Application of Classification Models and Spatial Clustering Analysis to a Sewage Collection System of a Mid-Sized City

I-S. Jung[1], J. H. Garrett Jr.[2], L. Soibelman[3] and K. Lipkin[4]

1 Graduate Research Assistant, Department of Civil and Environmental Engineering, Carnegie Mellon University; email: ijung@andrew.cmu.edu
2 Professor and Head, Department of Civil and Environmental Engineering, Carnegie Mellon University; 5000 Forbes Avenue, Pittsburgh, PA - 15213, PH (412) 268-5674, FAX: -7813; email: garrett@cmu.edu
3 Professor and Chair, Sonny Astani Department of Civil and Environmental Engineering, University of Southern California, 3620 S. Vermont Ave., KAP210, Los Angeles, CA – 90089, PH (213) 740-0609, FAX (213) 744-1426; email: Soibelman@usc.edu
4 Sr. Director of Business Development & Strategy, RedZone Robotics Inc., 91 43rd St, Pittsburgh, PA - 15201, PH (412) 476-8980 x236, email: klipkin@redzone.com

ABSTRACT

Improving asset management of infrastructure systems has been an ongoing issue in the United States. Oliveira et al. (2010, 2011) developed several approaches to better understand the nature and location of pipe breaks in a drinking water distribution system. In this paper, we applied these two approaches to another infrastructure system—the pipe network of a sewage collection system. We first applied several classification approaches to analyze factors associated with higher density regions of deteriorating pipes in the sewage collection system. Relevant attributes that cause poorly conditioned sewer pipes could be found using this approach. We then applied the network version of a density-based clustering algorithm created by Oliveira et al. (2010), used to detect clustered regions of pipe breaks in water distribution systems, to detect hierarchically clustered regions in one of the high density regions of pipe deterioration in the same pipe network. This latter approach was found to provide useful information and additional insight about the local attributes that might be related to the high-density of pipe deterioration.

INTRODUCTION

Improving asset management of infrastructure systems has been an ongoing issue in the United States. Oliveira et al. (2010, 2011) developed several approaches to better understand the nature and location of pipe breaks in a drinking water distribution system. In the analysis described by this paper, we first explored the application of several classification approaches to the data we had related to pipe deterioration problems in a sewage collection system. The goal of this classification analysis was to determine if there are any correlations among the available attributes within the higher density regions of pipe deterioration events, and to extract any relevant factors that are associated with the pipe deterioration process. Then, a

density-based clustering algorithm, adapted for networked domains, was applied to observe if applying the hierarchical clustering approach would provide additional insight into the local attributes that might be related to the higher densities of pipe deterioration events in parts of the sewage collection system.

This paper is organized as follows. The next section briefly presents the approach used in this paper that applies both classification and density-based clustering to better understand the correlations between local attributes and the clusters of pipe deterioration events. The section entitled "Case Study" then discusses our application of this approach to data collected from a specific sewage collection system provided to us by RedZone**,** which develops robotics and software technology to perform condition assessments of wastewater collection systems. This section first describes the data collection process and how the dataset was prepared. In addition, it describes the selection and application of the classification models applied to that data and the application of the network-OPTICS approach. The last section provides conclusions.

OVERVIEW OF APPROACH

There have been numerous approaches that have been explored for developing deterioration models to predict pipe breakage rates based on several factors (Piratla and Ariaratnam 2011). For instance, Clair and Sinha (2011) developed the weighted factor and/or fuzzy logic model to prioritize or predict the performance of water pipes. Chua et al. (2008) developed and evaluated the Proactive Rehabilitative Sewer Infrastructure Management (PRISM) model, which uses a Logistic Regression approach to determine the deficiency probability of sewer pipes. Oliveira et al. (2010, 2011) applied several different classification models, such as C4.5 Decision Tree, Logistic Regression, Boosted Decision Stump, etc. to identify relevant factors associated with abnormal regions of drinking water pipe failures.

In the research described in this paper, two classification models were applied to data provided by RedZone concerning the pipe network of a sewage collection system. During inspection, RedZone uses the Pipeline Assessment and Certification Program (PACP) developed by NASSCO to record the condition of the sewer pipes being inspected. Grades of 4 and 5 are the pipes of 'immediate attention' and 'poor' respectively. Thus, in our dataset, only pipes that are structurally graded as 4 or 5 were labeled as "Pipes with Defects" while the others were labeled as "Pipes with No Defects". Two classification models were applied to determine attributes that can be associated with clusters of pipes with defects and pipes with no defects: 1) a Bayesian Network Approach was applied to explain the distribution of the sewage collection system pipe deterioration data; and 2) Decision Trees were used to identify the relevance of particular attributes associated with the high density regions of pipe defects. This process was not used to indicate any cause-effect relation, but to identify factors that influence pipe deterioration or that may be good predictors of pipe deterioration (Oliveira 2010). After these two classification approaches were applied to the data, we then applied the density-based spatial clustering approach developed by Oliveira et al. (2010), referred to as Network-OPTICS to determine whether more specific relationships between attributes and the pipe defect classes could be discovered.

CASE STUDY

In this section, the process of preparing the dataset is presented. Then, the application of the two classification approaches and the application of Network-OPTICS are discussed in detail.

Data collection. Data used in this study were provided by RedZone Robotics. The data were collected using a Solo(R) robot, an autonomous self-operating inspection device, and RedZone's ICOM3 software to manage this inspection data and link each inspection result to a GIS-represented sewer asset tool. The data used in this research was from one of the cities that RedZone has inspected. The inspection data are recorded as the robot moves along the pipe, so the data contains the distance along each pipe and the corresponding structural condition grades. The data also contains physical characteristics of the pipes, all the information about inspection results, and additional pipe characteristics. Among the available attributes captured in the data, many of them could not be utilized, because they were not relevant for the prediction task or the data set was missing too many values for that attribute. Thus, in the end, only the attributes of pipe diameter and material were able to be drawn from the collected data. Geographical attributes, such as soil type and elevation, were collected from other sources and used. The soil type data was collected from the Natural Resources Conservation Service and the elevation data from the "County" Fiscal Office GIS (2004). We recognize this is an incomplete set of attributes and that if available, more attributes, such as traffic, age, demographics, usage, and many others that are possibly good indicators of deteriorating pipes should also be considered.

Data preparation. The number of defect points in the dataset was comparably very small, because pipes with defects can be thought of as 'abnormal' events that do not occur very often. In addition, every pipe of the network has not yet been inspected by RedZone such that applying the classification models and the clustering algorithm over the entire network was not appropriate to obtain any useful information. Thus, this case study was based on the four different dense areas of the defect points, as shown on the left of Figure1.

Because of this seriously skewed distribution of the dataset, several techniques had to be considered to address this issue. In this analysis, the two most common techniques, i.e., sampling and cost-sensitive classification were examined before applying the frameworks. There are several types of sampling, e.g., over-sampling and under-sampling. In Weka (Hall et al.. 2009), which is a collection of machine learning software tools, there is a sampling method, SMOTE, where the minority class gets over-sampled by creating synthetic examples rather than by over-sampling with replacement (Nitesh et al. 2002). Another technique, called cost-sensitive classification penalizes misclassification of anomalies more than misclassifying normal examples (Neill 2011). Both techniques would cause a bias towards the minority class, but since the number of pipes with defects was comparably very small, at least one of these techniques was necessary to observe any underlying relationships between the attributes and the target class.

Each high-density region has a different number of pipes and defect points, which means a different ratio of the number of members in the majority class to those in the minority class. In addition, selecting the sizes of the datasets was subjective; thus, even though the same technique was used, the number of times this technique was applied to each region had to vary in order to solve the class imbalance problem . In this study, SMOTE was used, because its precision and recall rates were higher than the cost sensitive classification approach. The SMOTE sampling was applied twice to region A (i.e., 200%), three times to B, once to C and twice to D, before applying the classification approaches. The ratio of the positive instances, or pipes with defects, to the negative instances, or pipes with no defects, was then better balanced. While each dataset was ensured to have a smaller number of pipes with defects than that of normal pipes, the percentage difference between the two classes was less than 50%. The skewness of the data tends to cause high false negative rates, resulting in falsely high overall accuracy. Thus, to evaluate classifiers with a skewed class distribution, either precision or recall must be assessed, instead of accuracy. In this case study, a high recall rate would be preferred to a high precision rate, because whenever there are seriously deteriorating pipes, it would be preferable to recognize all such situations.

Application of two classification algorithms to case study data. There are numerous classification algorithms available, so the choice of algorithms would be subjective. However, properties of the data, characteristics of the models, and the objective of the analysis were carefully considered to see which algorithms would be the most appropriate to perform the classification of the deteriorating pipes in the sewage collection system investigated in this case study.

Application of Bayesian Networks. Bayesian networks provide a useful graphical representation of the probabilistic relationships between many variables (Neill 2011). Thus, this algorithm was first applied to see dependencies between the variables and their relationships with the class variables, i.e., "Pipes with Defects" and "Pipes with No Defects" (Weka was used with "global score metrics" and the Hill Climbing search algorithm with a maximum of two parents.)

The Bayes Network created using the approach described above (as shown on the right of Figure 1) showed no direct dependencies between the class variables and the elevation. However, it did show dependencies among the other variables, i.e., the material, the diameter and the soil type. From this observation, it can be concluded that the elevation of the pipe is not a good indicator of pipe deterioration. The only variable that depended on the elevation was the soil type. R. Scharf states in the "Soil and Formation" website that "changes in elevation of only a few feet produce major changes on soil properties of the region, all attributable to the topography's effect on soil water". This statement enhances our result, which showed the dependency between the elevation and the soil type.

In the Bayesian Network, the diameter and the material of the pipe were found to be dependent on each other in every region. For example, in region A, the probabilities of 8 in. vitrified clay pipes (VCP) being in the "Pipe with Defect" and "Pipe with No Defect" classes were 0.764 and 0.957, respectively. Even in the other regions, the probabilities of 8 in. VCP pipes being in these two classes were relatively

higher than the combinations of other diameters and materials. In this dataset, VCP pipes dominate the entire dataset, and so do 8 in. pipes, so it was not surprising to see this result. Thus, we cannot simply conclude that "diameter = 8 in." and "material = VCP" are the most relevant factors in classifying the deteriorating pipes. However, this result would be enhanced if correlated with other variables, e.g., the age of the pipes. For instance, the 8 in. VCP pipes might have been installed in a similar time frame, which means there could be some trends in the installation of particular sizes or types of pipes.

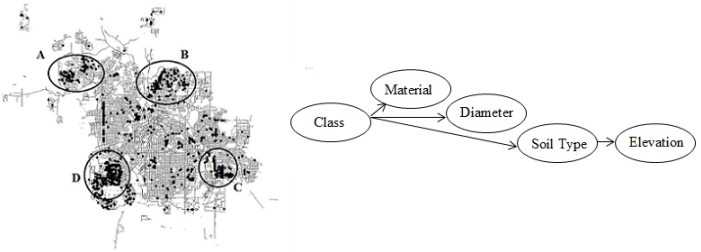

Figure 1. (Left) A network of the wastewater collection system where the points on the network indicate the defects (Right) The Bayes Network graph for Region A

Application of Decision Trees. Decision Tree creation is a widely known classification algorithm that creates a set of rules by assigning each attribute to a decision node of the tree. Depending on the value of the attribute, a different path in the tree is followed towards one class or another. Another advantage of the Decision Tree algorithm is that it is easy and simple to interpret its results. Since we were interested in interpretability of results, a Decision Tree algorithm was appropriate for application to our case study data.

The most widely known Decision Tree algorithms are ID3 and C4.5, created by Quinlan. An extension of ID3, the C4.5 algorithm has a pruning ability, so it was used instead of ID3 (in Weka, the J48 algorithm implements C4.5). The output of the J48 algorithm provides a sense of which variables are important in predicting the target class, because more important variables are located toward the top of the tree while unimportant variables get pruned (Neill 2011). The recall rates of "Pipes with Defects" in the four regions came out to be very similar except for the region C. They were 93.3%, 96.1%, 79.3% and 95.2% respectively. In most of the regions, diameter and/or material seemed to be the most important attributes, because they were the root nodes of the trees. The root node of the region A, for example, was the diameter of 8 in., and the subsequent nodes were the soil types followed by the elevations and materials, respectively. In region A, there were three different soil types that seemed relevant given the 8 in. pipes. Based on Table 1, while the soil type 'S1' has the moderate potential for frost action, 'S2' and 'S3' have high potential. However, it cannot be simply concluded that this soil property is always relevant. Other regions, for example, had soil types that have low potential for frost action. In addition, even though overall properties in the region are still the same, soil properties around the

pipe can be altered due to soil manipulation during pipe installation (Oliveira et al. 2011). There are likely complex combinations of factors that need to be considered before relating properties to deteriorating pipes. Thus, the interpretation of the soil data relating to the pipe deterioration may change as more available attributes are brought into this analysis. Nonetheless, it would be still worthwhile to be aware of the characteristics of the soils when predicting the conditions of the pipes.

Table 1. Examples of Soil Features

Soil Type	Potential for Frost Action	Corrosion Steel	Corrosion Concrete
'S1':	Moderate	Moderate	Moderate
'S2':	High	Moderate	Moderate
'S3':	High	High	High

Application of Network-OPTICS: In this analysis, the Network-OPTICS algorithm (Oliveira et al. 2010) was applied to the pipe deterioration events in region C to determine if different levels of clusters would provide any additional insight on the dataset. Region C was chosen because it had the most balanced distribution of pipes with and without defects out of the four regions, which means there is less bias. As can see be seen from Figure 2, the application of Network-OPTICS to the data from region C generated four levels of clusters, and levels 2 and 3 displayed the most interesting outputs. Since the variations among the physical characteristics of the pipes were very small, only the elevation and the soil data were compared between these two levels. As shown in Figure 2, clusters C9, C13, C5 and C7 are in level 2. Defect points that were part of cluster C7 are no longer part of any cluster at level 3, and several defect points of cluster C5 are also not part of cluster C3 at level 3. Moreover, cluster C13 disappears at level 3, while C9 was further broken down into C8 and C10.

Because the soil type varied so greatly geospatially, this clustering analysis was used to determine if a relationship between the soil types and the clusters at level 3 could be discovered. Figure 3 shows that C7, which is uniquely in level 2, has a different soil type from the clusters in level 3. The higher density clusters, i.e., clusters that remained up to level 3, are located on the same soil type, 'S4', as shown in Figure 3, and cluster, C7 on level 2 is located on soil type, 'S1'. This shows that the clusters in the different levels have different soil type. Thus, even though it might have been a coincidental result, analyzing the soil properties of these two areas and possible changes of their characteristics over time might be useful to determine if particular soil types are more likely to cause these pipes to deteriorate.

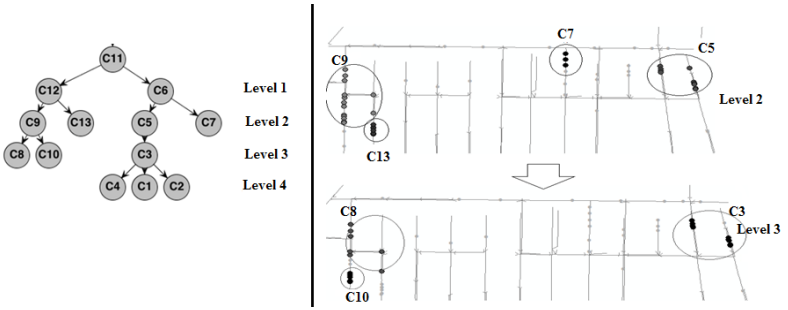

Figure 2. Left: the hierarchical clusters generated by the network-OPTICS for Region C. Right: comparison of the level 2 and level 3 clusters

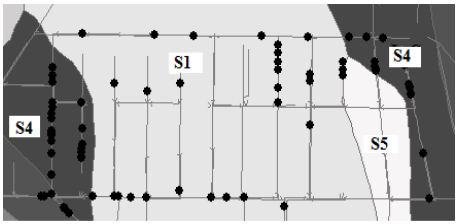

Figure 3. Soil types labeled in the part of the network for Region C

CONCLUSIONS

Bayesian Networks and Decision Trees were created for the data collected from a pipe network in a sewage collection system to see if there were any important local indicators of that deterioration. The performance of these two classification algorithms was assessed using the precision and recall rates and both seemed to perform reasonably well based on these metrics. Relevant factors associated with the pipes with defects, and the dependencies among the attributes, could be found. After the classification analysis, a density-based clustering algorithm was applied to one of the high-density regions of pipe deterioration. This exercise showed the value of examining the hierarchical clusters to get more meaningful associations between more dense clusters and local attributes that might explain the pipe deterioration. These results, however, need to be further explored because of the limited data availability and quality.

ACKNOWLEDGEMENT

The project is funded by a grant from the National Science Foundation (NSF), Grant CMMI0825964. NSF support is gratefully acknowledged. Any opinions, findings, conclusions or recommendations presented in this paper are those of the authors and do not necessarily reflect the view of the National Science Foundation.

REFERENCES

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999) "OPTICS: ordering points to identify the clustering structure." ACM SIGMOD international conference on Management of data table of contents, Philadelphia, Pennsylvania, 49-60.

Bouckaert R. R. (2005), "Bayesian Network Classifiers in Weka", Technical Report, Department of Computer Science, Waikato University, Hamilton, NZ

Chua, K., Ariaratnam, S.T., Ng, H., El-Assaly, A. (2008) "Wastewater Asset Management at the City of Edmonton, Alberta", *Pipelines 2008,* ASCE

Clair, Alison M. St., Sinha, Sunil K. (2011) "Development and the Comparison of a Weighted Factor and Fuzzy Inference Model for Performance Prediction of Metallic Water Pipelines", Pipelines 2011, ASCE

"County" Fiscal Office GIS (2004). "2000 Spot Elevations Annotation" FGDC Content Standards for Digital Geospatial Metadata

de Oliveira, D. P. (2010). Spatial data analysis of networked infrastructure failure data: An application for condition assessment of drinking water distribution systems. Carnegie Mellon University.

de Oliveira, D.P., Garrett, J.H. Jr., Soibelman, L. (2010) "A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage", Advanced Engineering Informatics 25, 380-389

de Oliveira, D. P., Neill D. B., Garrett J. H., Soibelman L. (2011) "Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network", *Journal of Computing in Civil Engineering, Volume 25, Issue 1*, ASCE

Hall, M, Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

NASSCO, PACP-NASSCO*,* http://nassco.org/training_edu/te_pacp.html

Neill, D.B. (2011). Lecture slides for Large Scale Data Analysis course (90-866), Carnegie Mellon University, available at http://www.cs.cmu.edu/~neill/courses/90866.html.

Nitesh V. Chawla et. al. (2002). Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16:321-357.

Oliveira D.P. d., Garrett Jr., J.H., Soibelman, L., (2009) "Spatial clustering analysis of water main break events." ASCE International Workshop on Computing in Civil Engineering, Austin, Texas

Piratla K. R., Ariaratnam S.T. (2011) "Criticality Analysis of Water Distribution Pipelines", Journal of Pipeline Systems Engineering and Practice, ASCE

Scharf, R., SCDNR Land, Water, and Conservation Division "Soil Composition and Formation"http://nerrs.noaa.gov/Doc/SiteProfile/ACEBasin/html/envicond/soil/slform.htm accessed Dec. 23, 2011

U.S. Natural Resources Conservation Service (NRCS) (2011), Soil Survey Geographic (SSURGO) Database for [Summit County, Ohio]. Available online at http://soildatamart.nrcs.usda.gov Accessed [11/07/2011]