Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network

Daniel P. de Oliveira[1]; Daniel B. Neill[2]; James H. Garrett Jr., F.ASCE[3]; and Lucio Soibelman, M.ASCE[4]

Abstract: Infrastructure systems of many U.S. cities are in poor condition, with many assets reaching the end of their service life and requiring significant capital investments. One primary requirement to optimize the allocation of investments in such systems is an effective assessment of the physical condition of assets. This paper addresses the physical condition assessment of drinking water distribution systems by analyzing pipe breakage data as the main source of evidence about the current physical condition of water distribution pipes over space. From this spatial perspective, the primary questions are whether data sets present unexpected clustering of pipe breaks, and where those break clusters are located if they do exist. This paper presents a novel approach that aims to detect and locate clusters of break points in a water distribution network. The proposed approach extends existing spatial scan statistic approaches, which are commonly used for detection of disease outbreaks in a two-dimensional spatial framework, to data collected from networked infrastructure systems. This proposed approach is described and tested in a data set that consists of 491 breaks that occurred over six years in a 160-mi water distribution network. The results presented in this paper indicate that the adapted spatial scan statistic approach applied to points in physical networks is able to detect clusters of noncompact shapes, and that these clusters present significantly higher than expected breakage rates even after accounting for pipe age and diameter. Several possible hypotheses are explored for potential causes of these clusters.

DOI: 10.1061/(ASCE)CP.1943-5487.0000079

CE Database subject headings: Spatial analysis; Statistics; Water distribution systems; Water pipelines; Failures.

Author keywords: Spatial analysis; Scan statistic; Physical network.

Introduction

Most critical infrastructure systems have been rated as in poor condition by ASCE (2009). This fact has raised concerns regarding potential effects of failure and has driven improvements in management practice. The proper assessment of the physical condition of infrastructure assets is a necessary measure required to improve or optimize the management of such systems (Hassanain et al. 2003). Physical condition assessment typically requires the inspection of assets, the derivation of some condition index, and the identification of possible causes of distresses and poor condition.

Infrastructure managers have increasingly sought new means to monitor and assess the overall performance of infrastructure systems, including their physical condition. These means include better sensing and monitoring technologies, which allow infrastructure managers to better assess the condition of individual infrastructure components. However, improvements in data analysis methods are also necessary, in order to make better use of available data and to allow the identification of more general and unforeseen deterioration trends in the system.

Several studies have addressed the analysis of physical condition and deterioration modeling in time (Deb et al. 2002; Kleiner and Rajani 2001). Their main goal is to provide prediction of the time to next failure or remaining service life of a component, which is an input to the optimization of capital investments.

One alternative and complementary approach is to observe the location of failures (or some sensor-based measurement of physical distress) over the extent of a given infrastructure system. In this paper, a water distribution network is considered and the identification of spatial clusters of failures is addressed. Such clusters are regions with an anomalously high number of pipe breakages. The term pipe breakage is used in this paper to describe a set of pipe failure events, including pipe body cracks or splits, joint failures, and hydrant valve failures, which result from deterioration mechanisms, such as internal or external corrosion, and surface loads. These failure events are the ones detected by the utility management and that required a repair record. Such

[1] Project Manager, Office of Campus Planning, The University of Texas at Austin, 5404 Grover Ave., Austin, TX; formerly, Ph.D. Candidate and Graduate Research Assistant, Dept. of Civil and Environmental Engineering, Carnegie Mellon Univ. (corresponding author). E-mail: danielpinhodeoliveira@gmail.com

[2] Assistant Professor of Information Systems, H.J. Heinz III College, School of Public Policy and Management, School of Information Systems and Management, Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: neill@cs.cmu.edu

[3] Professor and Head, Dept. of Civil and Environmental Engineering, Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: garrett@cmu.edu

[4] Professor, Dept. of Civil and Environmental Engineering, Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: lucio@andrew.cmu.edu

Note. This manuscript was submitted on November 16, 2009; approved on May 28, 2010; published online on June 7, 2010. Discussion period open until June 1, 2011; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Computing in Civil Engineering***, Vol. 25, No. 1, January 1, 2011. ©ASCE, ISSN 0887-3801/2011/1-21–30/$25.00.

events do not include undetected small leaks in the system. It is important to note that in July 2008 the utility management started a proactive detection of leaks, which would increase the number of failure events considered as breakage, since they commonly created a repair activity and record. As a result, data from this period were not considered in this paper.

A water distribution pipe network is a specific case of a networked infrastructure system in which exploratory spatial analysis is expected to provide relevant outcomes. The presence of clustering and the location of clusters are primary issues to be addressed in an exploratory spatial data analysis of breakage data (Baddeley 2008; Haining et al. 1998; Smith et al. 2008). The pipes in the vicinity of clusters of pipe breaks are natural candidates for replacement and capital investment planning. Once identified, cluster locations can provide useful information to decision makers, including the identification of the population at risk in such critical areas, and the specification of assets to be considered in capital investment benefit/cost analysis. Also, a possible advantage of cluster detection is the ability to explore each cluster in terms of local factors leading to high failure rate and the changes in these failure rates over time, as discussed in Oliveira et al. (2009).

The process of mining trends in spatially referenced data has been broadly adopted in several fields. One example is the public health field, in which epidemiologists detect clusters of disease cases to assist in identification of disease outbreaks. This paper will demonstrate that the field of infrastructure management can also benefit from such spatial analysis.

The goal of this paper is to present an approach for detecting spatial clusters of pipe breaks in drinking water distribution systems, and to demonstrate its application to an actual breakage data set. More specifically, this approach is based on an adaptation of the spatial scan statistic approach developed by Kulldorff (1997), which detects regions of space with unexpected clusters of events. The proposed approach aims to identify the most interesting regions in a physical infrastructure network, defined as those regions that present higher than expected breakage intensity, perhaps as evidence of some source of distress that might be then controlled. These regions consist of subsets of nearby and connected pipe breaks. The (x,y) coordinates of each pipe break are known, i.e., their location in a two-dimensional (2D) coordinate system, but these breaks are constrained to lie on the underlying network of pipes.

Differently from other areas in which cluster detection methods have been used, e.g., epidemiology, the deferred maintenance existing in most American water distribution systems requires attention not to one single cluster, but rather to a set of significant clusters. Therefore, multiple clusters are of interest when observing the results of cluster detection, rather than only the most interesting region, as in the case of analyzing a disease outbreak.

The natural representation of a water distribution system is, therefore, a planar graph in which edges represent pipes and nodes represent both pipe breaks and intersections. Clusters of breaks are expected to have flexible shape, given the distribution of breaks along pipes and the consequent possibility of noncompact clusters which follow the irregular disposition of pipes in space (as opposed to the circular spatial clusters detected by typical spatial scan statistic approaches). Fig. 1 illustrates such a representation. The proposed approach can be also extended to other networked infrastructure systems, although such extension is not addressed in this paper.

The proposed spatial scan approach for points in a physical network is applied to the data set illustrated in Fig. 2. The data set consist of a system with 268 km of pipes in which 491 breaks occurred in the period from 2002 to 2008 in a small municipality in Western Pennsylvania with an average breakage rate of 1.85 breaks/km.

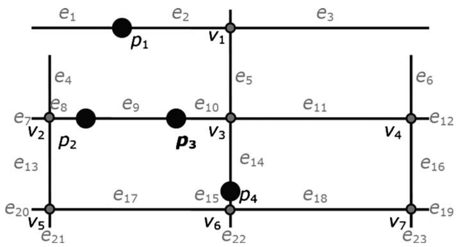


Fig. 1. Representation of hypothetical water distribution system as a graph to be used as a running example in this paper. Edges (e_i) represent pipes while nodes can represent both intersections (v_i) and breaks (p_i). Throughout the paper only break nodes will be represented, intersections being omitted for the sake of clarity.

The remainder of the paper is organized as follows. "Detection of Clusters and Infrastructure Management" describes the relationship between the detection of clusters and water distribution system condition assessment. "Previous Work" discusses previous work on the spatial scan statistic approach, and on spatial analysis of infrastructure systems. "Spatial Scan Statistic Approach" describes the spatial scan statistic approach for point events in a physical network. "Application Results and Discussion" presents the results and discussion of the application of this modified spatial scan statistic approach to the actual data set illustrated in Fig. 2. "Conclusions" presents conclusions and future work.

Detection of Clusters and Infrastructure Management

The goal of this section is to explain the relationship between the detection of clusters and infrastructure management needs (e.g., the physical condition assessment of drinking water distribution systems). In this paper, the random variable of interest is the occurrence of a break event in space, represented as a point constrained to lie on the underlying network of pipes. This point representation is substantially different than the aggregated count

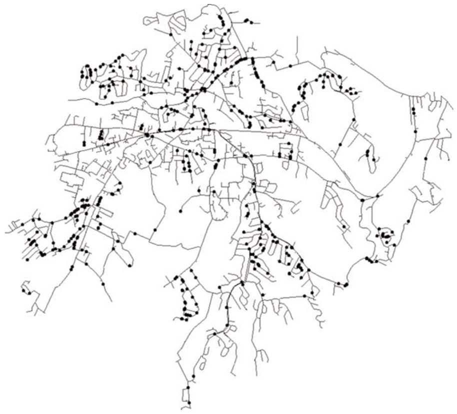


Fig. 2. Drinking water distribution system analyzed in this paper; dots are breaks and lines are pipes

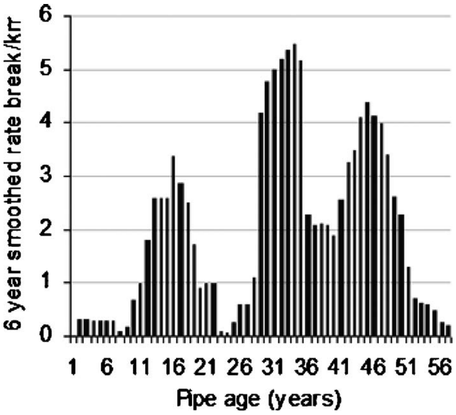


Fig. 3. Plot of pipe age in years versus average number of breaks per kilometer for the data set shown in Fig. 2

data commonly used when applying the spatial scan statistic approach, (e.g., the number of observed disease cases in each zip code).

A cluster is defined as a region (i.e., in the case of a water distribution system, a connected subgraph of the pipe network) in which the density of breaks is significantly higher than expected. The notion of what constitutes a high density of breaks depends on a subjective assessment of the expected number of breaks for a given pipe segment. In the simplest case, assuming a network of homogeneous pipes, the expected number of breaks per unit length is just the ratio of the total number of breaks in the data over the total length of pipes. More generally, the effect of some covariates that can affect the occurrence of the event of interest is often accounted for. For instance, when considering the population of regions, a larger population is more likely to present more cases of a disease compared to a smaller population. In the case of water distribution pipes, the age of a pipe is expected to affect its breakage rate. Generally, older pipes are expected to break more often, although the data used in this paper indicate low breakage rates for pipes over 45 years old, as shown in Fig. 3. This is probably a result of both of the lack of updates in the database, i.e., some pipes were replaced within the period covered by the data set, but were not updated in the geographic information systems (GIS), and the fact that the remaining old pipes are those in less aggressive environments. If the intensity of breaks is still high after adjusting for the expected rate for pipe age, there is evidence that some additional factors must be affecting breakage, and the identification of these factors can be addressed by further investigation.

The primary goal of this work is to identify spatial regions that have an abnormally high density of breaks after controlling for factors that are assumed to affect breakage, e.g., age and pipe diameter in the case of water distribution pipes. These regions are important because by knowing their location, the physical condition of a specific group of pipes can be determined, instead of the physical condition of hypothetical pipe segments of a certain type, e.g., a 40-year old, 6-in. cast iron pipe, to the analysis of the physical condition of a specific group of pipes.

The identification of clusters can provide useful information on: (1) local indicators of physical condition, which can be used to assess benefit/cost analysis of replacement and (2) the linkage of critical pipes with consumer location along the network and the identification of cases in which critical consumers are vulnerable to pipes in critical conditions, which can assist prioritization of maintenance, operation, and replacement decisions.

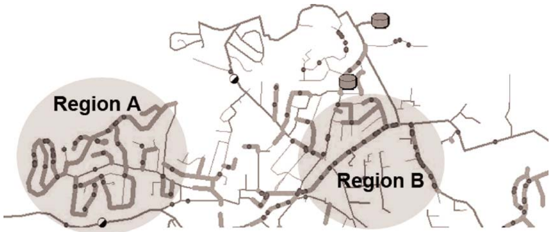


Fig. 4. Two similar groups of pipes (both consisting of 6-in.-diameter, 50-year-old, cast-iron pipes), represented as the dark thicker segments, which have visually different intensity of pipe breaks

The presence of clusters is also evidence that there are dependencies in the data set. These dependencies might result from: (1) the interaction between breaks, i.e., one break might be, for a number of reasons, causing subsequent breaks; and (2) from the nonindependent distribution of environmental factors over space, e.g., high values of soil conductivity, which is the electrical conductivity of an extract from saturated soil paste [U.S. Natural Resource Conservation Service (NRCS) 2007]. Fig. 4 illustrates the case of two regions that are similar in terms of pipe characteristics, but by visual inspection display different breakage intensities. Region B (Fig. 4, right) consists of a dense set of breaks along an elongated extension of 6-in., 50-year old cast iron pipe. Region A (Fig. 4 left) consists of a less dense collection of breaks over a more compact region containing a greater total length of pipe. Later in this paper, such differences will be more rigorously addressed.

While allowing the derivation of local indicators and immediately useful information for infrastructure management decision making, cluster detection is still a tool for exploratory analysis of spatial data. By detecting clusters and identifying the affected locations, the analysis allows the rejection of the hypothesis that breakage is a random process along the network, and to conclude that intensities are significantly higher within clusters. However, this conclusion does not provide an explanation of the actual factors causing clusters. Regarding factors correlated to breakage, the analysis can only conclude that the factors used to adjust the expected rate, e.g., pipe age and size, are not sufficient to explain the high rate within a cluster. "Application Results and Discussion" presents follow-up analyses to examine several other possible hypotheses which may explain the anomalous clusters of pipe breaks.

Previous Work

Relevant research related to this paper falls into two categories: the use of spatial scan statistics for analysis of data in other domains, and the spatial analysis of infrastructure systems, mainly drinking water systems. The former research provides the methodological points of departure for the spatial scan statistics approach presented in the next section, while the latter research discusses and considers alternative approaches in the infrastructure management domain.

The spatial scan statistic approach, originally proposed by Kulldorff (1997), has been frequently used in the analysis of epidemiological data in order to detect outbreaks of disease (Kulldorff et al. 2005; Stevenson et al. 2008). The widely used SaTScan software (SaTScan is a trademark of Martin Kulldorff), which was developed under the joint auspices of Martin Kulldoff, the National Cancer Institute, and Farzad Mostashari at the New York City Department of Health and Mental Hygiene, is based on the spatial scan statistic approach developed by Kulldorff.

A typical use of the spatial scan statistic approach aims to identify the most interesting spatial region(s) in a given larger search area. For example, in a given county or state, a spatial scan statistic approach might be used to identify a subset of zip codes that are indicative of an outbreak of some disease. The algorithm uses a moving window of a given shape (e.g., circular or rectangular) and varying dimensions, which scans the area of interest. In each spatial location, e.g., zip code, some count variable, such as the number of disease cases, is measured and a search is performed for spatial regions (groups of nearby zip codes) with significantly higher than expected counts.

In each step of the spatial scan, the scanning window captures a set of observed realizations of the random variable and generates a score to measure how likely the observed realizations are compared to the expected distribution of this variable. The window with highest score consists of the most interesting subset of locations, as measured by the likelihood ratio statistic described below.

From this process, three issues are important to be emphasized. The first is the search window shape and size, which defines the shape of cluster that can be detected. For instance, a circular scanning window limits the algorithm to detect compact clusters, (i.e., 2D circular or oval clusters defined according to assumed Gaussian distributions), while a rectangular window allows detection of elongated spatial clusters (Neill 2006). As discussed below, neither circular nor rectangular clusters are appropriate for the water distribution network data considered here, since neither of these scan window shapes take the network structure into account.

The second issue is the choice of models for the null hypothesis, i.e., the probability model for the expected outcome in a given window, and the alternative hypothesis, i.e., a probabilistic model for the outcome of interest (e.g., a cluster of disease cases, or of pipe breaks, in some spatial region). These models can be defined, for instance, as either a Bernoulli or Poisson process. For the Poisson case, under the null hypothesis H_0, the breakage rate λ is assumed to be identical at all locations. The alternative hypothesis $H_1(S)$ assumes a breakage rate λ_1 inside region S and λ_o outside region S, where $\lambda_1 > \lambda_0$. Kulldorff (1997) provided a detailed explanation of the derivation of the Bernoulli and Poisson model. In the research described in this paper, a variant of the Poisson model proposed by Neill (2006) is used.

The third issue is the definition of a score $F(S)$ to be calculated for each region S, in order to estimate whether the observed outcome deviates from the expected outcome under the null hypothesis. One typical approach is to compute the likelihood ratio, i.e., the probability of the observed data under the alternative hypothesis, $\Pr[\text{Data}|H_1(S)]$, divided by the probability of the observed data under the null hypothesis, $\Pr(Data|H_0)$

$$F(S) = \frac{\Pr[\text{Data}|H_1(S)]}{\Pr(\text{Data}|H_o)} \tag{1}$$

The problem addressed in this paper requires the search of regions of flexible shape over the edges of a graph. Spatial scan statistic approach that handle searches over regions of flexible shape in graph structures were proposed by Patil and Taillie (2004) and Janeja and Atluri (2008). However, in these approaches, the subjects of interest are the nodes of a graph G, which represent the center of an areal entity with a count attribute, e.g., a county or zip code. This is different from the case of this paper, in which the events of interest are points over the edges of the graph.

Shi and Janeja (2009) described a scan statistics approach that addresses a problem similar to the one addressed in this paper, e.g., the detection of clusters in a physical network, but with a different search strategy and score function. Shi and Janeja (2009) applied their framework to the detection of clusters in traffic accidents. Their approach includes several search strategies that search over subsets of possible subgraphs on a network. The search strategy relies on the linear referencing representation of point events along predefined routes of a network, which is an approach commonly used in geographic information systems for transportation. Such an approach enables any event along a network, e.g., a traffic accident, to be assigned to a mark along a route. The presence of routes enables, in turn, the inclusion of the idea of flow direction along the network, which might be a factor accounting for the occurrence of clusters. Marks are regularly spaced along a route and each mark will hold a random variable of interest, i.e., the count of events that occurred close to it. This is different from considering the location of an event (accident or pipe break) as a random variable, as presented in this paper.

Furthermore, the statistics used by Shi and Janeja (2009) rely on the comparison of rates inside and outside a given region of interest, which are likely to provide overestimated scores and increase false positives in the detection of abnormal regions. The approach described in this paper uses a different score measurement in order to avoid underestimating the expected counts, by controlling for variables of interest, such as pipe age and size. Also, the use of such a search strategy is not adequate in the absence of explicitly defined routes (such as railroads and highways) and flow direction (in the case of water distribution systems, flow direction might change depending on the pressure on some points in the network).

One study has addressed spatial analysis of the physical health of infrastructure assets, including analysis of the space-time clustering of water pipe breaks (Goulter and Kazemi 1998). Goulter and Kazemi analyzed data from Winnipeg by using an ad hoc approach to assess the presence of clusters. For instance, they observed that 22% of breaks occurred within a distance of 1 m of another break. These results, according to the writers, demonstrate the presence of spatial autocorrelation in the breakage process. However, the writers do not indicate how accurate the locations of break points are, nor the potential error in the distance between any two points. It is likely that small clusters of breaks can occur in space and yet, on a larger scale, no major deviations from randomness would be detected. Therefore, a more statistically robust analysis of clustering seems necessary.

Spatial Scan Statistic Approach

The problem informally presented in previous sections can be more formally posed as follows. Given that: (1) a spatial framework represented as a graph $G(V,E)$, in which water distribution pipes are represented as edges in the set $E=\{e_1,e_2,\ldots,e_n\}$, and pipe intersections are represented as a set of nodes $V=\{v_1,v_2,\ldots,v_n\}$; (2) a set P of breaks, $P=\{p_1,p_2,\ldots,p_n\}$, represented as points occurring on the edges E; and (3) models of the null hypothesis H_o and alternative hypothesis $H_1(S)$ for any given region S in graph G; and (4) a likelihood ratio statistic $F(S)$, which allows the deviation of the observed breakage within each

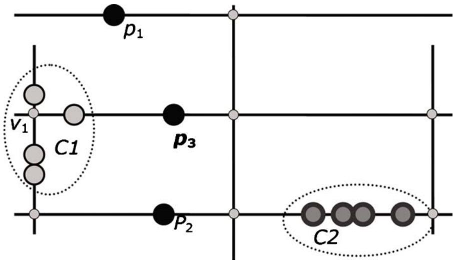


Fig. 5. Illustration of the concept of connected points. Vertices do not interfere in the connectivity between breaks. Break points not assigned to any clusters, (e.g., points $p_1 \cdots p_3$ above) impede the connectivity of the clusters, (e.g., C1 and C2 above).

region from an expected realization of breakage to be evaluated.

The goal of the algorithm is to find all regions S (each consisting of a connected set of break points P_j, as shown in Fig. 5) with abnormally high likelihood ratio scores $F(S) > F_{\text{thresh}}$. The threshold score for a region to be statistically significant, F_{thresh}, is computed using the randomization testing approach described below.

Since break points are located on the segments of graph G, clusters are expected to have elongated rather than compact shapes, following the edges in E. This fact drives the definition of one primary requisite of the algorithm, the capability to detect clusters of complex non-compact shapes, which will be addressed by the definition of an adequate choice of search window.

The method presented in this section follows the framework of the generic scan statistic approach proposed by Neill (2006), consisting of the following steps: (1) acquisition of data consisting of the network structure and the set of spatially referenced break points p_{i}; (2) choice of a set of spatial regions to search over, which are subsets of connected points in P; (3) choice of the models of the data under H_o and $H_1(S)$; (4) definition of a score function $F(S)$ based on H_o and $H_1(S)$; (5) definition of the most "interesting" regions; and (6) assessment of the statistical significance of the most interesting region(s) identified in Step 5. These steps are detailed in the following subsections.

Acquisition of Data for a Set of Pipe Segments e_{i} and Break Points p_i

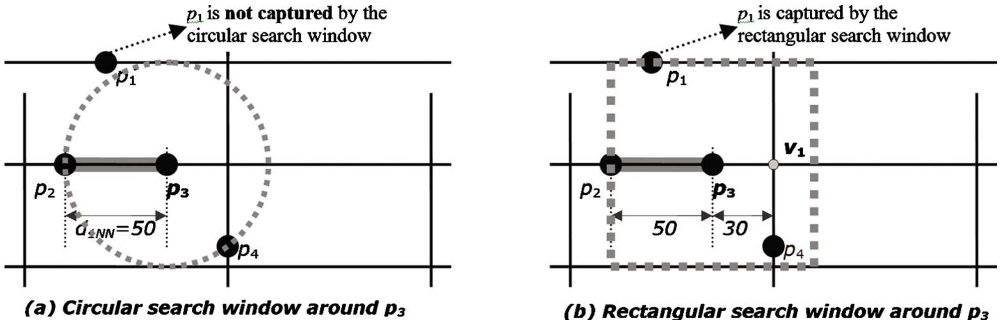
Generally, a breakage data set consists of data points whose main attribute is the pair of (x, y) coordinates, which defines the location of a break on the edge of a segment e_{i}. This location, for the research described in this paper, was estimated by a street address data record, which is taken as an approximation of the true location of break occurrence. Regarding the location of breaks, significant uncertainty is expected from the mapping procedure, which first locates breaks as points over a street segment and subsequently to the pipe that lay under the street. These are procedures that can be performed in standard GIS. Therefore, the location of the break along a pipe segment presents more uncertainty, while the association of the break with a given pipe is less uncertain, as in the data presented in this paper. Pipe age and size, also attributes included in the data set, can be used as covariates, as discussed in the next section.

Choose a Set of Spatial Regions to Search Over

The search strategy adopted in traditional spatial scan statistic-based tools, such as SaTScan, uses circular or rectangular windows. Figs. 6(a and b) illustrate a circular and a rectangular window, respectively, which are centered on a point event p_3. Both windows neglect the fact that the events of interest are constrained by a network space: using a distance metric which does not take the network into account will harm performance by finding false positive "clusters" in areas with a large total length of pipe, and will fail to detect true clusters in areas with a small total length of pipe. Thus an alternative approach is needed, and a new search strategy that relies on shortest path distances between points given the underlying network representation has been developed.

Fig. 7 illustrates a search window over the network space that aims to correctly account for the network constraint on break points. The search strategy is defined by traversing the set P of break points in a graph G. For each point $p_{\text{i}} \in P$, regions are defined by the connected subsets that include point p_i and its k nearest neighbors, where k ranges from 1 to $|P|-1$. For each of these regions, the *length* of the search window is defined by the distance d_{kNN} between p_{i} (the region center) and its kth nearest neighbor (NN) (Fig. 7). The search window includes all edges or portions of edges reachable within a shortest path distance d_{kNN} from p_{i}.

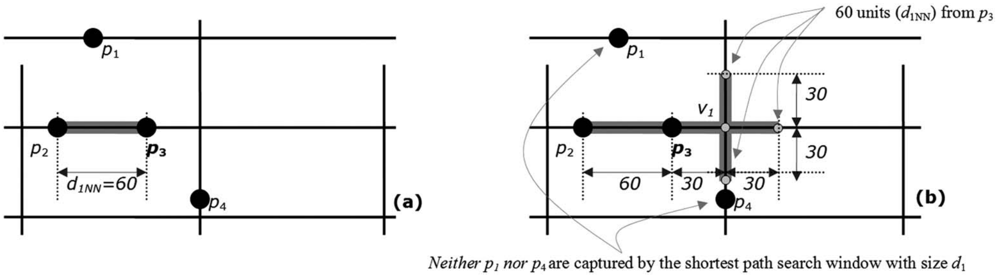
First, for each break point p_i, the shortest path from p_i to each other point (intersections and break points) is found using Dijkstra's shortest path algorithm (Dijkstra 1959). Then for each break point p_i and neighborhood size k, the search steps are: (1) identify the kth nearest neighbor of p_i, which will determine the window size d_{kNN} and (2) find whether each edge e_j is (fully or partially) contained in the search region based on the shortest paths from p_i to the two nodes connected by that edge. Let L_j denote the length of edge e_j, and let d_{ij1} and d_{ij2} denote the shortest path distances


p_1 is not captured by the circular search window
p_1
p_2
p_3
d_{1NN}=50
p_4
p_1 is captured by the rectangular search window
p_1
v_1
p_2
p_3
50
30
p_4

(a) Circular search window around p_3

(b) Rectangular search window around p_3

Fig. 6. Illustration of conventional scan statistics search strategies, given a distance d_{1NN} between p_3 and its nearest neighbor p_2. (a) A circular window of radius d_{1NN}; (b) rectangular window of side $2d_{1NN}$ capture different sets of points around p_3 in 2D Euclidean space.



60 units (d_{1NN}) from p_3

Neither p_1 nor p_4 are captured by the shortest path search window with size d_1

Fig. 7. Search window S defined by a center point p_3 and its nearest neighbor (1NN) p_2. Illustration of the search strategy: (a) computing distance d_{1NN}=60 between p_3 and p_2; (b) search window extending up to the next node (intersection v_1) but not reaching a distance d_{1NN} from p_3. Therefore, the search window extended along all possible paths until it reaches a distance d_{1NN} from p_3 for all paths.

from p_i to the two nodes connected by edge e_j. Then the length of edge e_j contained in region S, $L_j(S)$, can be calculated as $\min[L_j, \max(0, d_{kNN} - d_{ij1}) + \max(0, d_{kNN} - d_{ij2})]$. This computation is illustrated in Fig. 8.

The computation time of the search procedure described above is $O(bn^2)$, where b=total number of break points and n =number of nodes (break points and intersections) in the graph. For each of the b break points, the score function $F(S)$ can be maximized over all regions centered at that break point p_{i} by performing the following steps: (1) perform a single-source shortest path computation to obtain the distance from p_{i} to each node, requiring $O(n^2)$ time; (2) sort the edges by distance to p_{i}, where the graph is planar and there are $O(n)$ edges, and thus this step requires $O(n \log n)$ time; and (3) for each of the b values of the neighborhood size k and the corresponding distance d_{kNN}, find the set of edges contained within a distance d_{kNN} of p_i, and step through them. In the worst case, Step 3 requires $O(n)$ time for each neighborhood size, and thus $O(nb)$ time in total. Thus the complexity of the algorithm is $O(n^2)$ for each of the b break points, giving a total complexity of $O(bn^2)$.

In order to reduce the search time, the number of regions to be searched is reduced by limiting the length of the search window (defined by the distance to the kth nearest neighbor of a given center point) to a maximum of 4.2 km. Clusters for windows longer than 4.2 km are unlikely to satisfy the assumption that the intensity of breaks is homogeneous under the null or alternative hypothesis, due to variation in underlying environmental factors. Nevertheless, this threshold definition potentially limits the detection of clusters that are larger than 8.4 km in diameter.

Choice of Models of the Data under H_o and $H_1(S)$

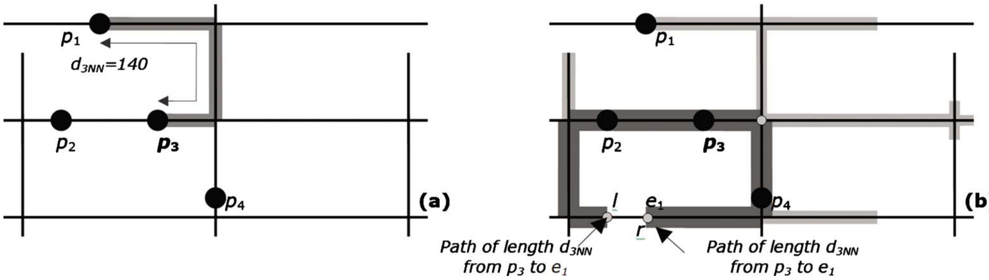
Given a region S captured by the scanning window and formed by a set of connected break points p_i in P, the observed count c_{i} and expected count b_i of breaks for each edge e_i that is fully or partially contained in region S are computed. Then the likelihood ratio test will compare the following hypotheses:

H_o: $c_i \sim \text{Poisson}(b_i)$ for all edges e_i.

$H_1(S)$: $c_i \sim \text{Poisson}(qb_i)$ for all edges (or parts of edges) e_i contained in region S, and $c_i \sim \text{Poisson}(b_i)$ for all edges (or parts of edges) e_i not contained in region S, for some constant $q > 1$.

The expected count b_i of breaks for an edge e_i under the null hypothesis is computed as: $b_i = \lambda_i L_i$, where L_i=length of edge e_i in feet, and λ_{i} is the expected breakage rate per foot of pipe. The spatial distribution of breaks over pipes can be assumed to follow either a homogeneous Poisson process (HPP) or non-HPP (NHPP). For the HPP, an equal breakage rate $\lambda_i = \lambda$ for all pipe segments e_i is assumed, while for the NHPP, the breakage rate λ_i can vary between pipe segments.

Since there are some factors that are known to be correlated to breakage rate, such as age and pipe diameter (Kleiner and Rajani 2001; Pelletier et al. 2003), the breakage rate can be adjusted to different choices of factors. Therefore, several alternatives for assumptions about the underlying process are available, and the NHPP adjusted for pipe age (NHPP_{age}) and adjusted for pipe diameter ($\text{NHPP}_{\text{size}}$) are considered in the research described in this paper. Adjusted breakage rates were calculated for age alone, i.e., λ_{age} for the NHPP_{age}, and for size alone, i.e., λ_{size} for the $\text{NHPP}_{\text{size}}$. Rates were learned from the whole data set available



Path of length d_{3NN} from p_3 to e_1

Path of length d_{3NN} from p_3 to e_1

Fig. 8. Search window S defined by a center point p_3 and its 3rd-nearest neighbor (3NN) p_1. Illustration of the search strategy: (a) computation of distance d_{3NN}=140 from p_3 to p_1; (b) paths from p_3 containing the right side and the left side of edge e_1 (thicker line) and the complete search window with center point p_3 and length d_{3NN}=140.

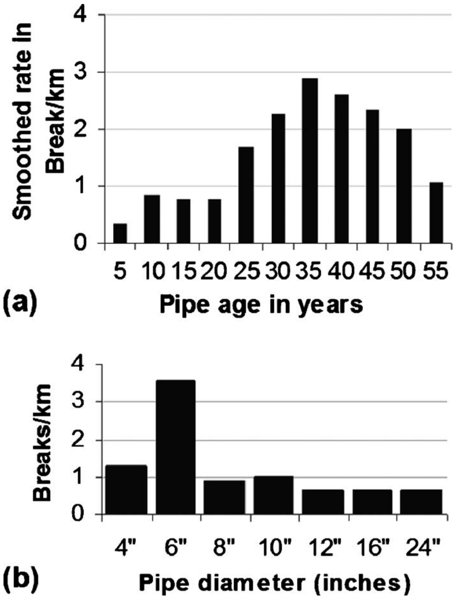


Fig. 9. Breakage rates (breaks/km): (a) smoothed rates by pipe age; (b) by pipe diameter

for the network considered in this paper and smoothed rates were generated and used in the algorithm. Fig. 9 provides the rates for the data set for different pipe attributes.

While breakage rates can potentially be conditioned on any subset of attributes, it is worth noting that as sample sizes in each category are reduced when more attributes are considered, estimates will incorporate more noise and therefore be less reliable. When considering rates for the combined effect of pipe diameter and age, several groups contain few instances. For example, the data set contains only 0.06 km of 15-year old pipes with 10 in. diameter, and thus the two observed breaks in these pipes would produce an unlikely estimate of 33 breaks/km. The sparsity of our data did not allow us to reliably produce smoothed estimates, and thus only separate estimates for pipe diameter and age are used here.

Define a Score Function $F(S)$ based on H_o and $H_1(S)$

The score function to be assigned to each region S is the likelihood ratio $F(S)=\Pr(\text{Data}|H_1(S)/\Pr(\text{Data}|H_o)$, where the null and alternative hypotheses were defined in an earlier section. For the Poisson process considered in this paper, the likelihood ratio statistic is

$$F(S) = \max_{q>1} \frac{\prod_{e_i \in S} \Pr[c_i \sim \text{Poisson}(qb_i)]}{\prod_{e_i \in S} \Pr[c_i \sim \text{Poisson}(b_i)]} \tag{2}$$

This simplifies to the following expression (Neill 2006):

$$F(S) = \frac{\max_{q>1} e^{-qB} q^C}{e^{-B}} = \left(\frac{C}{B}\right)^C e^{B-C} \text{ if } C > B, \text{ 1 otherwise}$$

In this expression, C and B denote the total observed count Σc_{i} and the total expected count $\Sigma b_i = \Sigma \lambda_i L_i$ of region S, respectively, and the maximum likelihood estimate $q=\max(1, C/B)$ has been used. It is important to note that only the edges fully or partially contained in region S are included in these summations, and for partially contained edges, only the length of pipe and the breaks actually contained in region S are included.

Identification of the Most Interesting Regions

The definition of the most "interesting" regions consists of the identification of the region in graph G with highest value of $F(S)$, and any nonoverlapping secondary clusters which also have significantly increased count. It is important to note that small regions, here arbitrarily defined as those with less than four breaks, were not considered as possible cluster candidates. This is a choice based on the assumption that the distances between a small number of breaks can be underestimated due to uncertainty in break location, leading to false positives in region detection. As the minimum number of breaks in one region increases and breaks in different pipes are included, the uncertainty in distances is expected to be reduced.

Assessment of the Statistical Significance of Identified Regions

After searching the set of regions s_i in graph G, a set of no-overlapping regions and their corresponding log-likelihood ratio scores are obtained. However, it is important to consider that high scores can occur just by chance, even when the true distribution of points follows the null hypothesis H_o. Therefore, it is necessary to make an assessment of how often we would expect to see a score as high as or higher than each of the scores in the most interesting regions set. Such assessment can be performed by randomization testing, in which a large number of simulated data sets are generated under the null hypothesis of no clusters, the maximum region score is computed for each simulated data set, and the original region scores are compared to the distribution of simulated maximum scores.

In the research described in this paper, the randomization test consisted of 999 runs of the realization of a process according to the models under the null hypothesis H_o, for each of the three cases (NHPP_{age}, $\text{NHPP}_{\text{size}}$, and HPP). For each simulated data set, breaks were created by traversing the set of pipes and randomly creating breaks for each pipe segment according to the Poisson process. More precisely, for a given edge e_i with length L_i and breakage rate λ_i, the number of breaks for that edge was randomly drawn from a Poisson distribution with mean $b_i = \lambda_i L_i$, and each break was assigned a location on that edge uniformly at random.

The maximum score $F^* = \max_S F(S)$ is computed for each replica data set. Then, to compute the p-value of a given region S from the original data set, its score $F(S)$ is compared to the distribution of simulated maximum scores F^*. If R is the total number of simulated data set, and N is the number of simulated data sets with $F^* > F(S)$, then the p-value is calculated as

$$p\text{-value}(S) = \frac{N+1}{R+1}\text{Eq.} \tag{3}$$

This p-value gives an estimate of how likely the region score is to be generated randomly if the null hypothesis of no clusters of breaks is true. Regions with p-values smaller than the significance level $\alpha = 0.05$ are considered statistically significant, and are reported as significant clusters.

Application Results and Discussion

This section presents the results of the analysis of the previously described water pipe breakage data set by the described spatial scan statistic algorithm. The most significant regions for the HPP are presented in Fig. 10, along with the log-likelihood ratio score and the p-value for each region.

In Fig. 10, the six statistically significant clusters, along with the highest-scoring nonsignificant cluster (Cluster 7, which corresponds to Region A in Fig. 4) are shown for the HPP case. Although several regions could, by visual inspection, be considered possible candidates to form clusters, the most interesting region, i.e., Region 1 in Fig. 10, is visually deceptive, since it appears as only one break in the map (actually, there are six nearby breaks in a very small length of pipe). Other regions, such as Regions 2 and 4, would be intuitively expected to form a single cluster, but are split due to different densities of breakage in the different segments.

These results also demonstrate that the approach is able to differentiate the two areas presented in Fig. 4. In the HPP results presented in Fig. 10, Region 7 (corresponding to area A in Fig. 4) does not provide sufficient evidence that its realization is significantly different from the underlying null hypothesis of a HPP (i.e., its p-value is larger than 0.05), while Regions 2 and 4 (corresponding to Area B in Fig. 4) are found to be statistically significant. This suggests that there is some underlying factor in Area B that is causing the higher than expected intensity of breaks, which is most likely absent in Area A.

Fig. 10 also presents the results for the NHPP assumptions, adjusting separately for pipe age and diameter ($\mathrm{NHPP}_{\mathrm{age}}$ and $\mathrm{NHPP}_{\mathrm{size}}$). The results of the randomization test for the assessment of significance of the likelihood ratio scores is provided in Table 1, which shows the likelihood ratio scores and respective p-values for the six most interesting regions found across the three different processes, obtained through randomization testing with Monte Carlo simulation.

It is interesting to observe that the top five most significant clusters in the HPP case were significant for $\mathrm{NHPP}_{\mathrm{age}}$, while the sixth cluster was not found to be significant. Similarly, four of the six significant clusters in the HPP case were also found to be significant for $\mathrm{NHPP}_{\mathrm{size}}$. These four clusters were significant in all three analyses, which means that, even accounting for the higher breakage rates in old pipes and in smaller pipes, these regions still present an unexpectedly high breakage rate. When accounting for pipe diameter, Regions 4 and 6 in the HPP (Fig. 10) disappear, suggesting that the variation in pipe size explained these deviations from the expected number of breaks. Adjustment for pipe diameter also reduced the scores of Regions 2 and 3, thus accounting partially for their deviation from the expected number of breaks, but was not sufficient to explain their anomalous counts. Interestingly, adjustment for pipe diameter increased the scores of Regions 1 and 5, suggesting that these clusters took place in areas where the pipe size would lead us to predict a low breakage rate.

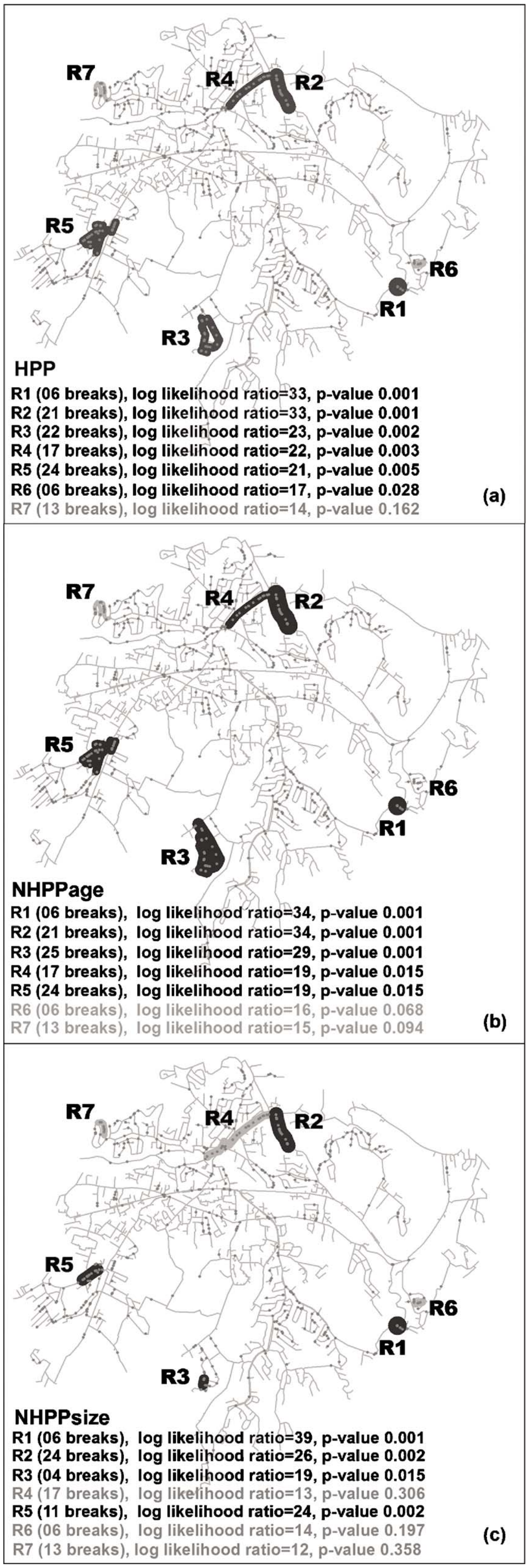

R7
R4
R2
R5
R6
R1
R3
HPP
R1 (06 breaks), log likelihood ratio=33, p-value 0.001
R2 (21 breaks), log likelihood ratio=33, p-value 0.001
R3 (22 breaks), log likelihood ratio=23, p-value 0.002
R4 (17 breaks), log likelihood ratio=22, p-value 0.003
R5 (24 breaks), log likelihood ratio=21, p-value 0.005
R6 (06 breaks), log likelihood ratio=17, p-value 0.028
R7 (13 breaks), log likelihood ratio=14, p-value 0.162
(a)
R7
R4
R2
R5
R6
R1
R3
NHPPage
R1 (06 breaks), log likelihood ratio=34, p-value 0.001
R2 (21 breaks), log likelihood ratio=34, p-value 0.001
R3 (25 breaks), log likelihood ratio=29, p-value 0.001
R4 (17 breaks), log likelihood ratio=19, p-value 0.015
R5 (24 breaks), log likelihood ratio=19, p-value 0.015
R6 (06 breaks), log likelihood ratio=16, p-value 0.068
R7 (13 breaks), log likelihood ratio=15, p-value 0.094
(b)
R7
R4
R2
R5
R6
R1
R3
NHPPsize
R1 (06 breaks), log likelihood ratio=39, p-value 0.001
R2 (24 breaks), log likelihood ratio=26, p-value 0.002
R3 (04 breaks), log likelihood ratio=19, p-value 0.015
R4 (17 breaks), log likelihood ratio=13, p-value 0.306
R5 (11 breaks), log likelihood ratio=24, p-value 0.002
R6 (06 breaks), log likelihood ratio=14, p-value 0.197
R7 (13 breaks), log likelihood ratio=12, p-value 0.358
(c)

Fig. 10. Scan statistic results for: (a) clusters detected under the HPP assumption; (b) clusters detected under the $\mathrm{NHPP}_{\mathrm{age}}$ assumption; and (c) clusters detected under the $\mathrm{NHPP}_{\mathrm{size}}$ assumption

Limitations and Uncertainties

It is important to consider the uncertainties that are present in the analysis, either resulting from the assumptions of the proposed approach or from the data. As indicated before, problems related to the location of breaks are a critical source of uncertainty. Break location problems were handled on a case by case basis in order to reduce errors in the geocoding process. For each break with an inconsistent address, staff members in the local authority that provided the data were contacted in order to clarify such inconsistencies. Therefore, the location of breaks in terms of the pipe

Table 1. Description of Most Significant Clusters of Pipe Breaks under Different H_0 Models

	HPP			NHPP$_{age}$			NHPP$_{size}$		
	Log-likelihood	P -value	Size	Log-likelihood	P -value	Size	Log-likelihood	P -value	Size
R1	32.7823	0.001	6	33.5259	0.001	6	39.3022	0.001	6
R2	32.6398	0.001	21	28.9056	0.001	21	26.194	0.001	24
R3	23.4326	0.001	22	29.629	0.001	25	19.1803	0.014	4
R4	22.2936	0.002	17	19.3669	0.009	17	ns	ns	ns
R5	20.8497	0.005	24	19.2364	0.009	24	24.4094	0.002	11
R6	17.4254	0.031	6	ns	ns	ns	ns	ns	ns

segment associated with the break occurrence is considered to be reliable. However, the location of breaks along the pipe segment is more uncertain.

Another significant source of uncertainty is the estimation of the rates for the null hypothesis model for the different types of processes considered, i.e., NHPP$_{age}$ and NHPP$_{size}$. Due to noise in the data and increases in the variance of estimates for the adjusted rates resulting from smaller data sets, these estimates might not accurately represent the true rates under each process assumption. Finally, one last issue to be mentioned is the uncertainty resulting from the heuristics used in the search procedure, which were critical to the computational feasibility of the proposed search algorithm.

Exploring Possible Hypotheses for the Presence of Clusters

Evaluation of hypotheses regarding the causal factors influencing the observed clusters of pipe breaks is difficult since only observational data are available rather than data resulting from a controlled experiment. Nevertheless, some hypotheses about the factors associated with or potentially related to the causes of breaks can be explored, as presented in this section. Such hypotheses build on domain knowledge in order to search for interesting associations between breaks and attributes.

One hypothesis relies on the assumption that pressure has an important role in the occurrence of breaks. In this case, pressure on pipes is not available in this data set and therefore the elevation at which the breaks occur is taken as a surrogate variable. This assumption is valid for most of the system, but a portion of the northwest portion of the network is known to be a low pressure zone. If the elevation for each break is considered and a histogram of the elevation distribution is built as shown in Fig. 11, it is possible to observe that Region 1 in the HPP (Fig. 10) has very low elevation, corresponding to very high hydraulic pressure. Elevation was 256 m on average for the breaks in the cluster, compared to the overall distribution of elevation in the system as shown in Fig. 11.

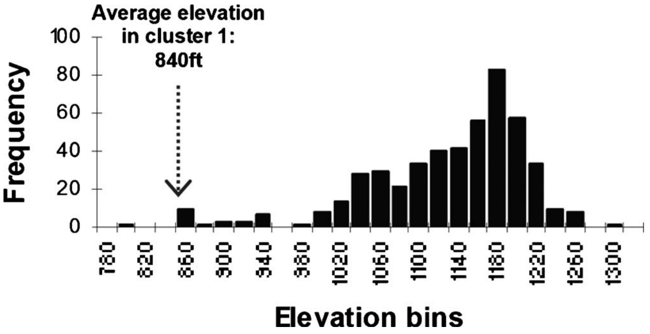


Fig. 11. Histogram of the elevation of break points

A second hypothesis relies on the assumption that surface loads might cause stresses on pipes that eventually lead to breakage. Such loads might be a single event as in the case of a load caused by an activity at a construction site, or a periodic load as in the case of traffic in roads above pipes. Data on unique events were not available for this study, and no reasonable proxy was identified. Periodic loads can result from several sources, and heavy traffic is a major source of load. The locations of bus routes are presented in Fig. 12, and these routes match some of the clusters (Regions 2, 4, and 5). However, there are several route lengths that are not associated with a cluster as well as clusters that do not coincide with bus routes. Therefore, while the overlap between clusters and transportation routes is suggestive of a possible association between these attributes, the data are not conclusive. Moreover, the distress caused by heavy traffic load depends on the depth of pipes and pipe bedding, which are unknown for the present data set. The utility management knows that many older pipes do not have adequate bedding, but there is no data available regarding precisely which pipes have adequate bedding. Pipe depth is unknown and the best information available is that, on average, pipes are 4 feet (1.2 m) deep. Furthermore, construction quality is expected to play a major role, but is not captured in any asset-related database. Other factors might also interact with surface loads and cause local effects, such as soil type, and pipe age and material, and a more detailed follow-up analysis might examine the joint impact of these variables.

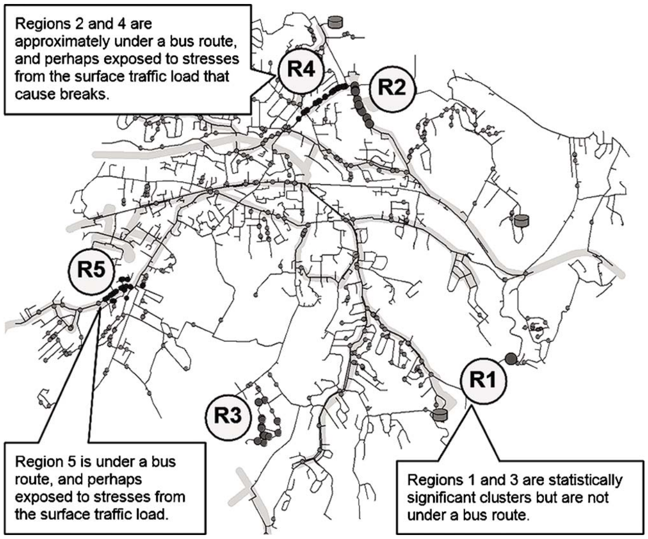

Regions 2 and 4 are approximately under a bus route, and perhaps exposed to stresses from the surface traffic load that cause breaks.
R4
R2
R5
R1
R3
Region 5 is under a bus route, and perhaps exposed to stresses from the surface traffic load.
Regions 1 and 3 are statistically significant clusters but are not under a bus route.

Fig. 12. Location of bus routes and their overlap with regions with abnormally high breakage

Conclusions

This paper presented a spatial scan statistics approach for detecting clusters of point events occurring on the edges in a physical network represented as a planar graph. This approach enabled an exploratory analysis of the occurrence of break events on a water distribution system, and can also be applied to other networked infrastructure assets. After performing this exploratory analysis to detect anomalous patterns of break events, one possible subsequent step is the development of a model to predict breakage in the network while accounting for the presence of clusters. This modeling step, however, is beyond the scope of this paper, though preliminary analysis suggested some possible variables which might be appropriate for such a model.

The novel features of the proposed approach are: (1) the use of spatial analysis techniques to assess breakage data sets, in order to detect regions of high breakage density and (2) the development of the cluster detection approach, building on the spatial scan framework of Kulldorff (1997) and Neill (2006), and its adaptation to the specific challenges of the pipe breakage problem. The results indicate that the adapted spatial scan statistics approach presented in this paper was able to detect potentially useful regions of noncompact shape and to account for the expected effects of pipe age and diameter on the breakage process.

From an asset management perspective, detected regions can be prioritized for maintenance and replacement, and can be used in benefit-cost analysis for capital investments. Additionally, the results presented in this paper are relevant for the insights they provide into factors leading to the observed abnormal breakage rates in the water distribution network data set considered here.

While results indicate that the algorithm is able to detect regions with statistically significant and abnormally high occurrence of breaks, future work will extend the algorithm capabilities to provide: (1) detection of space-time patterns, i.e., emerging clusters, and dynamic changes in cluster shape and size; (2) incorporation of sensing data, particularly leak detection data provided by listening devices known as "correlators," as an extra source of information to assess physical condition of pipes; and (3) multivariate modeling in order to assess the environmental factors associated with interesting regions that might account for abnormal breakage patterns observed within clusters.

Acknowledgments

The project is funded by a grant from the National Science Foundation (NSF), Grant No. CMMI0825964. Additionally, Dr. Neill was funded by NSF Grant Nos. IIS-0916345 and IIS-0911032. NSF support is gratefully acknowledged. Any opinions, findings, conclusions or recommendations presented in this paper are those of the writers and do not necessarily reflect the view of the National Science Foundation.

References

ASCE. (2009). *2009 report card for America's infrastructure*, Reston, Va.

Baddeley, A. (2008). *Modelling spatial point patterns in R (Workshop Notes)*, CSIRO and Univ. of Western Australia, Collingwood, Victoria, Australia.

Deb, A. K., Hasit, Y. J., Schoser, H. M., Loganathan, G. V., and Agbenowsi, N. (2002). *Decision support system for distribution system piping renewal*, AWWA Research Foundation, Denver.

Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs." *Numerische Mathematik*, 1, 269–271.

Goulter, I. C., and Kazemi, A. (1998). "Spatial and temporal groupings of water main pipe breakage in Winnipeg." *Can. J. Civ. Eng.*, 15(1), 91–97.

Haining, R., Wise, S., and Ma, J. (1998). "Exploratory spatial data analysis in a geographic information system environment." *Statistician*, 47(3), 457–469.

Hassanain, M. A., Froese, T. M., and Vanier, D. J. (2003). "Framework model for asset maintenance management." *J. Water Resour. Plann. Manage.*, 17(1), 51–64.

Janeja, V. P., and Atluri, V. (2008). "Random walks to identify anomalous free-form spatial scan windows." *IEEE Trans. Knowl. Data Eng.*, 20(10), 1378–1392.

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

Kulldorff, M. (1997). "A spatial scan statistic." *Comm. Statist. Theory Methods*, 26(6), 1481–1496.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., and Mostashari, F. (2005). "A space-time permutation scan statistic for disease outbreak detection." *PLoS Med.*, 2(3), 216–224.

Neill, D. B. (2006). *Detection of spatial and spatio-temporal clusters*, Carnegie Mellon Univ., Pittsburgh.

Oliveira, D., Garrett, J. H., Jr., and Soibelman, L. (2009). "Spatial clustering analysis of water main break events." *Proc., ASCE Int. Workshop on Computing in Civil Engineering*, ASCE, Reston, Va.

Patil, G. P., and Taillie, C. (2004). "Upper level set scan statistic for detecting arbitrarily shaped hotspots." *Environ. Ecol. Stat.*, 11(2), 183–197.

Pelletier, G., Mailhot, A., and Villeneuve, J.-P. (2003). "Modeling water pipe breaks—Three case studies." *J. Water Resour. Plann. Manage.*, 129(2), 115–123.

Shi, L., and Janeja, V. P. (2009). "Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP)." *Proc., Int. Conf. on Knowledge Discovery and Data Mining*, Paris, 767–776.

Smith, M. J., Goodchild, M. F., and Longley, P. A. (2008). *Geospatial analysis: A comprehensive guide to principles, techniques and software tools*, Winchelsea Press, Leicester, U.K.

Stevenson, M., Stevens, K. B., Rogers, D. J., and Clements, A. C. A. (2008). *Spatial analysis in epidemiology*, Oxford University Press, New York.

U.S. Natural Resource Conservation Service (NRCS). (2007). *Soil data mart—PA003—Allegheny county, Pennsylvania*, USDA—NRCS, Washington, D.C.