Predictive and Spatial Analytics for Planning Inspections of Sewer Infrastructure

Richard Harvey[*1], Edward McBean[2]

School of Engineering, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada

[*1]rharvey@uoguelph.ca; [2]emcbean@uoguelph.ca

***Abstract*- Aging sewer pipes inevitably deteriorate to a point where raw, untreated wastewater leaks out of the pipe and into the surrounding soil and nearby sources of groundwater. Municipalities can utilize closed-circuit television (CCTV) inspection technology to identify individual pipes within a sewer network in bad structural condition. Although CCTV inspections provide essential information on pipe condition, they are expensive and often limited to small portions of an entire sewer network. Consequently, any threat to the environment posed by uninspected pipes remains unknown. Predictive analytics can leverage existing inspection datasets so that reliable predictions of condition are available for individual pipes not yet inspected. The predictive capability of the "random forests" data mining algorithm is demonstrated using a case study of sanitary sewer pipe condition data collected by a municipality in Ontario, Canada. A comparison of class-imbalance learning strategies (undersampling and threshold adjustment) is carried out to evaluate the potential to increase predictive accuracy for bad condition pipes representing the minority class of interest. Threshold adjustment is found to provide an optimal level of performance for the classification problem – with the trained predictive model achieving a false negative rate of 18%, false positive rate of 29% and an excellent area under the receiver operating characteristic (ROC) curve of 0.81 (considered to be excellent given the nature of the inspection dataset). An analysis of the predictive capabilities of the random forests algorithm trained using a dataset one-third the size of the one originally available to the case study municipality indicates the algorithm has utility for municipalities outside of the case study area. Network spatial analytics are implemented to visualize clusters of bad condition pipes in the sewer system. Visualization of pipe condition in this manner is found to be an effective tool for screening candidate pipes for inspection and for conveying the necessity of inspection to members of a municipality responsible for approving sewer inspection-related budgets.**

Keywords- Exfiltration; Inspection; Predictive Analytics; Random Forests; Wastewater

I. INTRODUCTION

The majority of North American sewer infrastructure was installed during the period of rapid economic expansion that followed the conclusion of the Second World War. The first phase of this infrastructure is rapidly approaching the end of its useful life and a growing list of evidence suggests many sewers are in an advanced state of disrepair. In the 2013 Report Card for American Infrastructure, a poor condition grade was assigned to the 700,000 miles of publicly owned sewer mains currently in operation [1]. Aging pipes represent the largest capital investment need – comprising three-quarters of the estimated $298 billion of capital investment required over the next twenty years to address wastewater infrastructure [1].

Structural defects in sewer pipes allow raw, untreated wastewater to leak into the surrounding environment. This phenomenon is referred to as exfiltration and is of great concern to municipalities as a range of groundwater pollutants are found in typical municipal wastewater flows [2]. There are numerous historical instances of contamination to validate this concern – *e.g.* a leaking sewer contaminated an aquifer in the Bath region of the United Kingdom in 1928, resulting in a typhoid outbreak affecting 50 people [3] and a total of 54 separate incidents of sewer-related groundwater contamination have been recorded in the United Kingdom alone [2]. Although published data on the rates of exfiltration of raw sewage from aging sewers is rather limited, estimated rates have been established for some international locations – *e.g.* 18 of average daily flow in Munich, Germany [4], 10% of dry weather flow in Rome, Italy [5], and 8% of total pumped volume in Hong Kong [6].

In municipalities with separate sanitary and stormwater networks, the issue of sewage exfiltration is exacerbated when sewage leaking from broken sanitary sewers passes untreated into any broken stormwater pipes located nearby. Sewage contamination of a dedicated stormwater system poses a significant risk to human health as stormwater pipes release their untreated contents to surface waters, such as lakes used for recreational pursuits. A number of recent studies indicate sewage contamination of stormwater systems is nearly ubiquitous in the urban environment. Approximately half of 18 stormwater outfalls studied in Milwaukee, Wisconsin contained more than 25% sanitary sewage by volume and were linked to poor water quality of receiving waters in the area [7]. Individuals swimming in recreational waters located near a stormwater outfall studied in [8] were 50% more likely to experience adverse health effects (*e.g.* gastroenteritis) than those swimming further away from the same outfall. Sewer leaks located near building connections were a significant source of faecal contamination in stormwater systems recently studied in central Singapore [9]. A study conducted in California indicates stormwater pipes act as conduits for raw sewage leaking from failed sanitary sewers, even during periods of dry weather [10].

Raw sewage that flows untreated into surface waters is potentially very dangerous as it may contain a wide range of bacteria, pathogenic microorganisms and endocrine disrupting compounds (EDCs - commonly found in human pharmaceuticals, detergents and chemical disinfectants). These EDCs threaten long-term environmental and human health as they interfere with natural processes in the body responsible for reproduction, development and behaviour. The

pharmaceuticals carbamazepine (an anticonvulsant and mood-stabilizing drug) and sulfamethoxacol (an antibiotic commonly used to treat urinary tract infections) were not completely eliminated in soil columns studied in [11], suggesting they are capable of passing from leaking sanitary sewers to nearby stormwater pipes. Ultimately, chronic faecal contamination of recreational waters ultimately reduces the potential economic benefit of water-based recreation and tourism activities as municipalities are forced to close recreational waters to protect users from illness and disease [12]. Wastewater leaking out of structurally damaged sewer pipe is also capable of contaminating drinking water distribution systems during transient low or negative pressure events, further compounding the potential consequences of aging infrastructure [13].

Until quite recently, most municipalities in North America carried out sanitary sewer-related work primarily as part of basement flooding relief studies (generally limited to qualitative inspections of select sewer pipes representing a small portion of the entire sewer system), as part of roadway rehabilitation projects (where older pipes are dug-up and replaced when new roads are constructed), or in in response to sewer failure or blockages. A major shift in municipal accounting practice has had a dramatic impact on the way sewers are now managed in North America. Historically, tangible assets such as sewer pipes were only recorded as expenditures in their year of installation and the value of existing assets did not appear on annual financial reports. With the introduction of *GASB Statement 34* in the United States in 1999 and *PSAB Statement 3150* in Canada in 2009, capital asset inventory is now reported using accrual accounting methods reflecting the fact that assets such as sewer pipes have a value long after their initial cost of construction is incurred, but this value depreciates over time [14, 15]. These changes in accounting practice have improved the accountability of local government to their citizens as the value of existing sewers are now amortized over their useful life with depreciation recorded as an expense on the municipality's statement of operations. Accurate information needed to satisfy these new legislative requirements is typically obtained by determining the condition of individual pipes in a sewer system using closed-circuit television (CCTV) inspection technology – where a video record of condition is captured using a camera mounted on a remote-controlled robot driven along the length of the pipe. Although CCTV provides valuable information on defects present inside each inspected pipe as shown in Fig. 1, these inspections are expensive and budgetary restrictions force most municipalities to limit inspection-related work to portions of their entire network of sewer pipes.

In this paper, the random forests data mining algorithm is proposed as an efficient means of leveraging information from existing inspection datasets so that predictions of condition can be made for the remaining pipes that have not yet been inspected. As the case study described herein demonstrates, predictions of pipe condition made available by random forest models can be combined with network spatial analytics that seek out clusters of bad condition pipes. The new information made available by this novel application of predictive and spatial analytics can serve as a valuable screening tool for planning future inspection-related work.

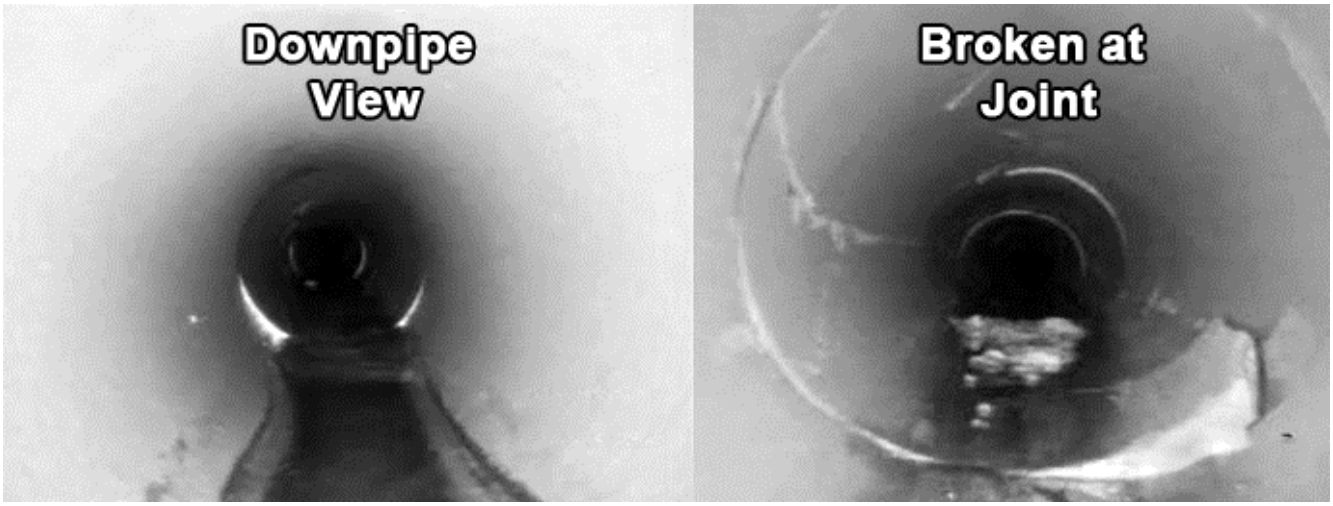

Fig. 1 CCTV technology is used to inspect the internal condition of sewer pipes currently in operation

II. AN OVERVIEW OF MODELS DEVELOPED TO PREDICT SEWER PIPE CONDITION

The majority of sewer pipe deterioration models have been based on statistical theory, with the output typically being a binary response (e.g. a yes or no answer to the question of whether or not a pipe will fail), categorical responses, or a matrix of probabilities for the transition of pipes between condition states. No goodness-of-fit tests were provided for logistic regression models developed in [16]. Multiple regression models were developed in [17] using a small dataset of pipe condition (22 training and 7 for testing asbestos cement; 79 training and 30 for testing concrete; 34 training and 16 for testing for PVC). Fitting data-sensitive sample means to a full population mean is inherently problematic, particularly when very few poor condition pipes are available for model development and validation [18]. Reference [19] explored a variety of regression models for predicting the condition of sewer pipes in Cincinnati, Ohio - binary logistic regression had the highest prediction accuracy with a correct prediction rate of 46% for structurally poor pipes. Reference [20] used 45 km of inspection data

collected in Niagara Falls, Ontario to develop network level logistic models for predictring the probability of concrete and vitrified clay pipes being in various condition states. A linear regression model developed in [21] was believed to greatly under-estimate the length of deficient pipe in a California sewer system. Reference [22] used Markov-chain models to describe the deterioration of large combined sewers in Indianapolis, Indiana – limited data made model development problematic. Markov models were used in [23] to estimate surface condition deterioration curves for groups of concrete sewer pipes in Waterloo, Ontario, Canada. Markov models developed to describe the deterioration of groups of stormwater pipes in Australia could only reasonably predict the future condition for cohorts of pipes [24]. Markov models and ordinal regression models developed using 27 km of stormwater pipe inspection data collected by an Australian municipality were useful at the system-level (predictive accuracy of the ordinal regression model for individual pipes was 42%) [25]. Survival analysis was used to predict structural condition at the network-level for sewer pipes in Quebec City, Quebec in [26]. As noted in the state-of-the-art of statistical deterioration methods presented in [27], cohort survival models are difficult to develop (computationally tedious and they require extensive datasets) and tend to underestimate the number of pipes in the poorest condition states.

While statistical techniques have their own unique sets of advantages, they often require assumpions that limit utility of the developed predictive tool. Data mining techniques, drawing on the fields of artificial intelligence and machine learning, can serve as an alternative when the inherent bias and sparseness of inspection datasets prevent statistical model development. Reference [28] investigated the importance of parameters related to sewer pipe deterioration in Pierrefonds, Quebec using artificial neural networks. Neural networks were found to be superior to both Markov models and ordinal regression models when modeling stormwater pipe deterioration in [29]. Reference [30] used support vector machines to predict individual pipe condition in South Australia. References [31, 32] used decision tree classifiers to investigate sewer pipe condition.

In general, there is still a significant need to develop efficient methods for learning from existing inspection datasets so that location-specific predictions of condition can be made for individual pipes that have not yet been inspected. In the majority of cases, existing modelling techniques (e.g. multiple linear regression, logistic regression and Markov-chains) are of limited utility at the individual-pipe level. As such, this manuscript investigates the possibility of reliably predicting individual pipe condition using the "random forests" data mining system. The random forests algorithm consists of growing hundreds of decision tree classifiers and then combining the predictive capabilities of these individual trees into an ensemble, or forest of trees. Each individual tree is grown to a maximum size by first selecting a random subset of predictor inputs (i.e. pipe attributes) to spilt the dataset on and then calculating the best split based on the CART algorithm first described in [33]. Hundreds of these trees are grown in a similar fashion, and classification predictions are made by having each tree in the forest cast a vote and then determining the mode of these votes.

Models consisting of a single decision tree classifier may exhibit high variance and tend to be unstable as their structure will change depending on the instances available for model training. A more powerful approach to modeling deterioration is the random forests system as it has proven to be one of the best performing classification algorithms on a variety of tasks [34]. Random forests are a logical choice for modeling sewer pipe condition as they are efficient for large databases, require minimal parameter tuning, are insensitive to outliers in a dataset, capable of high levels of performance even when faced with class imbalanced datasets, require no assumptions of pipe behaviour over time and are inherently capable of identifying the most important input predictors out of a larger candidate set. Some example applications in other fields of research have included ecology (classification of invasive plant species in [35]), marketing (predicting customer retention in [36]) and genetics (analysis of genome wide association datasets in [37]). Further information on random forests can be found in [38, 39].

III. CASE STUDY

A municipality located in south-western Ontario, Canada was used as a case study area. The 120,000 residents of the municipality rely on a fractured bedrock aquifer as the source of their drinking water. A recent study indicates shallow over-burden thickness and fast groundwater velocities make the aquifer vulnerable to the deterioration of the municipality's aging sanitary sewer system, with 90% of 22 sampled drinking water wells found to contain at least one sewage-derived contaminant and 45% of the wells exhibited human enteric viruses derived from the exfiltration of domestic sewage flows [40].

The municipality's sanitary system increased by approximately 100 km every 15 years after WWII to satisfy the demands of increased residential development. Although the average sanitary sewer pipe in the municipality has only been in operation for 38 years as shown in Fig. 2, approximately 1,900 pipes were installed more than 50 years ago. Pipes are made of a variety of materials, including asbestos cement (11% of system length), concrete (16.1%), PVC (44.6%), reinforced concrete (3.1%), and vitrified clay (25.2%). In general, the oldest pipes in the network are made of vitrified clay and most pipes installed within the past 30 years are PVC.

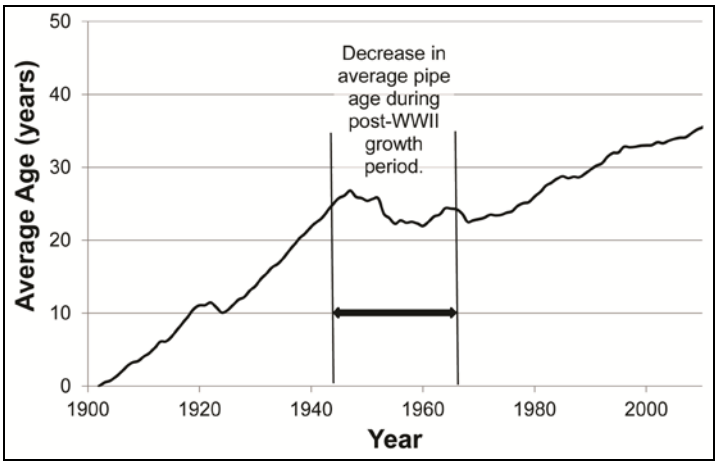

Fig. 2 Change in average sanitary sewer pipe age over time

An engineering consultancy was retained from 2008 - 2011 to help the municipality prepare for compliance with PSAB Statement 3150 by CCTV inspecting a portion of the municipality's more than 7,000 gravity sanitary sewer pipes. A total of 221 km of inspections were completed over this time period, with pipes selected for inspection based on expert opinion that determined those pipes would have an increased likelihood of being in poor structural condition. Structural defects inside each pipe were identified using the third edition of the Water Research Centre Manual of Sewer Condition Classification (WRc MSCC) and severity scores were assigned to defects using the fourth edition of the Water Research Centre Sewerage Rehabilitation Manual (WRc SRM). An internal condition grade (ICG) of 1, 2, 3, 4 or 5 was assigned to each inspected pipe using thresholds established in the WRc SRM for the highest severity scores accumulated in any one meter length of the pipe. The consultancy carried out comprehensive quality assurance/quality control (QA/QC) to ensure accuracy of the collected inspection data. Approximately 90% of the inspected pipes were assigned an ICG of 1, 2, or 3 as shown in Table 1. As a result, the inspection dataset is class imbalanced, with significantly more pipes in some condition classes than others. Class imbalance is common within pipe inspection datasets (as indicated in [19, 41]) and is problematic for predictive modeling as most algorithms will pursue high levels of accuracy by focusing their learning effort on correctly predicting instances belonging to the majority class at the expense of misclassifying minority instances. Fortunately, the data mining community offers a variety of class imbalance learning techniques – the majority designed with the intention of solving two-class problems, as accounting for class imbalance is exceedingly complex beyond two classes.

TABLE 1 INTERNAL CONDITION GRADES ASSIGNED TO INSPECTED PIPES

Material	**Internal Condition Grade**					**Structural Condition**	
	1	**2**	**3**	**4**	**5**	**Good (ICG 1-3)**	**Bad (ICG 4-5)**
Asbestos Cement	373	14	24	5	0	411	5
Concrete	364	79	73	36	7	516	43
PVC	110	9	1	2	1	120	3
Reinforced Concrete	56	8	0	0	0	64	0
Vitrified Clay	213	135	165	121	29	513	150
Total	*1116*	*245*	*263*	*164*	*37*	*1624*	*201*

Pipes were reclassified as being in *good* (ICG 1 – 3) or *bad* (ICG 4-5) structural condition to accommodate the following class-imbalance learning strategies:

- Down-sampling (*i.e.* balanced random forests): where for each tree grown in the random forest, a bootstrap sample of instances is drawn from the minority class and the same number of instances are randomly drawn (with replacement) from the majority class.
- Alternative classification cut-offs: where a random forest is grown but the model is tuned to improve performance on the minority class by identifying a classification cutoff that is more effective than the baseline threshold of 0.50 typically used to convert class probabilities to discrete class predictions [34].

Pipes with an ICG of 4-5 are the minority class of interest as their defects make them susceptible to failure and exfiltration-related issues. An analysis of the inspection data indicates reduced structural integrity was most often the result of cracks, fractures and defective joints as they represent 42%, 38% and 11% of all recorded structural defects, respectively as shown in Fig. 3. This appears to be the case for many Canadian sewer systems, as cracks were previously found to be the most common structural defect in a study of inspection datasets obtained from six other Canadian municipalities in [42].

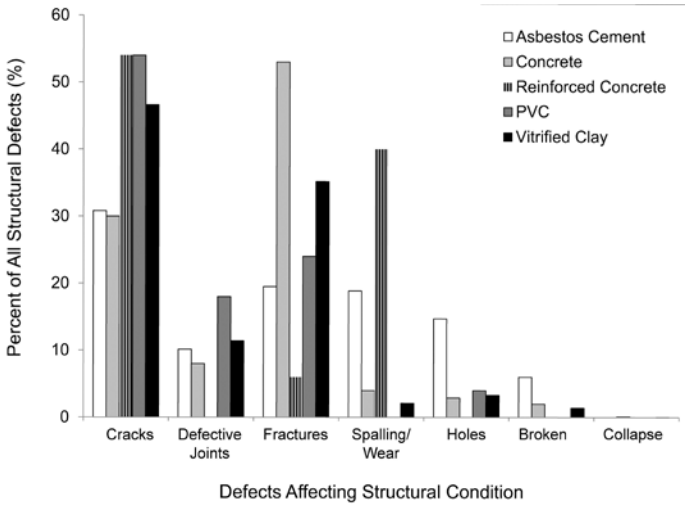

Fig. 3 Distribution of structural defects present in various sewer pipe materials

CCTV inspection records were integrated with existing spatial data in GIS to prevent under-utilization of the data. ESRI ArcMap™ was found to be an excellent document management system that saved considerable time in locating, organizing and confirming the accuracy of field-inspection information. Pipe-specific attributes obtained from GIS used as input predictors for model development include: material of construction (asbestos cement, concrete, PVC, reinforced concrete or vitrified clay), age at the time of inspection (years), installation era (pre-WWI (< 1914), WWI (1914-1918), inter-war (1919-1938), WWII (1939-1945), post-WWII (1946-1966), or modern (1967-present)), sewer type (trunk or branch), diameter (mm), length (m), slope (%), slope change (maximum % change in slope at either the upstream or downstream pipe connection), upstream invert elevation (m), orientation change (maximum change in pipe orientation at either the upstream or downstream connection), burial depth (m), road coverage (portion of a pipe's total length covered by an overlying roadway), number of nearby watermain breaks that have occurred within 3 m of the sewer pipe, land use (agricultural, commercial, industrial, institutional, park or residential), and census tract (one of 27 districts within the municipality pre-determined by government for census purposes). This set of predictor inputs was analysed to ensure there were no near-zero variance predictors, nor was there any strong correlations between predictor variables. Further identification of important input predictors (*e.g.* feature elimination) was unnecessary due to capabilities inherent in the random forests algorithm for such purposes.

Integration of inspection data within a GIS environment afforded an opportunity to identify clusters/hotspots of *bad* condition inspected pipes. Network spatial analysis techniques can be used for cluster identification that establish proximity using the shortest-path distance between events located on the network. A recently introduced third party application for ESRI ArcMap™ known as SANET [43] can be used to efficiently implement network spatial analysis without the burden of excessive mathematical manipulations of the dataset. Fig. 4 presents a hotspot map of *bad* condition inspected pipes developed using SANET (equal-split continuous kernel density function, midpoint of a *bad* pipe considered to be a point of failure, and a shortest path distance between events of 200 m). Hotspot severity can be gauged using threshold values of standard deviation from the mean network density – *e.g.* pipes having a density greater than 2.5 standard deviations above the mean can be considered to be a hotspot of serious concern due to the number of structurally poor pipes located within close proximity.

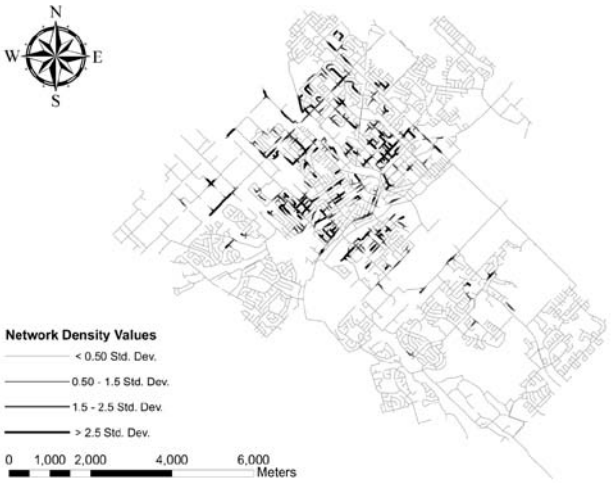

Fig. 4 Density estimation of inspected pipes found to be in bad condition (ICG 4-5)

IV. ASSESSING PREDICTIVE CAPABILITY

Random forests generate both a continuous-valued prediction (in the form of a class membership probability between 0 and 1) and a discrete class prediction based on the class membership probability (using a baseline probability cut-off of 50%). Based on the discrete class predictions, predictive capability was evaluated using the confusion matrix in Table 2 - where bad condition pipes are defined as the positive, minority class of interest for inspection planning purposes. The inspection dataset was partitioned using stratified random sampling into training, evaluation and test sets using a 70-10-20 split ratio.

TABLE 2 CONFUSION MATRIX USED TO EVALUATE PREDICTIVE PERFORMANCE ON A BINARY CLASSIFICATION TASK

		Predicted	
		Bad Pipe [+ class]	**Good Pipe [- class]**
Actual	**Bad Pipe [+ class]**	True Positive (TP)	False Negative (FN)
	Good Pipe [- class]	False Positive (FP)	True Negative (TN)

In addition to overall accuracy (which can be an unreliable indicator of predictive capability when dealing with class imbalanced data), model performance was assessed using the metrics true positive rate, true negative rate, false positive rate, and false negative rate. The area under the receiver operating characteristic (ROC) curve provided further indication of predictive power (with excellent models generally defined as those achieving an area under the ROC > 0.8). The ROC curve, a commonly used tool for evaluating the success of a data mining investigation, is a plot of the true positive rate versus the false positive rate achieved when different probability thresholds are used to establish classifications for instances in a dataset. For each candidate threshold (e.g. 50%) the true positive rate and false positive rate are plotted against each other. More information on the use of these metrics for gauging model performance can be found in [44].

V. RESULTS

Random forests were trained, tuned and tested using the randomForest [45] and caret [46] packages developed for R. Table 3 contains the test set classification results for Random Forest Model 1 (using the entire set of available input-predictors, no class-imbalance learning strategies, the baseline classification cutoff of 0.50, and maximum area under the ROC curve used during model tuning to establish the optimal number of input predictors randomly select to grow each tree in the forest). The confusion matrix indicates class imbalance had the negative consequence of a predictive model that focused almost entirely on correctly classifying good pipes in the majority class in the pursuit of high overall accuracy. The area under the ROC curve achieved during three-repeats of ten-fold cross-validation of the training set was 0.81. Area under the ROC curve for the evaluation and tests sets was 0.82 and 0.81, respectively. These ROC results indicated the model would be capable of being an excellent classifier for bad pipes if the model could be optimally tuned using an alternative cutoff that is more appropriate than the baseline classification cutoff.

Table 4 contains the test set classification results for Random Forest Model 2 (the baseline random forest Model 1 with an optimal cutoff of 0.125 derived from the evaluation set ROC curve). Using this new optimal cut-off, any instance predicted by Model 1 to have a probability of being bad > 0.125 is classified as being bad. Using this new cutoff, the model achieved a true positive rate of 0.82 (up from 0.11), a true negative rate of 0.73 (down from 0.99, but still acceptable), and an overall accuracy of 0.74. As the existing structure of Random Forest Model 2 is the same as Model 1, it has the same excellent area under the ROC curve characteristics.

Table 5 indicates the test set classification results for Random Forest Model 3: where instances in the majority class were down-sampled during model training. This balanced random forest model was found to be less suited to correctly identifying bad condition pipes than Model 2, achieving a test set true positive rate of 0.71 and a test set area under the ROC curve of 0.80.

TABLE 3 TEST SET CONFUSION MATRIX FOR RANDOM FOREST MODEL 1: BASELINE CLASSIFICATION CUTOFF OF 0.50

		Predicted		**Performance Metrics**
		Bad Pipe [+ class]	**Good Pipe [- class]**	
Actual	**Bad Pipe [+ class]**	4	34	True positive rate = 0.11, False negative rate = 0.89
	Good Pipe [- class]	2	316	True negative rate = 0.99, False positive rate = 0.01 Accuracy = 0.90, Area under ROC curve = 0.81

TABLE 4 TEST SET CONFUSION MATRIX FOR RANDOM FOREST MODEL 2: OPTIMALLY TUNED CLASSIFICATION CUTOFF OF 0.125

		Predicted		**Performance Metrics**
		Bad Pipe [+ class]	**Good Pipe [- class]**	
Actual	**Bad Pipe [+ class]**	31	7	True positive rate = 0.82, False negative rate = 0.18
	Good Pipe [- class]	86	232	True negative rate = 0.73, False positive rate = 0.27 Accuracy = 0.74, Area under ROC curve = 0.81

TABLE 5 TEST SET CONFUSION MATRIX FOR RANDOM FOREST MODEL 3: DOWN-SAMPLING OF MAJORITY CLASS

		Predicted		**Performance Metrics**
		Bad Pipe [+ class]	**Good Pipe [- class]**	
Actual	**Bad Pipe [+ class]**	27	11	True positive rate = 0.71, False negative rate = 0.29
	Good Pipe [- class]	75	243	True negative rate = 0.76, False positive rate = 0.24 Accuracy = 0.75, Area under ROC curve = 0.80

VI. DISCUSSION

A. Planning Future Inspections through Combined Spatial and Predictive Analytics

The increased true positive rate and slightly higher area under the ROC curve indicate Model 2 (optimal probability cut-off) performs better than Model 3 (down-sampling) when predicting the minority class of interest in the test set. Overall, the performance metrics obtained for Model 2 indicate it would be capable of providing reliable predictions of condition for uninspected pipes in the sanitary sewer network. Of the more than 4,000 pipes in the system that have not been inspected, Model 2 predicts 1073 will be in bad structural condition. Although the immediate inspection of these 1073 would be preferred, budgetary restrictions will most likely limit the number of pipes that can be inspected within the coming year.

A variety of approaches could be used to plan the next round of inspections. One option would be to consider the individual predicted probability of belonging to the bad pipe class in tandem with risk-of-failure concepts (e.g. WRc critical sewers). It would also be possible to select a portion of the most likely bad pipes as the focus for the next round of inspection (i.e. gain-chart analysis as indicated in [47]). Alternatively, the municipality can use the combined power of the validated predictive model with the information learned from network spatial analytics to establish a subset of pipes for immediate inspection. In essence, this establishes a risk assessment framework whereby future inspections are based on a predicted probability of being bad and a proximity to existing network hotspots, making the consequence of their failure greater than pipes located outside the hotspot. As an example, of the 1073 uninspected pipes predicted to be bad, 52 are within a hotspot established according to a network density value greater than 2.5 standard deviations above the mean density value for the entire network as shown in Fig. 5.

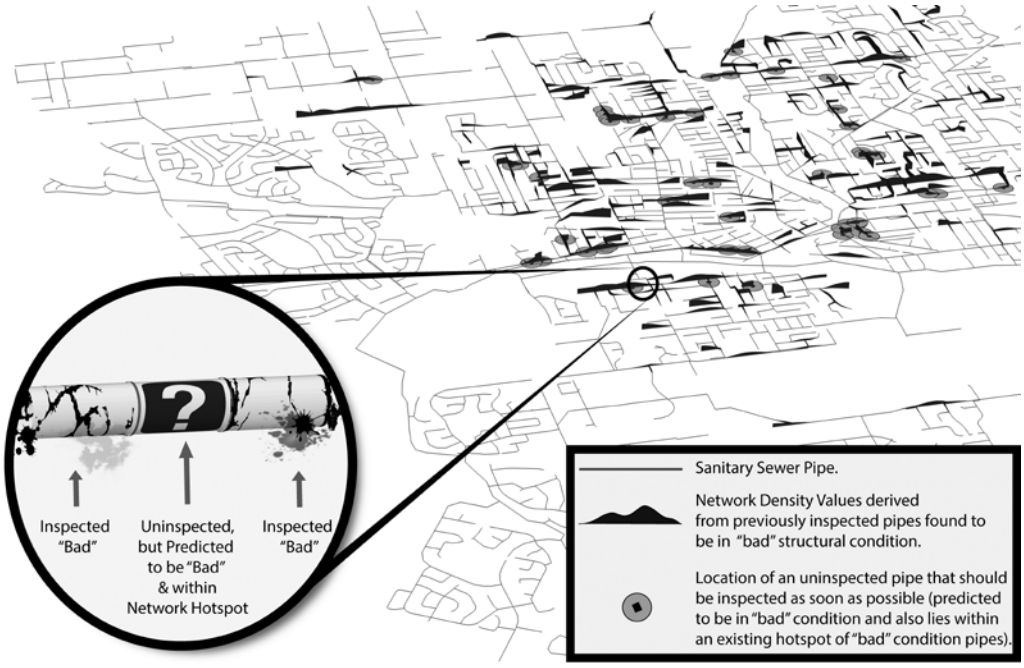

Fig. 5 Predictive and spatial analytics can be combined to identify uninspected pipes that should be targeted for immediate inspection

B. Sensitivity Analysis

A sensitivity analysis was performed to evaluate the impact on the predictive capabilities of the random forests algorithm when using a dataset containing significantly fewer inspections than the one available for analysis by the case study municipality. A scenario was established by using one-third of the inspection dataset (consisting of 531 good pipes and 61 bad pipes) to train a random forest model and then keeping the remaining two-thirds of the inspection dataset as a hypothetical set of 1,195 pipes that have not yet been inspected (but would actually contain 1064 good and 131 bad pipes if they were eventually inspected). The model trained using one-third of the dataset achieved a cross-validated area under the ROC curve of 0.76 and a test set true positive rate and accuracy of 75% and 72%, respectively. Using this trained model to predict the condition of the 1195 pipes in the hypothetical set of uninspected pipes would have resulted in an overall accuracy of 73% (866 pipes correctly classified) and a true positive rate of 79% (104 out of 131 bad pipes correctly identified). This result indicates that although models developed using a significantly a smaller dataset may not have as high a level of performance, trained predictive model can still be of significant utility to a municipality.

C. Inherent Restrictions on Predictive Capability for Individual Pipes

Beyond basic characteristics such as material of construction and age, sewer pipe deterioration rates are likely highly dependent on a variety of factors that are often unavailable for analysis including original design quality control, local

environmental conditions, history of extreme loading events, operating context, and maintenance history since installation. An absence of this information serves to make the characterization of any time-dependent relationship between asset condition and failure difficult. Although drawn from experience within the medical field, reference [48] provides some explanation as to why predictions of individual pipe condition will, most likely, always be associated with some degree of uncertainty. Individual pipe models are similar to models developed in the medical field for individual patients, where even the best models are of little use when predicting survival times for terminally ill patients (with expected error rates usually 50% at best as human survival is so uncertain). In much the same way, the behaviour of individual sewer pipes over time is proving to be difficult to model with the same levels of accuracy as generalized network level models. The individual pipe condition modeling task is further complicated by the nature of the data used for condition classification. CCTV inspections are carried out using standardized systems of defect detection but there is still considerable potential for operator subjectivity or experience to introduce noise into the modeling dataset (i.e. some operators may miss certain types of defects resulting in a less severe ICG score for a pipe).

VII. CONCLUSIONS

Widespread evidence of environmental contamination caused by sewer deterioration and higher expectations for sewer performance both call for a more proactive approach to sewer inspection and management. An analysis of CCTV inspections carried out by a municipality in Ontario, Canada indicates the most common structural defects are cracks, fractures and defective joints, representing 41%, 38% and 11% of all recorded structural defects, respectively. The inspection dataset was found to be class imbalanced - with approximately 90% of all pipes in good structural condition. Network spatial analytics in GIS was found to be a useful tool for identifying clusters/hotspots of inspected sanitary sewer in bad structural condition.

The random forests data mining algorithm was used to predict the structural condition of individual sanitary sewer pipes using a set of input predictors derived using GIS. Two class imbalance learning techniques were utilized to improve the predictive capabilities of the algorithm for bad condition pipes representing the minority class of interest. Although down-sampling (balanced random forests) performed well on the classification task at hand, it was outperformed by the simpler approach of establishing a new classification cut-off for the predictive model – an approach that achieved an excellent area under the receiver operating characteristic (ROC) curve of 0.81, an overall accuracy of 74% and a true positive rate of 82%.

Although accurate predictions of condition of individual pipes within a sewer network are often difficult to achieve (due to class imbalance common in inspection datasets, the uncertain nature of the inspection data, and the considerable variability that can exist in environmental conditions and loading events experienced by individual pipes), the combined application of the random forest algorithm with network predictive analytics was found to be a useful tool for efficiently screening uninspected pipes for future rounds of inspection.

ACKNOWLEDGMENT

The University of Guelph, the Natural Sciences and Engineering Council of Canada and the Canada Research Chairs program funded this research.

REFERENCES

[1] ASCE, America's infrastructure report card, American Society of Civil Engineers, 2013.

[2] P. Bishop, et al., “Impacts of sewers on groundwater quality,” Journal of the Institution of Water and Environmental Management, vol. 12(3), pp. 216-223, 1998.

[3] D. Halliday, Sewers as a source of groundwater pollution in urban areas, Hydogeology Research Group, 1992, University of Birmingham: Birmingham, United Kingdom.

[4] R. Mull, F. Harig and M. Piekle, “Groundwater management in urban areas of Munich,” Journal of the Institution of Water and Environment, vol. 6(2), pp. 199-206, 1992.

[5] A. Cardoso, et al., Assessing the impact of infiltration and exfiltration in sewer systems using performance indicators - case studies of the APUSS project, 10th International Conference on Urban Drainage. 2005, Copenhagen, Denmark. p. 1-8.

[6] D. Lerner, The use of marker species to establish the impact of the City of Nottinghma, UK on the quantity and quality of its underlying groundwater, in Groundwater in the Urban Environment - Problems, Processes and Management, Balkema: Rotterdam, Holland, 1997.

[7] E. Sauer, et al., “Detection of the human specific Bacteroides genetic marker provides evidence of widespread sewage contamination of stormwater in the urban environment,” Water Research, vol. 45(14), pp. 4081-4091, 2011.

[8] R. Haile and J. Alamillo, An epidemiological study of possible adverse health effects of swimming in Santa Monica Bay - final report, Santa Monica Bay Restoration Project: Monterrey Park, CA, USA. p. 1-183, 1996.

[9] J. Doshi, An investigation of leaky sewers as a source of fecal contamination in the stormwater drainage systems in Singapore, in Engineering, Massachusetts Institute of Technology, p. 1-49, 2012.

[10] B. Sercu, et al., “Sewage exfiltration as a source of storm drain contamination,” Environ Sci. Technol., vol. 45(17), pp. 7151-7157, 2011.

[11] G. Cordy, et al., “Do pharmaceuticals, pathogens and other organic waste compounds persist when waste water is used for recharge?” Ground Water Monitoring and Remediation, vol. 24, pp. 58-69, 2004.

[12] NRDC, Testing the Waters - The Impacts of Beach Pollution, Natural Resources Defence Council: New York City, New York, USA,

2012.

[13] P. Teunis, et al., "Enteric virus infection risk from intrusion of sewage into a drinking water distribution network," Environmental Science and technology, vol. 44, pp. 8561-8566, 2010.

[14] GASB, Summary of Statement 34 2012, Governmental Accounting Standard Board.

[15] OMBI, Implementation of Accounting for Tangible Capital Assets - Reference Manual, 2007, Ontario Municipal Benchmarking Initiative: Ontario, Canada.

[16] S. Ariaratnam, A. ElAssaly and Y. Tang, "Assessment of infrastructure inspection needs using logistic models," Journal of Infrastructure Systems, vol. 7(4), pp. 160-165, 2001.

[17] F. Chugthai and T. Zayed, "Infrastructure condition prediction models for sustainable sewer pipelines," Journal of Performance of Constructed Facilities, vol. 22(5), pp. 333-341, 2008.

[18] D. Tran, A. Ng and K. McManus, Practical review - a discussion of deterioration models for stormwater pipe systems in Victoria, Australia, First International Conference on Structural Condition Assessment, Monitoring and Improvement, 2005, Perth, Australia.

[19] B. Salman, Infrastructure management and deterioration risk assessment of wastewater collection systems, in Department of Civil and Environmental Engineering, University of Cincinnati: Ohio. p. 194, 2010.

[20] R. Younis and M. Knight, "A probability model for investigating the trend of structural deterioration in wastewater pipelines," Tunneling and Underground Space Technology, vol. 25(6), pp. 670-680, 2010.

[21] L. Wright, J. Heaney and S. Dent, "Prioritizing sanitary sewers for rehabilitation using least-cost classifiers," ASCE Journal of Infrastructure Systems, vol. 12(3), pp. 174-183, 2006.

[22] R. Wirahadikusumah, et al., "Assessment technologies for sewer system rehabilitation," Automation in Construction, vol. 7(4), pp. 259-270, 1998.

[23] B. Sinha and R. McKim, "Probabilistic based integrated pipeline management system," Tunneling and Underground Space Technology, vol. 22(5), pp. 543-552, 2007.

[24] T. Micevski, G. Kuczera and P. Coombes, "Markov model for stormwater pipe deterioration," Journal of Infrastructure Systems, vol. 8(2), pp. 49-56, 2002.

[25] D. Tran, et al., "Prediction models for serviceability deterioration of stormwater pipes," Structure and Infrastructure Engineering, vol. 4, pp. 287-295, 2008.

[26] S. Duchesne, et al., "A survival analysis model for sewer pipe structural deterioration," Computer-Aided Civil and Infrastructure Engineering, vol. 28(2), pp. 146-160, 2012.

[27] E. Ana and W. Bauwens, "Modeling the Structural Deterioration of Urban Drainage Pipes," Urban Water Journal, vol. 79(1), pp. 1069-1079, 2010.

[28] Z. Khan, T. Zayed and O. Moselhi, "Structural condition assessment of sewer pipelines," Journal of the Performance of Constructed Facilities, vol. 24(2), pp. 170-179, 2010.

[29] D. Tran, Investigations of deterioration models for stormwater pipelines, in School of Architectural, Civil and Mechanical Engineering, Victoria University: Australia. p. 219, 2007.

[30] J. Mashford, et al., "Prediction of sewer condition grade using support vector machines," Journal of Computing in Civil Engineering, vol. 25(4), pp. 283-290, 2011.

[31] J. Jung, et al. Application of classification models and spatial clustering analysis to a sewage collection system of a mid-sized city, International Conference on Computing in Civil Engineering, 2012, Clearwater Beach, Florida: ASCE.

[32] S. Syachrani, H. Jeong and C. Chung, "Decision tree based deterioration model for buried wastewater pipelines," Journal of Performance of Constructed Facilities, vol. 27(5), pp. 633-645, 2012.

[33] L. Breiman, et al., Classification and Regression Trees, 1984: Chapman and Hall/CRC.

[34] M. Kuhn and K. Johnson, Applied predictive modeling, 1st Ed., 2013, New York: Springer Science and Business Media.

[35] Cutler, D., et al., "Random forests for classification in ecology," Ecology, vol. 88, pp. 2783-2792, 2007.

[36] B. Lariviere and D. V. d. Poel, "Predicting customer retention and profitability by using random forests and regression forests techniques," Expert Systems with Applications, vol. 29(2), pp. 472-484, 2005.

[37] B. Goldstein, et al., "An application of random forests to a genome-wide association dataset: methodological consideration and new findings," BMC Genetics, vol. 11(49), pp. 1-13, 2010.

[38] L. Breiman, Random forests, Machine Learning, vol. 45(1), pp. 5-32, 2001.

[39] A. Cutler, D. R. Cutler and J. Stevens, Random Forests, in Ensemble Machine Learning, C. Zhang and Y. Ma, Editors. 2012, Springer US. p. 157-175.

[40] A. Allen, Vulnerability of a fractured bedrock aquifer to emerging sewage-derived contaminants and their use as indicators of virus contamination, in Environmental Science, 2013, University of Guelph: Guelph, Ontario, Canada. p. 158.

[41] E. Ana, et al., Investigating the effects of specific attributes on sewer aging - a Belgian case study, 11th International Conference on Urban Drainage, 2008, Edinburgh, Scotland.

[42] L. Newton and D. Vanier, MIIP Report - The State of Canadian Sewers - Analysis of Asset Inventory and Condition, 2006, National Reseach Council Institute for Research in Construction: Ottawa, Ontario, Canada.

[43] A. Okabe and K. Sugihara, Spatial Analysis along Networks - Statistical and Computational Methods, 2012, United Kingdom: John Wiley and Sons Ltd.

[44] J. Han, M. Kamber and J. Pei, Data mining - concepts and techniques, 3rd ed., 2006: Morgan Kaufmann Publishers.

[45] A. Liaw and M. Wiener, Classification and regression by random forest. R News, 2002. 2/3: p. 18-22.

[46] M. Kuhn, The caret package. 2013 [cited 2013 October 1, 2013]; Available from: http://caret.r-forge.r-project.org/.

[47] R. Harvey and E. McBean, "Predicting the structural condition of individual sanitary sewer pipes with random forests," Canadian Journal of Civil Engineering, vol. 41(4), pp. 294-303, 2014.

[48] R. Henderson and N. Keiding, "Individual survival time prediction using statistical models," Journal of Medical Ethics, vol. 31(12), pp. 703-706, 2005.

Richard Harvey is a PhD candidate in water resources engineering at the University of Guelph (Guelph, Ontario, Canada).

Edward McBean is a Canada Research Chair in Water Supply Security. In that context, he relies upon statistical interpretation of data, fate and transport of chemicals and pathogens in the environment, and risk assessment/ management, to determine how features of water supply risk may arise. Hence, there are dimensions of a number of features including climate change and fate and transport modeling as applied to water resources phenomena. In addition to the above, Ed also has extensive experience in waste management, and greenhouse gas emissions as contributory to global climate change.