriebe) in the frame of the research project SEMA-Berlin. The collaboration with the Javeriana University in Bogotá has been supported by DAAD with funds from the German Federal Ministry of Education and Research (BMBF).

REFERENCES

Ahmadi, M., Cherqui, F., De Massiac, J. C. & Le Gauffre, P. 2015 Benefits of using basic, imprecise or uncertain data for elaborating sewer inspection programmes. *Structure and Infrastructure Engineering: Maintenance, Management, Life-Cycle Design and Performance* **11** (3), 376–388.

American Water Works Association 2012 *Buried No Longer: Confronting America's Water Infrastructure Challenge*. AWWA's Infrastructure Financing Report.

Ana, E. V. & 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.

Ana, E. V., 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.

ASCE 2011 *Failure to act: the Economic Impact of Current Investment Trends in Water and Wastewater Treatment Infrastructure*. ASCE report.

ATV M 143-2 1999 *Inspection, Repair, Rehabilitation and Replacement of Sewers and Drains – Part 2: Optical Inspection*. April 1999, DWA – German Association for Water, Wastewater and Waste.

Babovic, V., Drécourt, J. P., Keijzer, M. & Friss Hansen, P. 2002 A data mining approach to modelling of water supply assets. *Urban Water* **4** (4), 401–414.

Baur, R. & Herz, R. 2002 Selective inspection planning with ageing forecast for sewer types. *Water Science and Technology* **46** (6–7), 389–396.

Berardi, L., Giustolisi, O., Savić, D. A. & Kapelan, Z. 2009 An effective multi-objective approach to prioritisation of sewer pipe inspection. *Water Science and Technology* **60** (4), 841–850.

Berger, C., Falk, C., Hetzel, F., Pinnekamp, J., Roder, S. & Ruppelt, J. 2015 *Zustand der Kanalisation in Deutschland (Overview on Sewer Pipes Condition in Germany)*. DWA – German Association for Water, Wastewater and Waste, Hennef.

Bergstra, J. & Bengio, Y. 2012 Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13**, 281–305.

Black and Veatch 2013 *Strategic Direction in the U.S. Water Industry*. Black and Veatch report, https://www.bv.com/docs/reports-studies/sdr-water-industry.pdf (accessed 3 May 2018).

Breiman, L. 2001 Random forest. *Machine Learning* **45**, 5–32.

Breiman, L. 2002 *Manual on Setting Up, Using, And Understanding Random Forests V3.1*. https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf (accessed 3 May 2018).

Breiman, L., Friedman, J. H., Stone, C. J. & Ohlsen, R. A. 1984 *Classification and Regression Trees*. Taylor & Francis, Boca Raton, FL, USA.

Caradot, N., Rouault, P., Clemens, F. & Cherqui, F. 2017 Evaluation of uncertainties in sewer condition assessment. *Structure and Infrastructure Engineering: Maintenance, Management, Life-Cycle Design and Performance* **14** (2), 264–273.

Chen, C. & Breiman, L. 2004 *Using Random Forest to Learn Imbalanced Data*. University of California, Berkeley, CA, USA.

Chughtai, F. & Zayed, T. 2008 Infrastructure condition prediction models for sustainable sewer pipelines. *Journal of Performance of Constructed Facilities* **22** (5), 333–341.

Harvey, R. & McBean, E. 2014 Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering* **41** (4), 294–303.

IPK 2014 *Forderungskatalog 2014: Forderungskatalog für funktionsfähige öffentliche und private Abwasseranlage (Recommendations for the Successful Long-Term Management of Sewer Infrastructures)*. Impulse Pro Kanalbau.

Ishwaran, H., Kogalur, B., Gorodeski, Z., Minn, J. & Laue, S. 2010 High-dimensional variable selection for survival data. *Journal of the American Statistical Association* **105** (489), 205–217.

KfW – Deutsches Institut für Urbanistik – Difu 2016 *KfW-Kommunalpanel 2016 (KfW – Municipality Survey)*. KfW, Frankfurt am Main.

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

Kley, G. & Caradot, N. 2013 *Review of Sewer Deterioration Models*. KWB project SEMA, Report 1.2. Available on KWB website, http://www.kompetenz-wasser.de (accessed 3 May 2018), 43 pages, Berlin, Germany.

Le Gat, Y. 2008 Modelling the deterioration process of drainage pipelines. *Urban Water Journal* **5** (2), 97–106.

Liaw, A. & Wiener, M. 2002 Classification and regression by random forest. *R News* **2/3**, 18–22.

Oelmann, M., Roters, B., Hoffjan, A., Hippe, M. & Wedmann, T. 2017 Investitionsstau in der Abwasserentsorgung (Investment bottleneck in the disposal of wastewater). *KA – Korrespondenz Abwasser, Abfall* **2**, 131–138.

Rokstad, M. M. & Ugarelli, R. 2015 Evaluating the role of deterioration models for condition assessment of sewers. *Journal of Hydroinformatics* **17** (5), 789–804.

Salman, B. 2010 *Infrastructure Management and Deterioration Risk Assessment of Wastewater Collection Systems*. PhD thesis, University of Cincinnati, Cincinnati, OH, USA.

Savić, D. A., Giustolisi, O. & Laucelli, D. 2009 Asset deterioration analysis using multi-utility data and multi-objective data mining. *Journal of Hydroinformatics* **11** (3–4), 211–224.

Scholten, L., Scheidegger, A., Reichert, P., Mauer, M. & Lienert, J. 2014 Strategic rehabilitation planning of piped water networks using multi-criteria decision analysis. *Water Research* **49**, 124–143.

Sousa, V., Matos, J. P. & Matias, N. 2014 Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition. *Automation in Construction* **44**, 84–91.

Tran, D., Perera, A. W. M. & Davis, P. 2006 Application of probabilistic neural networks in modeling structural deterioration of stormwater pipes. *Urban Water Journal* **3** (3), 175–184.

Vitorino, D., Coelho, S. T., Santos, P., Sheets, S., Jurkovac, B. & Amado, C. 2014 A Random Forest algorithm applied to condition-based wastewater deterioration modeling and forecasting. *Procedia Engineering* **89**, 401–410.

Ward, B. & Savić, D. A. 2012 A multi-objective optimisation model for sewer rehabilitation considering critical risk of failure. *Water Science and Technology* **66** (11), 2410–2417.

Werey, C., Rozan, A., Wittner, C., Le Gat, Y., Le Gauffre, P., Nirsimloo, K. & Leclerc, C. 2012 Gestion patrimoniale des réseaux d'assainissement: de l'état des réseaux à la planification de leur réhabilitation – Outils, méthodes et perspectives (Asset management of sewer networks: from sewer condition to rehabilitation planning). *Sciences Eaux & Territoires* **9** (4), 44–53.

First received 21 February 2018; accepted in revised form 31 May 2018. Available online 18 June 2018