


## Article

# Sewer Condition Prediction and Analysis of Explanatory Factors

Tuija Laakso <sup>1,\*</sup> , Teemu Kokkonen <sup>1</sup>, Ilkka Mellin <sup>2</sup> and Riku Vahala <sup>1</sup>

<sup>1</sup> Department of Built Environment, Aalto University, P.O. Box 15200, 00076 Aalto, Finland; teemu.kokkonen@aalto.fi (T.K.); riku.vahala@aalto.fi (R.V.)

<sup>2</sup> Department of Mathematics and Systems Analysis, Aalto University, P.O. Box 11100, 00076 Aalto, Finland; ilkka.mellin@aalto.fi

\* Correspondence: tuija.laakso@aalto.fi; Tel.: +358-50-521-9568

Received: 24 July 2018; Accepted: 1 September 2018; Published: 13 September 2018



**Abstract:** Sewer condition is commonly assessed using closed-circuit television (CCTV) inspections. In this paper, we combine inspection results, pipe attributes, network data, and data on pipe environment to predict pipe condition and to discover which factors affect it. We apply the random forest algorithm to model pipe condition and assess the variable importance using the Boruta algorithm. We analyse the impact of predictor variables on poor condition using partial dependence plots, which are a valuable technique for this purpose. The results can be used in screening pipes for future inspections and provide insight into the dynamics between predictor variables and poor condition.

**Keywords:** Boruta algorithm; logistic regression; partial dependence plot; random forest; sewer condition; variable selection

## 1. Introduction

Utility companies commonly use closed-circuit television (CCTV) inspections to assess sewer condition. The inspection results can provide support for various levels of decision-making. In the simplest case, they support short-term decisions on whether or not to change a pipe. If the results are further processed and analysed, they can also provide support for mid- and long-term asset management decisions (e.g., [1–3]).

Pipe deterioration is known to be a complicated process affected by a multitude of factors [4]. Although many different factors have been identified as influential or potentially influential (see, e.g., [4,5]), it has not been possible to identify a clear set of explanatory variables [6]. CCTV inspection results have been used both in modelling pipe condition and in analysing influential factors. However, Khan et al. [7] note that there has been far less focus on assessing explanatory factors as compared to modelling pipe condition or deterioration.

Condition modelling can be carried out at different resolutions, varying from the network level to pipe cohorts and individual pipes. Network level prediction provides support for high-level asset management decisions, for example for estimating the volume of renovation needs and the related budgets. Compared to network level models, pipe cohort models enable improved accuracy in condition estimation and, consequently, better support for e.g., inspection decisions. When the aim is to support decisions on where to target condition inspections, predicting the condition of individual pipes is the best option [4]. In addition to the spatial resolution, the temporal resolution can also vary from predicting the current condition to predicting the future condition in the mid or long term.

Pinpointing of sewer inspections is a crucial step in sewer asset management, since the inspection results serve as input to renovation decisions and renovation is a costly procedure. Study of how

different pipe attributes and environmental variables affect pipe condition provides an understanding of the mechanisms that cause poor condition. This provides valuable information not only for inspection decisions, but also for future installations. This study focuses on modelling the prevailing condition of a pipe network, and hence lifespan models are discussed only when they address factors explaining pipe deterioration.

The methods previously applied for condition modelling include both traditional statistical methods and machine learning methods. Salman and Salem [8] compared the performance of binary logistic regression, multinomial logistic regression, and ordinal regression in modelling sewer condition. They found ordinal regression to be unsuitable for the study, while binary logistic regression provided the best results. Ariaratnam et al. [9], Ana et al. [10], and Fuchs-Hanusch et al. [11] also applied logistic regression to model poor conditions. Chughtai and Zayed [12] applied multiple regression to model a five-level sewer condition scale. They created four different models: three for predicting the structural condition of different materials and one for predicting the operational condition. Savic et al. [13] and Savic et al. [14] applied a different form of regression, a data-driven modelling algorithm called evolutionary polynomial regression, to model the blockage and collapse events in different pipe classes.

Khan et al. [7], Tran et al. [15], and Sousa et al. [16] applied neural networks to model how different variables affect the condition of the sewer. Tran et al. [15] modelled the structural condition of stormwater pipes, while Khan et al. [7] modelled sewer condition and Sousa et al. [16] the condition of sanitary sewers. Sousa et al. [16] compared the performance of artificial neural networks (ANNs) and support vector machines (SVMs) with that of logistic regression. They found that ANNs provided the highest classification performance. SVMs, on the other hand, provided excellent results in the study by Mashford et al. [17], who predicted sewer condition using a five-level scale.

Decision trees were applied by Syachrani et al. [18] and Harvey and McBean [19], and random forests by Harvey and McBean [20]. Syachrani et al. [18] found that decision trees consistently outperformed the regression and neural networks in predicting the “real age” of sewer pipes. Harvey and McBean [19] compared the performance of a decision tree with that of SVMs and they found that the decision tree outperformed the SVMs. Harvey and McBean [20] predicted pipe condition (good/poor) using random forests and obtained good results.

The reported model performance against data varied greatly from one study to another. The multinomial regression in Salman and Salem [8] correctly predicted the condition rating for 53% of the pipes, while for the SVM model in Mashford et al. [17] the share was 91%. The rest of the reported fits varied between these two. In spite of the interest in the field, few studies have managed to provide tools for understanding how different predictor variables affect pipe condition. This article studies the use of random forests for predicting an individual pipe’s condition, demonstrates the use of the Boruta algorithm for the analysis of variable importance, and explores partial dependence plots in assessing the effect of different predictor variables on pipe condition.

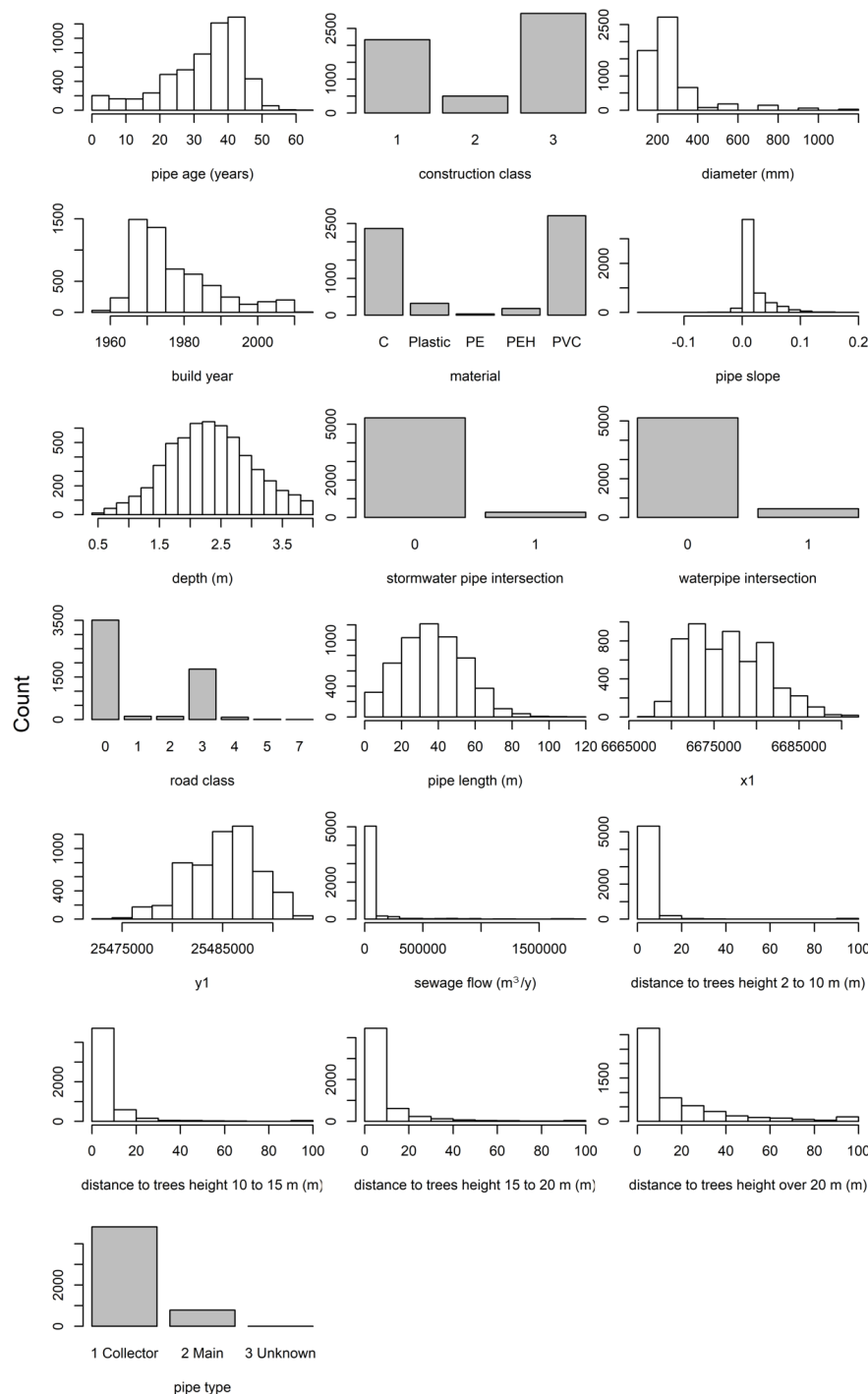
## 2. Materials and Methods

### 2.1. Data Sets and Model Setting

The studied network consists of 1241 km of foul sewers in southern Finland. The network serves a population of approximately 330,000 people and covers an area of ca. 230 km<sup>2</sup>. The main materials in the network are polyvinyl chloride (47%; PVC), other plastics (29%), and concrete (19%). The share of other miscellaneous materials (iron, glass fibre lining and epoxy lining) is 2% and the material is unknown for 3% of the pipes. The oldest pipes in the network were installed in 1955, and continuous construction has been going on since the 1960s. Approximately 30% of the network was inspected using CCTV between the years 2001 and 2016. So far, the utility company has decided on inspections based on expert judgement. The person carrying out an inspection grades the observations on a five-step scale following the Finnish guidelines for sewer inspections, which are an application of the European standard EN-13508-2 (2003). Score 0 indicates no defect, 1 “slight defect”, 2 “minor

defect”, 3 “moderate defect”, and 4 “serious defect”. In the whole data set, 54% of inspected pipes received a score of 2 or lower. The most common defect types in the data set were “deformation”, “root intrusion”, and “surface defect”. These three defect types covered ca. 65% of the total number of defects found.

The complete inspection data set comprises ca. 48,000 observations for ca. 10,000 pipes. In the current analysis, only gravitational pipes and the largest material groups were included. Figure 1 presents the features of the analysed data set.



**Figure 1.** Histograms of the analysed data set. C: concrete; PE: polyethylene; PEH: polyethylene high-density; PVC: polyvinyl chloride.

As can be seen in Figure 1, the data set included “concrete”, “PVC”, “PE”, and “PEH”, but also pipes only labelled “plastic”. “Plastic” was the original label and it was kept unchanged. After removing clear outliers, a data set with observations for ca. 6700 pipes remained. Some of these records contained missing attributes, which were estimated using random forest imputation (see Section 2.3 Random Forest). The data set contained ca. 16,600 maximum score observations, since sometimes there were many observations with the same maximum score for a pipe. Table 1 presents the 19 studied predictor variables for each pipe, i.e., a network section between subsequent manholes.

**Table 1.** Predictor variables studied.

Pipe attributes	Age at time of inspection (years)	numerical
	Installation year	numerical
	Diameter (mm)	numerical
	Material	categorical
	<i>x</i> -coordinate of pipe end number 1	numerical
	<i>y</i> -coordinate of pipe end number 1	numerical
	Slope (m/m)	numerical
	Depth (m)	numerical
Attributes related to pipe environment	Length (m)	numerical
	Soil construction class at the point of observation: 1 = soil types best suited for construction (rock, gravel, sandstone, moraine), 2 = intermediately suitable soils (certain types of silt and clay, fine sand), 3 = least suitable soils (certain types of silt and clay, peatland)	ordinal
	Road class 1 to 7, where road class 1 contains small roads with a carriageway of width 3 m and road class 7 contains one-way motorways with a carriageway of at least 8 m and the rest of the road classes fall between these two; the higher the road class, the larger the road. The classification is based on the topographic database of National Land Survey Finland.	ordinal
	Intersection with a stormwater pipe	categorical
	Intersection with a water supply pipe	categorical
Attributes related to the network structure	Distance to a tree stand (m; classified to four classes based on average tree height)	numerical
	Estimated annual sewage flow (m <sup>3</sup> /y (m <sup>3</sup> per year); water consumption of all water users upstream of the pipe)	numerical

The predictor variables included both continuous variables (such as slope) and categorical variables (such as material). The spatial variation of defects was studied using the *x*- and *y*-coordinates of the pipe end labelled as number 1 in the utility database. Intersection with water and stormwater pipes was considered, since according to the utility’s experience, defects may appear at pipe intersections. The installation year and age, albeit strongly correlated, are both interesting to study separately, since they reflect different aspects in condition modelling: age represents the impact of pipe age and installation year the impact of, for example, quality of work or material in a given year. For discovering whether the pipe was in poor condition we applied binary classification, where scores 3 and 4 implied “poor” condition and scores lower than 3 implied “acceptable” condition. For each pipe, we selected the maximum defect score from all observations available for that pipe to represent the pipe condition. The aim of the model was to locate pipes with serious defects with the reasoning that these would be in most urgent need of renovation or replacement. Therefore, this model does not consider defects of score 2 or 1 to indicate an inspection need, even if a pipe contains several of these less serious defects.

## 2.2. Binary Logistic Regression

Binary logistic regression is a statistical method for modelling the connection between a binary outcome (such as existence or absence of poor condition) and predictor variables, which can be numerical or categorical. The binary logistic regression model is formulated as follows [21]:

$$\log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta_1 x_1 + \dots + \beta_p x_p \quad (1)$$

where  $\pi(x)$  denotes the probability of an outcome  $x = (x_1, \dots, x_p)$ ,  $\alpha$  is the intercept term, and  $\beta_1, \dots, \beta_p$  are regression coefficients for predictor variables  $x_1, \dots, x_p$ . In the binary logistic regression model, the connection between the log odds of an outcome and the predictor variables is linear.

## 2.3. Random Forest

Random forest is a machine learning method developed by Breiman [22], and it can be used for classification and regression problems. Machine learning methods can cope with non-linear relationships both between the dependent and the predictor variables as well as between the predictor variables themselves. This is beneficial in the case of sewer condition prediction, since these connections may be non-linear and predictor variables may have interdependencies on many levels. In addition, no pre-categorization is needed.

A random forest is created by generating a multitude (an ensemble) of independent classification trees based on random bagging samples of observations. Additionally, only a subset of variables is selected to form each tree, which adds another layer of randomness to the model creation. The individual trees are weak classifiers and creating an ensemble of trees significantly improves the prediction accuracy compared to creating only one decision tree. Each tree creates a prediction for the class of each observation, i.e., “votes” for a class, and the random forest model chooses the prediction gaining the most votes [22]. A random forest can operate with data containing numerical and categorical variables and variable values in different scales.

Piragnolo et al. [23] found indications that random forests perform better than SVMs as training sets get larger. Liu et al. [24] discovered that random forests clearly outperformed both ANNs and SVMs in an electronic tongue data classification problem. They also state that random forests are able to deal with classification problems of unbalanced, multiclass, and small sample data without data pre-processing procedures.

Random forest imputation is a method for estimating missing variable values. The algorithm starts by a simple replacement of missing values by variable medians for numeric variables and by the most frequent levels for categorical variables. After this, the values are adjusted iteratively. For numeric variables, the estimated values are replaced by the weighted average of the non-missing observations, where the weights are the proximities. For categorical variables, the estimated values become the category with the largest average proximity.

## 2.4. Adjusting a Binary Model through Changing the Discrimination Threshold

The diagnostic ability of a binary classifier can be assessed using a confusion matrix. A confusion matrix presents predicted and observed labels for all observations as illustrated in Table 2.

**Table 2.** Confusion matrix.

		Predicted	
		Negative	Positive
Observed	Negative	True negative	False positive
	Positive	False negative	True positive

The confusion matrix presents the number of observations in each class. The positive class refers to the existence of an observation such as poor condition and a negative class to its absence. The true positive (*TP*) cell shows the number of correctly classified positive cases and the true negative (*TN*) cell the number of correctly classified negative cases. The false negatives (*FN*s) refer to “missed positive cases” and equal the observations that received a negative label although they are positive. The false positives (*FP*s), the “false alarms”, refer to those observations classified as positive, although they are negative.

The model accuracy is defined as the fraction of correctly classified observations (Equation (2)).

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

The true positive rate (*TPR*) is the share of positives correctly labelled as positive (Equation (3)) and the true negative rate (*TNR*) is the share of negatives correctly labelled as negative (Equation (4)).

$$TPR = \frac{TP}{TP + FN} \quad (3)$$

$$TNR = \frac{TN}{TN + FP} \quad (4)$$

Similarly, the false positive rate (*FPR*) is the share of all negatives incorrectly labelled as positive and the false negative rate (*FNR*) is the share of positives incorrectly classified as negative. The rates relate to each other through equations  $FNR + TPR = 1$  and  $FPR + TNR = 1$ .

When a random forest is applied on a binary classification problem, each tree votes whether or not each observation in the data set belongs to a given class. Typically, the observation is considered to belong to the positive class if the share of positive votes is equal to or higher than 50%, that is, the discrimination threshold is set to 0.5. However, the discrimination threshold can be altered to adjust the diagnostic ability of the classifier. The selection for the optimal threshold depends on the problem context. Often the goal is to maximize the model accuracy, which means maximizing the sum  $TPR + TNR$ . Alternatively, the discrimination threshold can also be set to match a desired *FNR* or *FPR*.

## 2.5. Boruta Algorithm and Partial Dependence Plots

Variable selection is the process of finding those predictor variables that are essential for the modelled phenomenon. The Boruta method [25] is a variable selection algorithm suitable for finding all the variables relevant for the problem. Variable selection is the process of identifying those predictor variables that are essential for explaining the modelled phenomenon. The Boruta method [25] is a variable selection algorithm for detecting relevant explanatory variables. The algorithm evaluates the effect of each variable on the model result. It finds relevant predictor variables by comparing the original variables' importance with importance achievable estimated using randomly permuted copies of the predictor variables. It outputs the loss of classification accuracy (a value between 0 and 1) resulting from permutation. In order to get a statistically valid result, the procedure of adding attributes with permuted values and estimating their importance is repeated until the importance is calculated for each of the attributes or, alternatively, the algorithm reaches a predefined number of model runs. The Boruta algorithm is implemented in the R package randomForest [26].

Friedman [27] suggested the use of partial dependence plots to visualize and analyse the results produced by a gradient-boosting machine. However, partial dependence plots can be applied to interpret the results produced by any black box learning method [27]. Partial dependence plots are a means of graphically presenting the relationships between individual predictor variables and a given outcome while the effect of other predictor variables has been averaged out. The partial dependence function is defined as follows [28]:

$$f_s(X_s) = E_{x_c} f(X_s, X_c) \quad (5)$$

In Equation (5), the partial dependence of  $f$  on a subset of variables  $X_s$  is defined as the expectation of  $f$  over the marginal distribution of all variables excluding  $s$ . The subscript  $c$  refers to complementary. The calculation of a partial dependence plot is carried out by averaging over the training data  $\{X_i, i = 1, \dots, n\}$  with fixed  $X_s$  [29]:

$$\bar{f}_s(X_s) = \frac{1}{n} \sum_{i=1}^n f(X_s, X_{ic}) \quad (6)$$

In a case where a random forest classifier is used to classify observations on a binary scale (such as from poor to good condition) a partial dependence plot depicts the proportion of votes in favour of class 1 (poor condition) against the values of one predictor variable [28]. The limitation of partial dependence plots is that they can only present the outcome against one or two predictor variables and the low-order interactions between the outcome and the predictors, whereas they are not useful for characterizing or interpreting high-order interactions [30].

### 3. Results

#### 3.1. Model Accuracy

We created two models: one with binary logistic regression and one with a random forest, both using ten different data samples. In all cases, 70% of the data were used for training the model, while the remaining 30% were reserved for model validation. Because for many pipes there was more than one observation matching the maximum score, on each round the maximum observation was selected randomly. Practically, the differences between the records representing the maximum score for a pipe were very small, since all the pipe attributes remained the same and only some location-specific attributes, such as location under a road, varied.

The binary logistic regression classified correctly ca. 62% in validation, whereas the random forest model reached a share of ca. 67%. After creating the model, we set the false negative rate to approximately 20%. The reason for the 20% limit was that the aim was to target pipes in poor condition; having a higher share of false negatives would have resulted in higher model accuracy, but a lower usability of the model when selecting pipes for condition inspections. The utility's original wish was that the model would not miss more than 10% of all defects. However, as this would have resulted in an unacceptably high rate of false positives, the *FNR* was set to 20% instead. Setting the *FNR* to 20% dropped the share of correctly classified observations to 56% for binary logistic regression, and to 62% for random forest. Because, in this particular case, the random forest provided better results and since it is likely that random forest can account for a wider set of dependencies between variables, we decided to concentrate on the random forest model.

Table 3 presents how the threshold values, *FNR*, *FPR*, and the model accuracy relate.

**Table 3.** The connection between the discrimination threshold, false negative rate (*FNR*), false positive rate (*FPR*), and accuracy.

	Discrimination Threshold	<i>FNR</i>	<i>FPR</i>	Accuracy
Original model	0.50	0.50	0.20	0.67
<i>FNR</i> set to 20%	0.35	0.20	0.53	0.62
Discrimination threshold set to 0.6	0.60	0.73	0.07	0.64

Table 3 shows how the original model with the highest accuracy yields a *FNR* of 0.5. Lowering the discrimination threshold results in a higher number of pipes being estimated to have poor condition by the model. As a result, a higher share of them is found (lower *FNR*) but more false alerts also occur (higher *FPR*)—the share of times that a pipe is inspected but is found not to have any serious defects. For comparison, setting the discrimination threshold 0.6 would yield an *FPR* of 7% and an *FNR* of 73%.

The sample was found to affect the results only to a limited degree: when the *FNR* is set to 20%, the models achieve an *FPR* between 49% and 56% on 100 different samples and an average *FPR* of 52%.

### 3.2. Variable Importance

Table 4 lists the predictor variables in the order of average importance determined by the Boruta algorithm. The results show the minimum, maximum, and average importance of each variable derived from the 100 different rounds of data sampling. The importance varies between 0 and 100, where 0 means the variable has no impact on classification results and 100 means that the variable fully determines the classification results.

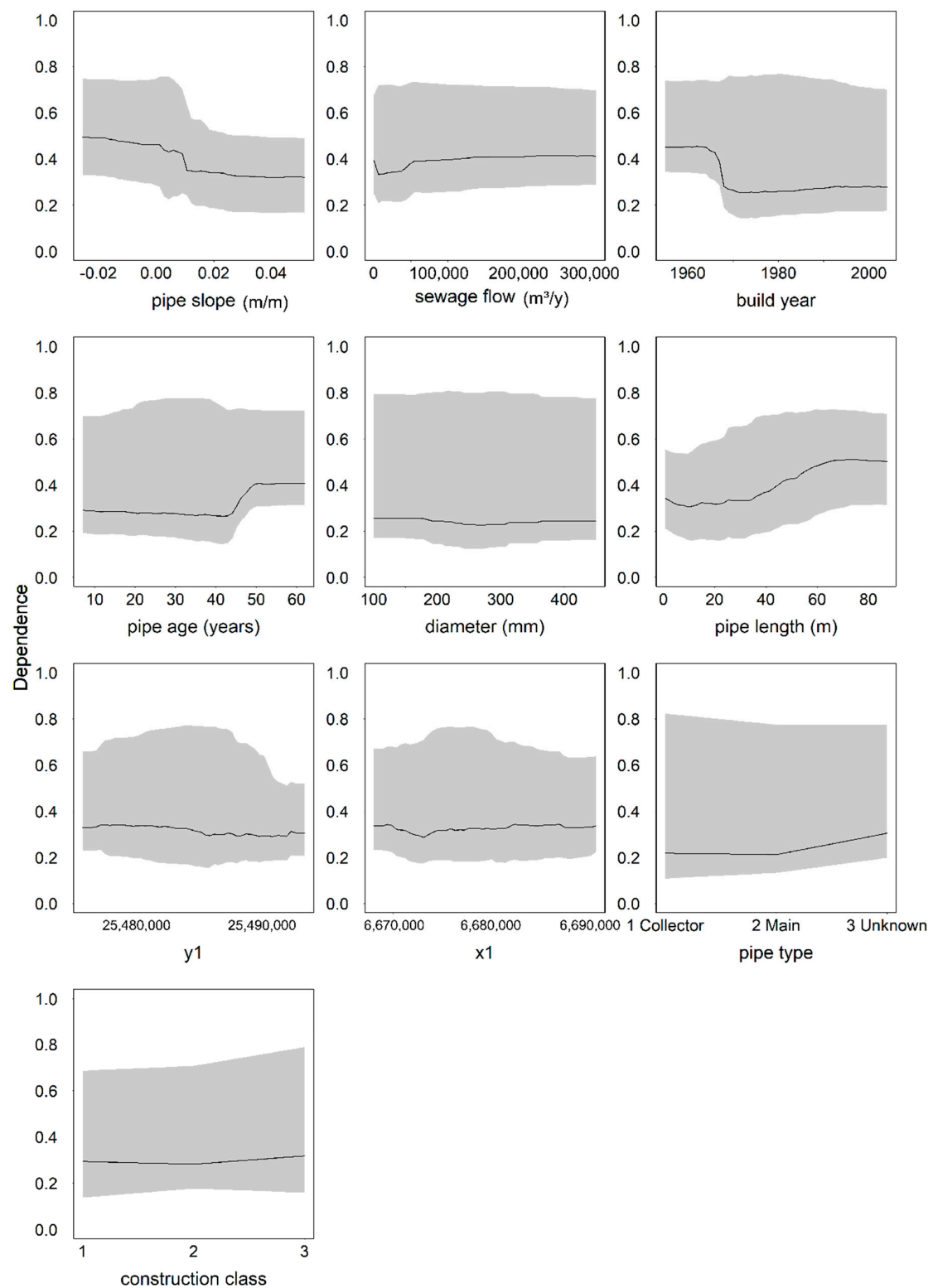
**Table 4.** Variable importance according to the Boruta method.

	Min	Max	Average
Pipe slope	17.26	28.91	23.34
Sewage flow (m <sup>3</sup> /y)	10.18	18.30	13.88
Pipe age (years)	11.14	16.83	13.71
Pipe length (m)	9.37	17.57	12.92
Build year	9.97	15.13	12.60
x1	8.25	12.91	10.48
y1	7.16	11.78	9.72
Construction class	6.34	12.63	9.22
Diameter (mm)	6.44	10.48	8.27
Distance to trees—height 15 to 20 m	3.66	7.84	5.36
Pipe type	3.59	7.38	5.34
Material	3.44	7.18	5.30
Distance to trees—height over 20 m	2.96	8.87	4.86
Depth (m)	2.22	5.95	4.22
Road class	1.52	5.27	3.45
Distance to trees—height 10 to 15 m	0.93	3.99	2.27
Stormwater pipe intersection	0.16	5.82	2.19
Distance to trees—height 2 to 10 m	0.74	3.64	2.10
Waterpipe intersection	-0.95	1.55	0.05

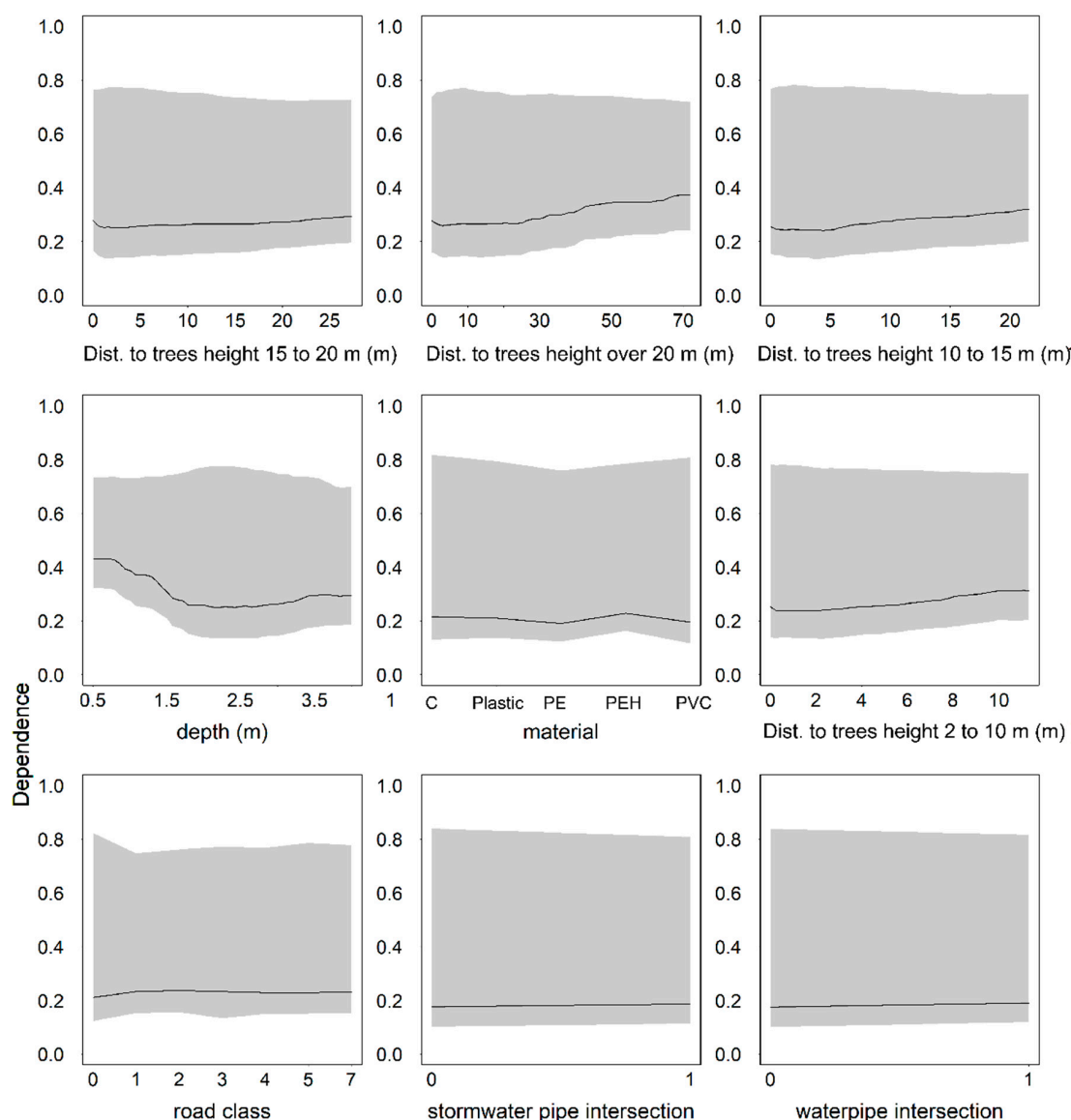
Table 4 shows that according to the Boruta method, almost all the predictor variables influenced poor condition—all but one variable, “waterpipe intersection”, scored an average relevance higher than zero. Additionally, “stormwater pipe intersection” was found uninfluential on some of the rounds performed. Table 4 shows how the variables with the highest importance values are among the highest ones on different rounds and similarly, the variables with the lowest importance remain among the lowest ones, even though the exact order of variables on different rounds may vary. For example, sewage flow scores between 10.18 and 18.30 in importance, which is always higher than the importance of material, which scores between 3.44 and 7.18.

### 3.3. Analysis of Partial Dependence

Figures 2 and 3 present the partial dependence plots for the predictor variables with 25% confidence intervals shown. In Figures 2 and 3, the *y*-axis depicts the vote average in favour of a positive outcome against each predictor variable while the effect of other predictor variables has been averaged out.



**Figure 2.** Partial dependence plots for 10 of the predictor variables.



**Figure 3.** Partial dependence plots for nine of the predictor variables.

The partial dependence plots in Figures 2 and 3 are in line with the results given by the Boruta algorithm: the variables deemed important by Boruta also show greater variability in partial dependence plots. The partial dependence plot on pipe length shows that pipes shorter than 40 m are most commonly in good condition, whereas pipes of length 60 m or more often are in poor condition. Pipe coordinates have an effect on pipe condition, which means that there is spatial variation in the locations of pipes having a poor condition. The plot on the dependence on the pipe age reveals that pipes can be in poor condition irrespective of their age. However, after age of 44 years there is a clear tendency towards a poor condition. Beyond the pipe age of 50 years there were very few inspected pipes and the maximum age of an inspected pipe in the data set was 56. Of all pipes in the data set, 93% have an annual sewage flow lower than 200,000 m<sup>3</sup>/year, even though there are pipes with a flow higher than 2 million m<sup>3</sup>/year. The pipes with the very highest flows have been excluded from the plot in Figure 2 to improve readability. The pipes with the lowest sewage flow have a higher tendency to poor condition. Apart from these pipes, there is a connection between higher flows and poorer condition. This is to be expected as the higher the flow is, the greater the physical stress the pipe is likely to be exposed to. The plot on how the pipe slope relates to a poor condition shows that for pipes with a slope slightly above zero, the likelihood of a poor condition is low. This is the

region where pipe slopes most likely are correct. The likelihood of defects increases dramatically if the slope is negative. Defects have the lowest dependence with pipes built around 1970. Before this point in time, pipes are more often in poor condition. Both the partial dependence plot of pipe age and that of the installation year illustrate that among the inspected pipes, the oldest ones are most often in bad condition. Additionally, there is a slight tendency towards poor condition among pipes that have been installed recently. New pipes experience early failures and defects, which the “bathtub curve” demonstrates. The “bathtub curve” illustrates how, after installation, a subgroup of assets exhibit failures and after this, the failure rate remains constant until it rises again at the end of the asset lifespan. Another possible explanation is that the younger pipes have been selected for inspection because of an operational problem they caused, so this does not necessarily imply newer installations being of a lower quality. Pipes installed at depths between 2 and 3 m are least frequently in poor condition. The minimum installation depth due to frost is 1.6 m for the network, so the results are in line with the guidelines. The dependence on pipe diameter is the lowest for pipes with the largest diameters. The plots on pipe distance to tree stand of different heights suggest that within a distance of 5 m from tree stand, the farther the pipe is, the better is its condition. Beyond a 5-m distance, the distance starts to affect in the opposite way: that the farther away the pipe is from trees, the more often it is in a poor condition.

The remaining partial dependence plots show only slight connections between the predictor variables and poor condition. Poor condition is slightly more common in the construction class with intermediately suitable soil types than in the other two classes. The intermediate class, however, contains far fewer data points than the other two, which could skew the results. Materials have only a small impact on condition, concrete and PEH pipes being slightly more often in poor condition than pipes made of other materials. Pipe intersections with stormwater and water distribution networks have very low to no effect on pipe condition. Pipe type shows some tendency towards poor condition but this is for the pipes, whose pipe type is unknown.

The confidence intervals in Figures 2 and 3 are quite wide for most of the variable values, meaning there is a lot of variation in the impact of specific variable values have on pipe condition. In general, the confidence intervals are narrower for the most important variables.

### 3.4. Model Accuracy with a Reduced Set of Predictor Variables

Not every utility may have the possibility to analyse the impact of all those predictor variables included in this study. In order to assess whether the random forest algorithm could provide useful results with fewer than 19 predictor variables, we carried out an analysis, where the number of predictors was reduced one by one. The predictors were removed one by one starting from the least important predictor as determined by the Boruta algorithm, and the removals were repeated until only the most important predictor remained. The aim was to find out, how many variables would suffice for creating a model with a similar explanatory power as the original, full model. The order was set from the least important to the most important variable to make sure that the variables with the highest predictive power would always be included in the model. We found that eight variables can suffice to achieve an accuracy similar to the complete model. Table 5 presents the results of the analysis.

Table 5 illustrates how an accuracy of 60% can be achieved with the following predictor variables: pipe slope, age, length, build year, sewage flow, y coordinate, construction class, and x-coordinate. Similarly, in order to exceed an accuracy of 50%, one would only need to include three variables: pipe slope, age and length. However, the FPR would then be 65%, meaning that the share of false alerts would be 65% and this may not be desirable.

**Table 5.** Results of the analysis on a reduced set of predictor variables.

Deleted Variable	FPR	Accuracy
Waterpipe intersection	0.52	0.62
Distance to trees—height 2 to 10 m (m)	0.56	0.59
Stormwater pipe intersection	0.56	0.59
Distance to trees—height 10 to 15 m (m)	0.56	0.59
Distance to trees—height over 20 m (m)	0.55	0.6
Material	0.55	0.6
Road class	0.58	0.59
Depth (m)	0.52	0.62
Pipe type	0.54	0.61
Distance to trees—height 15 to 20 m (m)	0.56	0.59
Diameter (mm)	0.55	0.6
x1	0.56	0.6
Construction class	0.59	0.58
y1	0.61	0.57
Sewage flow (m <sup>3</sup> /y)	0.65	0.55
Build year	0.67	0.53
Pipe length (m)	0.65	0.54
Pipe age (years)	0.8	0.46
Pipe slope	NA	NA

### 3.5. Model Application for the Uninspected Network

In addition to model validation against observed data, the model was also tested on the uninspected part of the network. For this purpose, the random forest model was created using six predictor variables (*x*-coordinate, *y*-coordinate, annual sewage flow, material, length, installation year) as these were available for the entire network.

Using the original discrimination threshold of 0.35, the model predicted that 72% of the remaining network was in poor condition and should be inspected. This is a high rate, but since the expected share of false positives was high as well, the actual share can be expected to be clearly lower. Inspecting 72% of the network is very resource intensive; however, the number of pipes to be inspected can be limited adjusting the discrimination threshold. Increasing the threshold value will result in a smaller set of predicted pipes in poor condition but with a higher confidence these pipes being in poor condition. One way of selecting pipes for inspections would be setting the discrimination threshold so that the number of pipes matches the given budget. The downside of increasing the threshold is that higher threshold values result in higher *FNR* (missed cases of pipes in poor condition).

## 4. Discussion

The created model was able to classify correctly approximately 62% of the previously unseen data, when the *FNR* was set to 20%. This is lower than what was hoped for, but in line with the accuracies reported on similar studies. The quality of CCTV inspection data has been found to be deficient in Dirksen et al. [31], where the authors found that approximately 25% of the defects present on a pipe were missed in an inspection. Even though the rate is likely to be lower for observations in the classes indicating most severe defects, which are the focus of this study, uncertainty in inspection results is a potential reason for decreasing the model accuracy. The reasoning behind why a particular pipe was selected for inspections was not recorded in the database, which causes uncertainty on the results. It would be beneficial to know, for example, whether a pipe was inspected because of a functional problem (e.g., blockage) or because the condition of a network area was assessed for a possible renovation. Boosted regression trees outperformed random forest in predicting water pipe failures in Winkler et al. [32]. A potential future research topic is to study, whether boosted regression trees or some other modelling method could provide better results than the logistic regression and random forest tested here. The most important predictor variables were identified using the Boruta

algorithm and partial dependence plots. In the following paragraphs, we discuss the results on the importance of different variables and compare them to those reported in the literature.

Pipe age has been found influential in various studies (e.g., [3,9,10,12,14,16,33,34]). In the network we studied, there was a clear tendency towards poor condition after a pipe age of 44 years. Harvey and McBean [19] found that the chance of being in poor condition was higher for pipes more than 50 years old, whereas in Egger et al. [35], the transition to the last condition phase representing the two worst condition classes took place much later: 50% of pipes reached this phase at the age of approximately 65 years.

Fuchs-Hanusch et al. [11] found that installation year affected pipe condition, whereas in the study by Ana et al. [10] the construction period had no effect. In our study, the partial dependence of pipe condition on the installation year was clear. Inclusion of both the build year and the pipe age was possible because the selected method allowed high correlation between predictor variables.

In our study, pipes beyond 40 m in length were most likely to have at least one defect indicating poor condition. This was an expected result, since in longer pipes potential length for defects is also higher. Additionally, longer pipes are more prone to bending stresses [12]. The studied network also has areas where laterals have been connected directly to sewer mains. Direct connections are a potential cause of structural damage and the longer the pipe is, the more connections it may have. Harvey and McBean [19] found a tendency towards poor condition after a pipe length of 33 m. In contrast, Khan et al. [7] found that pipes over 70 m in length were in good condition more often than shorter pipes, and the results by Baik et al. [36] indicated that longer pipes were less likely to deteriorate than shorter ones.

Installation depth was found to be insignificant for pipe condition in the work of Ariaratnam et al. [9] and Ana et al. [10], and significant in Sousa et al. [16] and Caradot et al. [3]. In the network we analysed, installation depths between 2 m and 3 m were least often connected with poor condition. In contrast, Harvey and McBean [19] found that the chance of being in poor condition was higher for pipes with an installation depth of 1.9 m or more. As Khan et al. [7] pointed out, an increase in depth implies greater dead load over the pipe and a higher probability of ground water table affecting the pipe. However, for the network we studied, the minimum recommended installation depth is 1.6 m due to frost in the winter, which could also cause the difference.

Many studies, e.g., [3,10–12,33,37,38] have found that material affects pipe condition, although some studies have suggested the opposite [9]. Khan et al. [7] found substantial differences in pipes of different types of concrete, and Syachrani et al. [18] found essential differences between the deterioration of vitrified clay pipes and PVC pipes. In our data set, only two of the materials, concrete and PEH, appeared as being more often connected with defects than others. This is in line with the work of Sousa et al. [16], who found differences between only some of the materials they studied. A possible explanation for the slight difference in condition between PEH and other materials is that, according to the utility's experience, the quality of certain batches of PEH pipes has been deficient. Concrete, on the other hand, has been selected for inspections based on the expert experience that it often may be in bad condition in the analysed network.

Pipe diameter had no effect on pipe condition in the study by Ana et al. [10]. In many cases, however, diameter has been identified as a significant factor, e.g., [3,12,33,37]. Khan et al. [7] found smaller diameters to be more stable as compared to the larger diameters. After 600 mm, there was a subtle tendency towards poor condition. Also, Tscheikner-Gratl et al. [34] found pipes with higher diameters to deteriorate more slowly. Harvey and McBean [19] found that the chance of being in poor condition was higher for pipes with a diameter smaller than 238 mm. In our data set, the defects were least common around 1500 mm, and around 300 mm there was a slight tendency towards better condition. According to the utility's experience, pipes with a diameter around 300 mm are correctly dimensioned considering current wastewater flows. The installation work of large-diameter pipes, on the other hand, is often more carefully supervised than that of other pipes. These reasons could explain their better condition.

Pipe slope was found to be insignificant in the works of Ana et al. [10] and Caradot et al. [3], whereas Chughtai and Zayed [12] and Ahmadi et al. [33] found a connection between pipe slope and condition. In our study, the connection was clear. Negative and very low slopes were the most harmful ones. Negative slopes and extremely low slopes cause inadequate rinsing, which can lead to debris accumulation and blockages, whereas steep slopes cause high flow velocities which can cause physical erosion in the pipe walls (e.g., [39]). Our results are in contrast with those by Tscheikner-Gratl et al. [34], who found pipes with steeper slopes to deteriorate more slowly.

Ana et al. [10] found location not to affect the pipe condition. In our study, the incorporation of spatial coordinates ( $x$ - and  $y$ -) into the analysis turned out to be useful: spatial variation took place in the occurrence of defects. Two reasons can explain this. First, the network itself exhibits spatial variation, as pipes are not equally spread over the study area. Second, networks are often built subarea-wise and the sub-areas may differ in, for example, quality of installation work, bedding material, land use, and number and type of water consumers served. However, there were pipes in poor condition all over the network, not concentrated only on specific subareas.

Traffic load was found insignificant in Caradot et al. [3], whereas Chughtai and Zayed [12] and Ahmadi et al. [33] found a connection between road class and pipe condition. In our study, the connection with the road class and pipe condition was weak, as was the connection between soil class and pipe condition. Soil type was found to be significant in Micevski et al. [37] and Savic et al. [14]. Even though the connection with the soil type was very vague, we must note that a fair number of uncertainties were related to the soil type estimate. Our analysis only covered the soil type at the very location of the observed defect based on soil maps that described the soil type at a 1-m depth (originally extrapolated from point measurements by the data provider). Therefore, more research is needed on the impact of soil type. Although intersection with stormwater pipes and water distribution pipes was originally assumed to affect pipe condition, based on the analysis results, they were insignificant.

We studied the effect of the tree stand using a fairly long maximum distance of 100 m. Most likely this is the reason to the somewhat counterintuitive trend seen in the partial dependence plots, where it seems that pipes are in good condition close to trees. However, when focusing on the distances of 5 m or less, we see that pipes within or close to tree stand are more often in poor condition than pipes farther away. Beyond this distance, there are likely to be many other urban structures between the pipe and tree stand, which complicates the connection. The results suggest that in future research, the distances studied should be limited to a couple of meters from the tree stand.

In our study, pipe and network features were more important than variables related to the soil type or to the vicinity to other infrastructure. The significance of pipe location was thus contradictory: the pipe coordinates were important, but factors related to pipe environment were not significant.

The results we gained on variable importance with the Boruta algorithm were in line with the outcomes provided by the analysis of partial dependencies. However, we denote that while the Boruta algorithm is useful for assessing importance, it provides little support for management decisions, since the structure of the connection between the variables and the condition remains unknown. We also found that even though all of the studied predictor variables were regarded as significant by the Boruta algorithm, a similar predictive power was reached using eight of the most relevant predictor variables.

The analysis on partial dependencies provided insight into the connections between the explanatory variables and poor condition. Since most of the articles reviewed for this study do not illustrate the effect of different predictor variables, a detailed comparison was not possible. Chughtai and Zayed [12] present the connection between the condition rating and pipe age for some of the predictor variables, considering average values for other variables. Their analysis, however, covers five condition classes, so the results cannot be compared directly with ours. Similarly, Khan et al. [7] studied the effect of different predictor variables on five condition classes and presented the normalised condition rating as a function of one predictor variable at a time.

## 5. Conclusions

Pinpointing of sewer inspections is a crucial step in sewer asset management, since the inspection results serve as input to renovation decisions. Condition modelling on pipe level is the optimal way of supporting inspection decisions. Random forest provides a useful method for sewer condition prediction. In this study, random forest (accuracy 62% when *FNR* was set to 20%) provided somewhat better results than the binary logistic regression (accuracy 56% when *FNR* was set to 20%). Even though the Boruta algorithm deemed the majority of the 19 studied predictor variables influential, a comparable predictive power was achieved using eight predictor variables. In spite of the fact that the most common defects in the data set—root intrusion, deformation, and surface defect—represent very different defect types, the random forest algorithm was able to capture their occurrence. A future research need is to study whether differences exist in how pipe condition depends on predictor variables, with analysis limited to defects of the same or similar types.

The analysis of partial dependencies proved to be a valuable means of visualizing the connection between predictor variables and poor condition and studying those connections. Although in this study the effect of a relatively large number of variables was analysed, many potentially influential variables could still be investigated. These include for example flow velocity, flow level, wastewater quality, groundwater level, and backfill type. Pipe-specific flow velocity and water level under different weather conditions can be studied using hydrodynamic modelling. For example Savic et al. [40] included them as variables to support asset management decisions. The impact of wastewater quality can be studied with a sensor network set up following the principles presented in Banik et al. [41].

The results of this study are valuable for planning both future condition inspections and new installations. The presented models are meant to be updated iteratively as more inspection results are collected. The greatest benefit of condition prediction is achieved when it is applied as a part of a risk-based sewer asset management program, where the consequences of sewer failure are also considered. Examples of how consequences can be assessed are given in e.g., [42–46].

**Author Contributions:** The work for this article was carried out as follows: Conceptualization, T.L.; Methodology, T.L., T.K., I.M., R.V.; Software, T.L.; Validation, T.L.; Formal Analysis, T.L.; Investigation, T.L.; Resources, R.V.; Data Curation, T.L.; Writing—Original Draft Preparation, T.L.; Writing—Review and Editing, T.K., I.M., R.V. and T.L.; Visualization, T.L.; Supervision, T.K. and R.V.; Funding Acquisition, R.V.

**Funding:** This research received no external funding.

**Acknowledgments:** The authors would like to acknowledge the water utility for providing the data for the study.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Baur, R.; Herz, R. Selective inspection planning with ageing forecast for sewer types. *Water Sci. Technol.* **2002**, *46*, 389–396. [[CrossRef](#)] [[PubMed](#)]
2. Le Gat, Y. Modelling the deterioration process of drainage pipelines. *Urban Water J.* **2008**, *5*, 97–106. [[CrossRef](#)]
3. Caradot, N.; Sonnenberg, H.; Kropp, I.; Ringe, A.; Denhez, S.; Hartmann, A.; Rouault, P. The relevance of sewer deterioration modelling to support asset management strategies. *Urban Water J.* **2017**, *14*, 1007–1015. [[CrossRef](#)]
4. Ana, E.V.; Bauwens, W. Modeling the structural deterioration of urban drainage pipes: The state-of-the-art in statistical methods. *Urban Water J.* **2010**, *7*, 47–59. [[CrossRef](#)]
5. Davies, J.; Clarke, B.; Whiter, J.; Cunningham, R. Factors influencing the structural deterioration and collapse of rigid sewer pipes. *Urban Water.* **2001**, *3*, 73–89. [[CrossRef](#)]
6. Scheidegger, A.; Hug, T.; Rieckermann, J.; Maurer, M. Network condition simulator for benchmarking sewer deterioration models. *Water Res.* **2011**, *45*, 4983–4994. [[CrossRef](#)] [[PubMed](#)]
7. Khan, Z.; Zayed, T.; Moselhi, O. Structural Condition Assessment of Sewer Pipelines. *J. Perform. Constr. Facil.* **2011**, *24*, 170–179. [[CrossRef](#)]
8. Salman, B.; Salem, O. Modeling Failure of Wastewater Collection Lines Using Various Section-Level Regression Models. *J. Infrastruct. Syst.* **2012**, *18*, 146–154. [[CrossRef](#)]

9. Ariaratnam, S.T.; El-Assaly, A.; Yang, Y. Assessment of Infrastructure Inspection Needs Using Logistic Models. *J. Infrastruct. Syst.* **2001**, *7*, 160–165. [[CrossRef](#)]
10. Ana, E.; Bauwens, W.; Pessemer, M.; Thoeys, C.; Smolders, S.; Boonen, I.; De Gueudre, G. An investigation of the factors influencing sewer structural deterioration. *Urban Water J.* **2009**, *6*, 303–312. [[CrossRef](#)]
11. Fuchs-Hanusch, D.; Günther, M.; Möderl, M.; Muschalla, D. Cause and effect oriented sewer degradation evaluation to support scheduled inspection planning. *Water Sci. Technol.* **2015**, *72*, 1176–1183. [[CrossRef](#)] [[PubMed](#)]
12. Chughtai, F.; Zayed, T. Infrastructure Condition Prediction Models for Sustainable Sewer Pipelines. *J. Perform. Constr. Facil.* **2008**, *22*, 333–341. [[CrossRef](#)]
13. Savic, D.; Giustolisi, O.; Berardi, L.; Shepherd, W.; Djordjevic, S.; Saul, A. Modelling sewer failure by evolutionary computing. *Water Manag.* **2006**, *159*, 111–118. [[CrossRef](#)]
14. Savic, D.A.; Giustolisi, O.; Laucelli, D. Asset deterioration analysis using multi-utility data and multi-objective data mining. *J. Hydroinf.* **2009**, *11*, 211–224. [[CrossRef](#)]
15. Tran, D.H.; Ng, A.W.M.; Perera, B.J.C.; Burn, S.; Davis, P. Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. *Urban Water J.* **2006**, *3*, 175–184. [[CrossRef](#)]
16. Sousa, V.; Matos, J.P.; Matias, N. Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition. *Autom. Constr.* **2014**, *44*, 84–91. [[CrossRef](#)]
17. Mashford, J.; Marlow, D.; Tran, D.; May, R. Prediction of Sewer Condition Grade Using Support Vector Machines. *J. Comput. Civ. Eng.* **2011**, *25*, 283–290. [[CrossRef](#)]
18. Syachrani, S.; Jeong, H.S.D.; Chung, C.S. Decision Tree-Based Deterioration Model for Buried Wastewater Pipelines. *J. Perform. Constr. Facil.* **2013**, *27*, 633–645. [[CrossRef](#)]
19. Harvey, R.R.; McBean, E.A. Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure. *J. Hydroinf.* **2014**, *16*, 1265–1279. [[CrossRef](#)]
20. Harvey, R.R.; McBean, E.A. Predicting the structural condition of individual sanitary sewer pipes with random forests. *Can. J. Civ. Eng.* **2014**, *41*, 294–303. [[CrossRef](#)]
21. Agresti, A. *Categorical Data Analysis*; Wiley: Hoboken, NJ, USA, 2013.
22. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
23. Piragnolo, M.; Masiero, A.; Pirotti, F. Comparison of Random Forest and Support Vector Machine classifiers using UAV remote sensing imagery. In Proceedings of the 19th EGU General Assembly, Vienna, Austria, 23–28 April 2017; p. 15692.
24. Liu, M.; Wang, M.; Wang, J.; Li, D. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sens. Actuators B* **2013**, *177*, 970–980. [[CrossRef](#)]
25. Kursu, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [[CrossRef](#)]
26. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
27. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
28. Hastie, T.J.; Tibshirani, R.J.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Series in Statistics; Springer: New York, NY, USA, 2001.
29. Causal Interpretations of Black-box Models. Available online: [https://web.stanford.edu/~hastie/Papers/pdp\\_zhao.pdf](https://web.stanford.edu/~hastie/Papers/pdp_zhao.pdf) (accessed on 20 August 2018).
30. Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [[CrossRef](#)] [[PubMed](#)]
31. Dirksen, J.; Clemens, F.H.L.R.; Korving, H.; Cherqui, F.; Le Gauffre, P.; Ertl, T.; Plihal, H.; Müller, K.; Snaterse, C.T.M. The consistency of visual sewer inspection data. *Struct. Infrastruct. Eng.* **2013**, *9*, 214–228. [[CrossRef](#)]
32. Winkler, D.; Haltmeier, M.; Kleidorfer, M.; Rauch, W.; Tscheikner-Gratl, F. Pipe failure modelling for water distribution networks using boosted decision trees. *Struct. Infrastruct. Eng.* **2018**, *27*, 1–10. [[CrossRef](#)]
33. Ahmadi, M.; Cherqui, F.; De Massiac, J.-C.; Le Gauffre, P. Influence of available data on sewer inspection program efficiency. *Urban Water J.* **2014**, *11*, 641–656. [[CrossRef](#)]
34. Tscheikner-Gratl, F.; Mikovits, C.; Rauch, W.; Kleidorfer, M. Adaptation of sewer networks using integrated rehabilitation management. *Water Sci. Technol.* **2014**, *70*, 1847. [[CrossRef](#)] [[PubMed](#)]

35. Egger, C.; Scheidegger, A.; Reichert, P.; Maurer, M. Sewer deterioration modeling with condition data lacking historical records. *Water Res.* **2013**, *47*, 6762–6779. [[CrossRef](#)] [[PubMed](#)]
36. Baik, H.S.; Jeong, H.S.; Abraham, D.M. Estimating Transition Probabilities in Markov Chain-Based Deterioration Models for Management of Wastewater Systems. *J. Water Resour. Plan. Manag.* **2006**, *132*, 15–24. [[CrossRef](#)]
37. Micevski, T.; Kuczera, G.; Coombes, P. Markov Model for Storm Water Pipe Deterioration. *J. Infrastruct. Syst.* **2002**, *8*, 49–56. [[CrossRef](#)]
38. Duchesne, S.; Beardsell, G.; Villeneuve, J.P.; Toumbou, B.; Bouchard, K. A Survival Analysis Model for Sewer Pipe Structural Deterioration: A sewer deterioration model. *Comput.-Aided Civ. Infrastruct. Eng.* **2013**, *28*, 146–160. [[CrossRef](#)]
39. Butler, D.; May, R.; Ackers, J. Self-Cleansing Sewer Design Based on Sediment Transport Principles. *J. Hydraul. Eng.* **2003**, *129*, 276–282. [[CrossRef](#)]
40. Savic, D.; Djordjevic, S.; Dorini, G.; Shepherd, W.; Cashman, A.; Saul, A. COST-S: A new methodology and tools for sewerage asset management based on whole life costs. *Water Asset Manag. Int.* **2005**, *1*, 20–24.
41. Banik, B.; Alfonso, L.; Di Cristo, C.; Leopardi, A. Greedy Algorithms for Sensor Location in Sewer Systems. *Water* **2017**, *9*, 856. [[CrossRef](#)]
42. Salman, B.; Salem, O. Risk Assessment of Wastewater Collection Lines Using Failure Models and Criticality Ratings. *J. Pipeline Syst. Eng. Pract.* **2012**, *3*, 68–76. [[CrossRef](#)]
43. Kleidorfer, M.; Möderl, M.; Tscheikner-Gratl, F.; Hammerer, M.; Kinzel, H.; Rauch, W. Integrated planning of rehabilitation strategies for sewers. *Water Sci. Technol.* **2013**, *68*, 176. [[CrossRef](#)] [[PubMed](#)]
44. Syachrani, S.; Jeong, H.D.; Chung, C.S. Advanced criticality assessment method for sewer pipeline assets. *Water Sci. Technol.* **2013**, *67*, 1302. [[CrossRef](#)] [[PubMed](#)]
45. Laakso, T.; Ahopelto, S.; Lampola, T.; Kokkonen, T.; Vahala, R. Estimating water and wastewater pipe failure consequences and the most detrimental failure modes. *Water Sci. Technol. Water Supply* **2017**, *18*, 901–909. [[CrossRef](#)]
46. van Bijnen, M.; Korving, H.; Langeveld, J.; Clemens, F. Quantitative Impact Assessment of Sewer Condition on Health Risk. *Water* **2018**, *10*, 245. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).