Int. J. Environ. Sci. Technol. (2015) 12:2687–2696
DOI 10.1007/s13762-014-0666-0

ORIGINAL PAPER

The application of the cluster identification method for the detection of leakages in water distribution networks

H.-Y. Lin · B.-W. Lin · P.-H. Li · J.-J. Kao

Received: 20 May 2013 / Revised: 18 July 2014 / Accepted: 12 August 2014 / Published online: 28 August 2014

Abstract To reduce the amount of water wastage caused by leakage, the utilities have to monitor and detect leakage of water distribution networks periodically. In order to identify leaking pipelines efficiently when limited resources are available, a cluster identification method (CIM) is proposed to establish a priority for leakage detection and to assess whether spatial clusters of high failure-prone areas exist. The proposed CIM evaluates the difference between the observed data and simulated trials to determine the statistical significance of each cluster; a method previously applied only in epidemiology studies to assess the occurrence probabilities of rare diseases for spatial clusters. The CIM suggested in this study is the overlapping local case proportions (OLCP) that uses grids to scope the entire area and then to simulate the number of failures in the neighborhood of each grid. The simulated failure ratios are then compared with the existing records to determine the statistical significance. The statistical significance represents the potential of the grid requiring further leakage detection. Three failure probability estimation methods, including local average, global average, and empirical equation, are utilized to analyze the suitability of the OLCP for use with various probability inputs. A case study in the central region of Taiwan was implemented to demonstrate the applicability of the proposed method. The results indicate that the rate of failure in the following year found within the spatial clusters determined by the OLCP was twice the average amount and thus provided valuable information used to prioritize the pipelines for further inspection.

Keywords Water leakage detection · Spatial clusters of high failure-prone areas · Overlapping local case proportions · Failure probability · Statistical significance

Introduction

In order to reduce the amount of water loss caused by leakages, the detection of potential leakages in water distribution networks (WDNs) must be periodically implemented and the deteriorated pipelines are then rehabilitated. To determine whether there is a leakage a WDN into district-metered areas (DMAs) (Farley and Trow 2003; Charalambous 2005) is usually delineated first and then acoustic devices are then applied to find the exact locations of the leakages. These tedious routines require considerable manpower and are generally time-consuming, especially in highly interconnected WDNs for which only a few DMAs can be implemented when a limited budget is available. If potentially faulty pipelines can be effectively

H.-Y. Lin (✉) · B.-W. Lin
Department of Environmental Engineering and Management, Chaoyang University of Technology, 168, Jifong E. Rd., Wufong District, Taichung 41349, Taichung County, Taiwan, ROC
e-mail: hylin@cyut.edu.tw

P.-H. Li
Green Energy and Environment Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, ROC

J.-J. Kao
Institute of Environmental Engineering, National Chiao Tung University, Hsinchu, Taiwan, ROC

identified, it would assist in the design of DMAs and would also reduce the need for on-site leakage detection.

Existing empirical regression equations (Walski and Pelliccia 1982; Su et al. 1987; Kleiner et al. 1998; Alvisi and Franchini 2005; Dandy and Engelhardt 2006; Mondéjar-Jiménez et al. 2013) based on the historical data of pipeline ages, diameters, and materials have been widely used to estimate the failure probabilities. These equations, although useful for the determination of the failure rate of pipelines in general, do not consider essential unforeseen or temporary external factors. For instance, external loads caused by trucks from nearby construction sites and soil erosion can also increase the failure rate of surrounding pipelines. These external factors may be temporary, case dependent or locally specific, which are generally neither included in equations nor detected by inspecting logging data. Pipeline leakages caused by persistent and locally external factors may contribute significantly to water wastage Therefore, a procedure that can taken into account external factors by exploiting maintenance data, including records concerning leakage and breakage rehabilitations is proposed in this study.

A labor-intensive leakage detection program for the entire area of a leaking DMA needs to be carried out if significant evidence for leakage-causing factors is insufficient. This situation also poses a dilemma in that the evidence of a specific factor can be assured only after sufficient inspection data have been accumulated. An area with a higher failure rate suggests that the probability of leakage is also higher. Therefore, a compromise solution for leakage detection is to look for abnormal phenomena in the data. Cluster identification methods (CIMs) are widely applied in rare disease epidemiology studies (Best et al. 2001) to identify spatial clusters of significantly high morbidity. Since leakage can be also regarded as a disease of a WDN and an appropriate CIM, it should also applicable for this study.

By using a CIM, the typical prevalence of different population groups (e.g., sex, age) for a disease is obtained first and then used to estimate the number of people likely to contract the disease within the given area. These estimated numbers are then compared with the observed numbers in the area using a statistic hypothesis to evaluate significance. Areas with statistical significance may merit further investigation for lurking risk factors or potential hazards. For instance, Yiannakoulias (2009) applied data relating to population attributable risk in order to analyze the spatial clusters of patients with lung cancer in Ontario, Canada. There are several CIMs designated for spatial epidemiology studies, including overlapping local case proportions (OLCP) (Rushton and Lolonis 1996), geographical analysis machine (Openshaw et al. 1988), and spatial scan statistics (Kulldorff and Nagarwalla 1995). Smith (2001) analyzed the three CIMs and established that the OLCP method is superior in data sensitivity and tends to be capable of finding critical spatial clusters efficiently. In this study, the OLCP method was thus selected for detecting clusters with potential leakages in a WDN. Rushton and Lolonis (1996) have utilized the OLCP in the spatial distribution of birth defects in urban areas to detect clusters with a high risk potential. In a similar manner, the WDN and maintenance records are analyzed by the OLCP herein to identify the spatial clusters with the potential for high failure rates. To assess the applicability of OLCP for WDN leakage detection problems, a research project with a 2-year period (2008–2010) has been conducted in Taichung City, Taiwan with the data of the local utility, the fourth branch of Taiwan Water Corporation.

Materials and methods

Figure 1 depicts the research procedure used in this study, including the four major steps. First, the data of a WDN were collected and then the failure probabilities of the pipelines were estimated. The estimated probabilities were then utilized in an OLCP to locate spatial clusters with a high leakage potential. Finally, the results of the OLCP

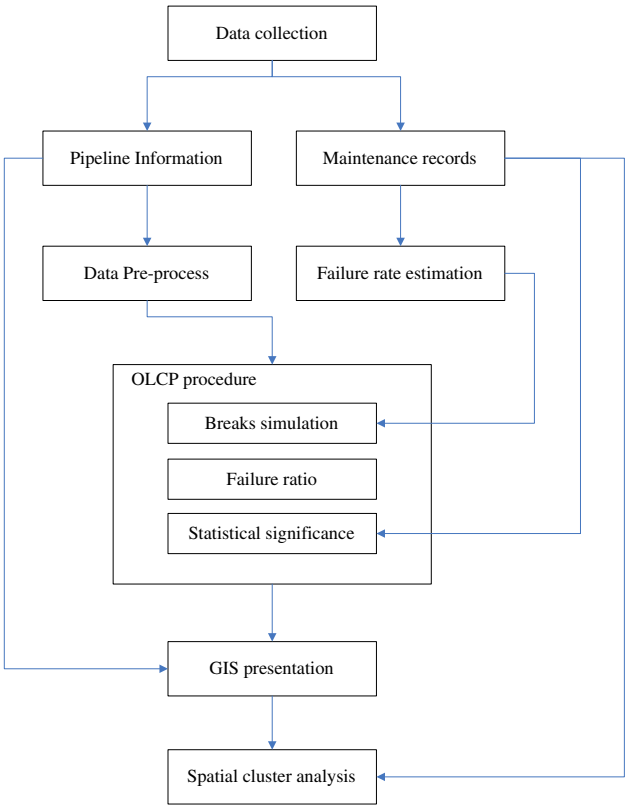

Fig. 1 Flowchart of OLCP analysis in WDN failure cluster detection

were illustrated by using GIS. Each step is described in the following sections.

Data collection

The data required for an OLCP analysis include the location coordinates, ages, materials, and commercial sizes of the installed pipes. A pipeline is split into numerous small sections based on its unit length. The role of a pipeline section is similar to that of a resident in the OLCP analysis. Each section, as for a resident, is regarded as an independent unit. The center coordinate of a pipeline section represents its spatial attribute in an OLCP, similar to the location where a resident dwells for epidemiology studies. Pipeline age, material, and commercial size are used to estimate the risk of leakage. In addition to the pipeline information of a WDN, its maintenance records, including date and coordinate, are also collected for implementing the OLCP analysis.

Failure probability estimation

The failure probability of a WDN pipeline, similar to the prevalence of a disease in spatial epidemiology studies, provides an indicator to assess whether the occurrence of concerned events, such as breakage and leakage, of an area is statistically significant. With the failure probability, an OLCP is applied to simulate failure events in order to compare them with the existing maintenance data. Three methods utilized to estimate failure probabilities in this study, including the empirical equation, local average, and global average, are presented in the following sections.

Empirical equation

There are several empirical equations available for estimating number of failures for each pipeline (e.g., Walski and Pelliccia 1982; Kleiner et al. 1998; Alvisi and Franchini 2005; Dandy and Engelhardt 2006). Walski and Pelliccia (1982) provide a useful regression model based on maintenance and pipeline data:

$$N_j(t) = ace^{b(t-k)} \quad \text{(Walski and Pelliccia 1982)} \qquad (1)$$

where t is the index of the year to be assessed; j is the index of a pipeline, $N_j(t)$ is the failure probability (number of failures/year/meter) of pipeline j in year t; a and b are both regression coefficients; c is a correction factor reflecting the effect of different pipeline diameters on failure rates; k is the installation year of pipeline j.

Based on Eq. (1), Chuang (2008) processed the maintenance data, including the records for both leakage and breakage rehabilitations of Taichung City, Taiwan, and concluded that the failure probability was expressed as:

$$N_j(t) = 1.29 \times 10^{-5} e^{0.1262(t-k)} \quad \text{(Chung 2008)} \qquad (2)$$

Local average and global average

Since the leakage related data of WDNs are not always available, the average rates of local and global breaks per year per length from existing maintenance records can be used to estimate the probability of failure. Local average represents the average failure rate obtained from the maintenance data in a specific area, while the global average represents the average of all the WDNs. It is worth noting that failure probabilities of all pipelines are identical if estimated using averages, either local or global, while different failure probabilities for various pipelines are expected when using an empirical equation.

Implementation of an OLCP for a WDN

The procedure to implement an OLCP in this study is briefly described as follows. First, the entire study area is divided into grids with equal spacing and a diameter is set which is slightly smaller than the interval between the two grids. Second, each grid is made the center of a circle and a circle with the determined diameter for each grid is drawn. Then the morbidity rate, number of breaks (patients), is divided by the number of total pipeline sections (residents) within each circle. Finally, the statistically significant spatial clusters with higher morbidity levels are determined. These spatial clusters are expected to have a higher potential leakage risk than others and are worthy of further investigation.

In applying an OLCP for the identification of WDN leakage problems, the pipeline information is preprocessed in accordance with an OLCP. Each pipeline section is treated as an independent unit. Figure 2 illustrates how the data are processed. The entire area is split into uniform grids with equal size, as indicated by crosses in the figure. A small circle represents a water pipeline section, while hollow and solid ones indicate functional and failure sections, respectively. A large circle indicates a spatial cluster area for which a local morbidity is counted. The center of a large circle is a grid, which is marked by a cross. The diameter of a large circle, as aforementioned, is slightly shorter than the interval between two grids to avoid any overlap of areas covered by two adjacent clusters, which may cause dependent results of local morbidities and consequently result in incorrect spatial clusters being identified.

After the WDN data are processed, pipeline sections are assigned with the aforementioned failure probabilities and

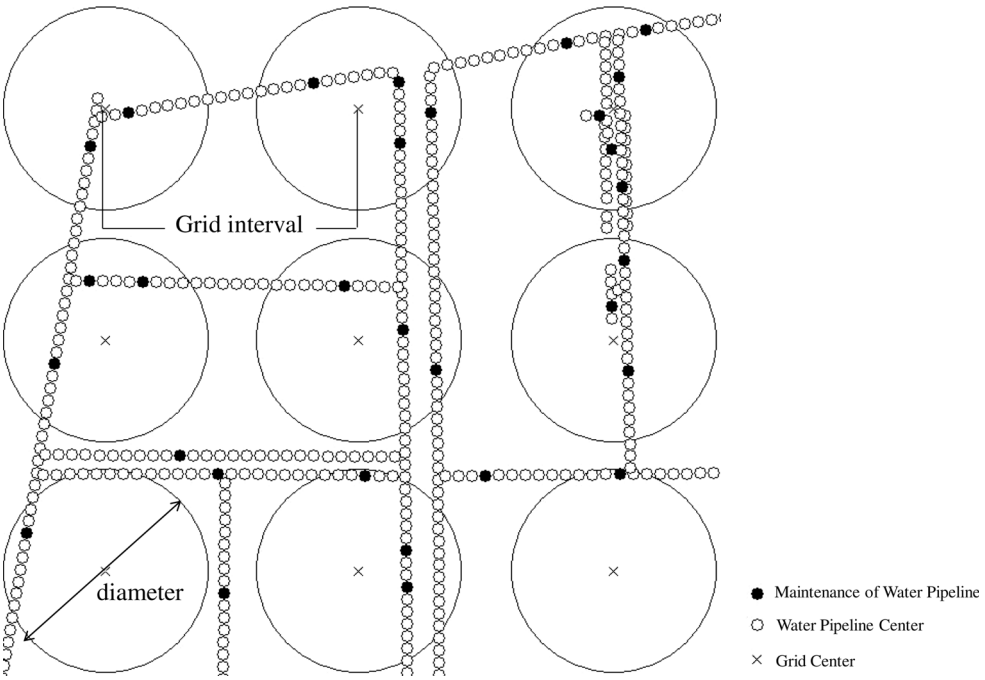

Fig. 2 Grid center, pipeline center, and cluster coverage diameter

used to simulate failure events. In a simulation trial, each pipeline section is randomly marked either functional or failure based on its given failure probability and the Poisson process. The failure rate of a cluster area is then calculated by dividing the number of failures by the number of pipeline sections inside the area. After sufficient simulation trials, the observed failure rate calculated from the existing maintenance records is compared with the simulated rates to evaluate the statistical significance of abnormity. The failure rates, including the simulated and observed ones, of each cluster area are gathered and sorted according to where the rank of the observed rate indicates its irregularity. For instance, if the rank of an observed failure rate of a cluster area is within the ten highest rates among 1,000 values, its probability to occur by random is <1 % and may be worth a further investigation. These statistical significance data, i.e., p value, are then presented by GIS to reveal spatial clusters with high potential for failure.

Odds ratio and scan ratio analysis

After the OLCP analysis, the statistical significance of the observed failure probability for each cluster area of a specific year was determined. To validate whether the failures of the following year for the spatial clusters with high significance are more than for other clusters, the existing maintenance records of the following year were used. An odds ratio was used to verify the performance of the proposed method:

$$ro_f = \frac{k_f/m_f}{K/M} \tag{3}$$

where ro_f represents the odds ratio, i.e., the probability to find the failures in the following year within the areas whose p value is less than f over the probability to find failures in the following year within the entire study area; k_f and K represent the number of failures in the following year within areas whose p value is less than f and within the entire study area, respectively; similarly, m_f and M represent the number of pipeline sections within areas whose p value is less than f and within the entire study area, respectively.

A single odds ratio under p value reflects the abnormality of the observation only and does not include the number of pipeline sections, which may not be sufficient to determine on-site detection loading. Therefore, in addition

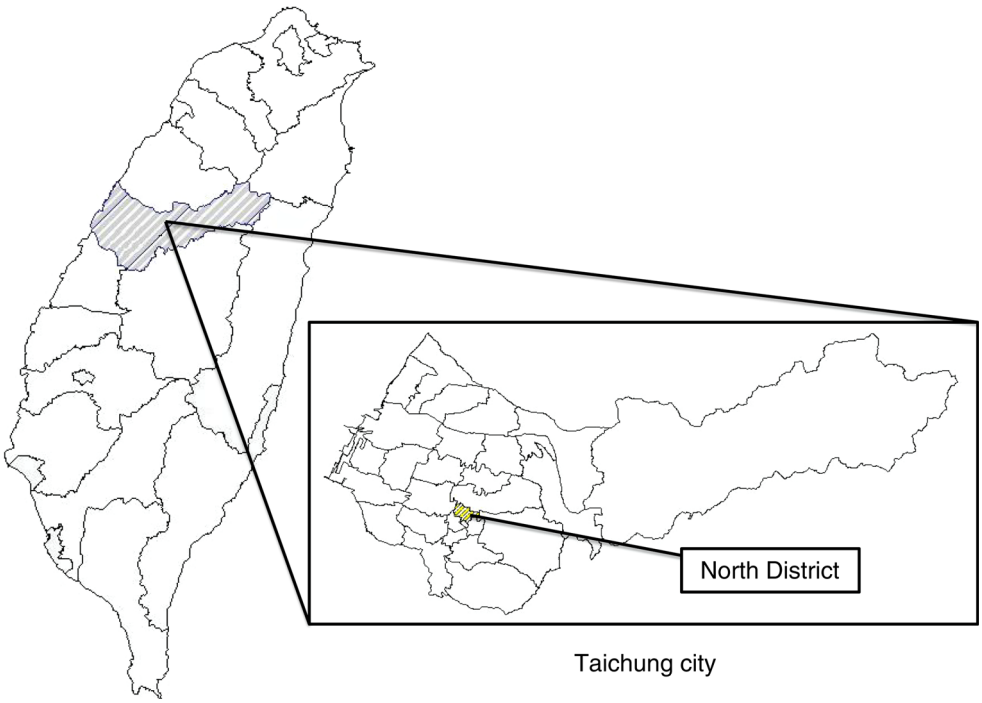

Fig. 3 Location of North District in Taichung City, Taiwan

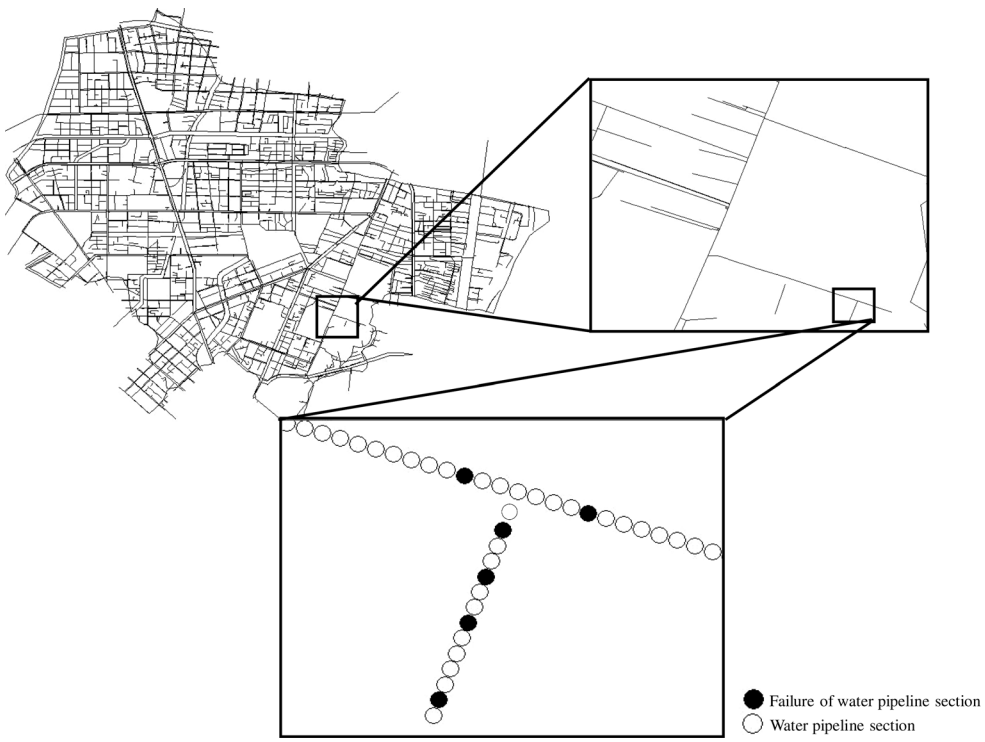

Fig. 4 The WDN and pipeline sections in North District

to the odds ratio being determined by cluster areas with the highest p value, a scan ratio defined by the percentage of potential pipeline sections is an alternative way to evaluate the suitability of OLCP for detecting failures, which was computed as follows:

$$rs_c = \frac{k_c/m_c}{K/M} \quad (4)$$

where c represents the accumulated percentage of pipeline sections with ascending p values to be included; rs_c represents the scan ratio, i.e., the probability to predict failures in the following year within the $c\%$ pipeline sections with ascending p values over the probability to predict failures in the following year within total pipeline sections; k_c represents the number of failures in the following year within the $c\%$ pipeline sections with ascending p value; m_c represents the number of pipeline sections within the $c\%$ pipeline sections with an ascending p value.

Results and discussion

Case background

A case study involving the North District in Taichung City, Taiwan, was carried out to demonstrate the applicability of the proposed method for detecting failure clusters in a WDN. Taichung City is the third largest city of Taiwan, with over 2.67 million residents dwelling in an area covering 2,215 square kilometers. The North District is located in the center of Taichung City, with about 150 thousands residents and in an area of 6.96 square kilometers. Figure 3 presents the location of North District, Taichung City, Taiwan. The software applied in this study to implement the OLCP procedure was Disease Mapping and Analysis Program, version 4 (DMAP IV, Department of Geography, The University of Iowa 2007). The pipeline information and maintenance data of Taichung City were provided by the local utility, the fourth branch of Taiwan Water Corporation. Figure 4 illustrates the WDN of North District and the pipeline sections, which were obtained by dividing the pipeline by its unit length. In this case, the length is 4 m and the total number of pipeline sections is 584,316. Table 1 lists the numbers of rehabilitations performed in 2003–2006. The decreasing number of rehabilitations has affirmed the efforts of the local utility in reducing pipeline failures.

Table 1 Number of pipeline rehabilitations between 2003 and 2006 for Taichung City

Year	Number of rehabilitations
2003	7,414
2004	6,569
2005	5,706
2006	4,755

To estimate the failure probabilities of the pipelines, three methods, including local average (North District), global average (Taichung City), and empirical equation (Chung 2008) are applied in this study. Table 2 lists the failure probabilities estimated by the three methods over 3 years (2003–2005). The failure probabilities for the local average and global average were calculated year by year and thus identical for all pipeline sections in each year, while the probabilities were determined by the empirical formula and varied with pipe ages and thus are different from one another. In this study, the local averages were higher than the global averages, which also indicated that the failure probability of North District was higher than the entire Taichung City in 2003–2005. The probability estimated by Chuang (2008) is much less than those calculated using the other two methods. In addition to the failure probability, the grid interval and diameter of a cluster area are also required in the application of an OLCP. The determination of the grid interval and the diameter is a subtle process. Since a failure rate is calculated for the entire area of each cluster, a larger grid interval and diameter tends to "average" the abnormal events, while a smaller grid interval and diameter often give sensitive but meaningless results due to the insufficient samples expected from a small cluster. In this study, the grid interval and diameter were set to 75 and 50 m, respectively.

In the simulation step of the OLCP, 999 trials were conducted for each failure probability estimation method in this study. In each trial, each pipeline section was marked randomly either success (functional) or fail (failure) based on its given failure probability under the Poisson process. Thus, the ratio of the number of failures over the number of pipeline sections within a cluster area was computed in each trial. The maintenance records, i.e., the failures observed in reality, were compared with the simulated data. For each cluster area and each failure probability estimation method, the statistical significance of the observed

Table 2 Three pipeline failure probabilities utilized in the OLCP analysis

Estimation method	Year	Failure probability
Empirical formula (Chuang 2008)	2003–2005	0.000021–0.011513
Local average (North District)	2003	0.019
	2004	0.017
	2005	0.015
Global average (Taichung City)	2003	0.013
	2004	0.011
	2005	0.010

ratio was determined by its rank among 1,000 ratios, including 999 trials and the observation itself. Figure 5 presents the statistical significance (p value) of the observed maintenance record for 2004 with the local average failure probability. From this it can be seen that the lesser the p value, the higher the ratio of failures. It is obvious that the northwest clusters in the figure have higher ratios of failures than the other clusters, which also implies that on-site detection in this area is expected to be more cost-effective than in other areas.

Results of odds ratios and scan ratios

The existing maintenance record of the following year was then used to compute the odds ratios. Figure 6 shows the 2004–2006 average odds ratios of the various p values for the results obtained using the three different probability estimation methods. In this case study, the odds ratios of all the estimation methods were higher than one, especially in the areas with a low p value, which meant that if on-site detection was conducted in a high p value area, the probability of finding failures would be higher than for other randomly selected areas. For different estimation methods, the North District's average (local) was better than those obtained using the other two methods. For the area with a p value <0.15 and using the local average estimation method, the probability of finding failures in the area was 2.47 (odd ratio) times that of a randomly selected area. It can also be observed that if the p value is too low, the associated odds ratio is not significant because the number of samples is too few.

The pipeline sections were scanned in ascending p values order, which meant that the pipeline section with the smallest p value was scanned first. Comparing with p values, pipeline scanned percentage is a better indicator to assess the WDN detection workloads. Figures 7 and 8 present the scan ratios of the results obtained using the three failure probability estimation methods in a single year and its average, respectively. Figure 7 presents the result of each single year. For instance, the top chart of Fig. 7 utilized the maintenance data of 2003 to determine the high leakage potential (low p value) area and then validated the scan ratios of different pipeline scanned percentage with the maintenance data of 2004. Because the practical

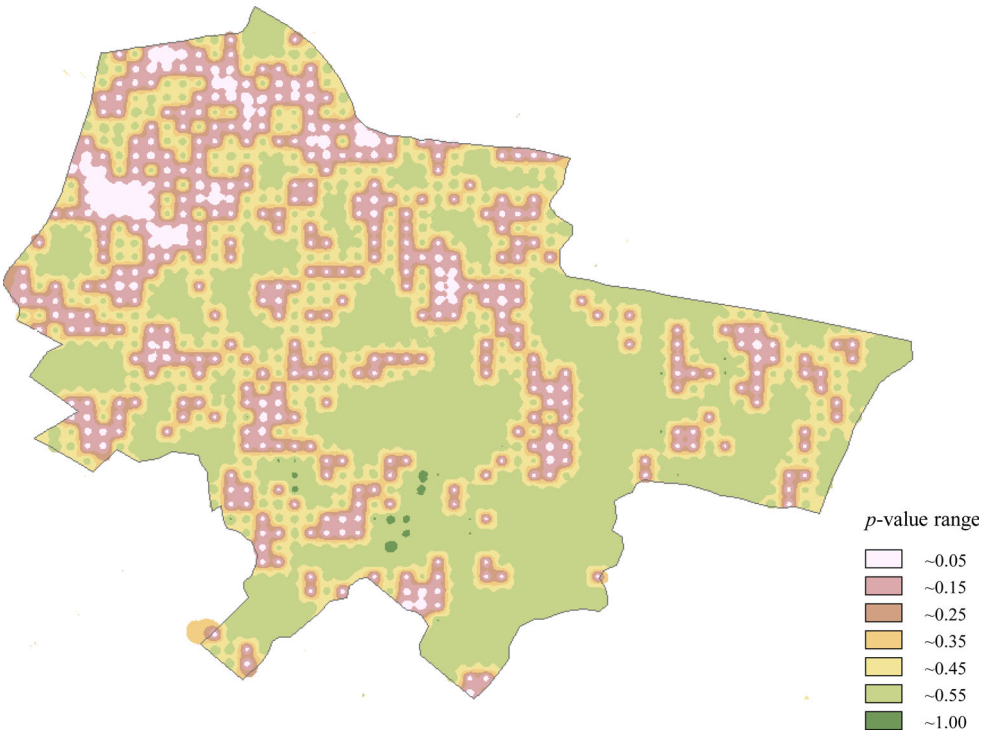

Fig. 5 The p values of all spatial clusters in North District (2004)

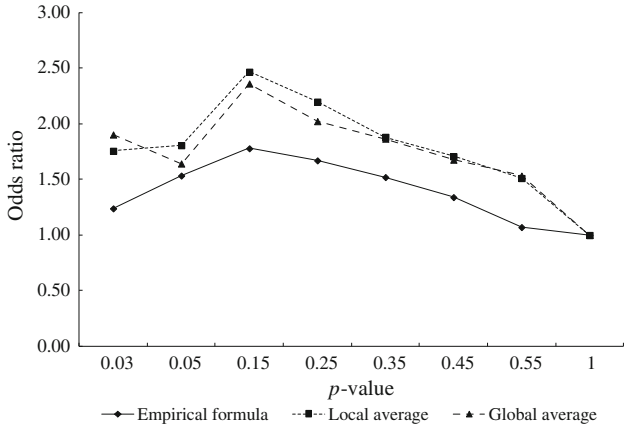

Fig. 6 Odds ratios of various p values for results obtained using three probability estimation methods (2004–2006 average)

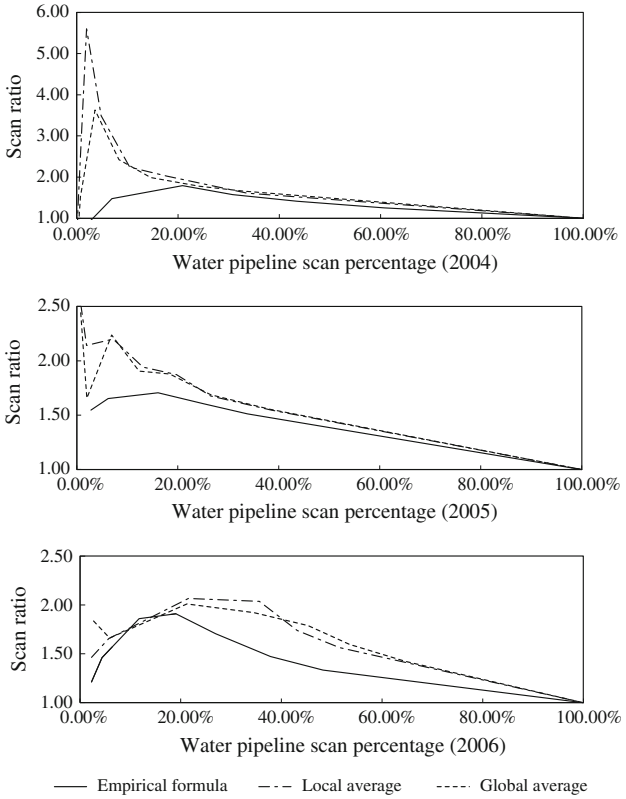

Fig. 7 Scan ratio versus pipeline scanned percentage for each single year

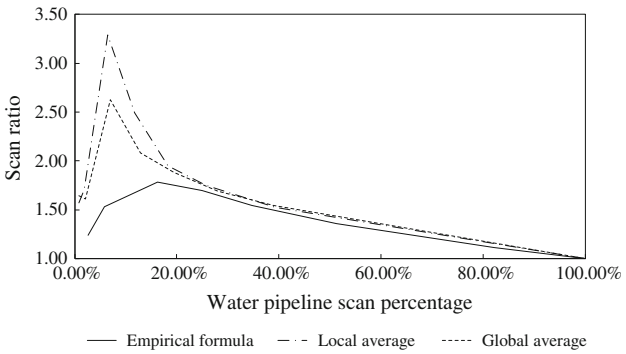

Fig. 8 Scan ratio versus pipeline scanned percentage (2004–2006 average)

maintenance plans were historical data and determined independently from OCLP, Fig. 7 presents the superiority of the proposed methods. The scan ratios for all the methods used for all 3 years are all higher than one, which indicates that if the pipeline sections were scanned in ascending p value order, the probability of finding failures in the following year is always higher than for an identical number pipeline sections being randomly scanned. Figure 8 presents the average performance that this effect achieved its peak at 3.3, 2.63 and 1.79 times for 6.5, 7, and 16.2 % of total pipeline sections scanned with ascending p value order for the local average, global average and empirical equation, respectively. If 10 % of total pipeline sections were to be scanned, the failures in the following year determined by using both the local and global averages in ascending p value order were more than twice of those being randomly scanned in this study. This confirms that the proposed method is a sound prescreen tool to narrow down the on-site detection area required. In both the comparisons based on odds ratios and scan ratios, the local average method outperformed the other methods in most cases.

Discussion and findings

In this study, the failure probability calculated using the empirical equation did not perform as well as the other two methods. The reason for this might be that the empirical equation, which is sensitive to pipe age and diameter, was established from the maintenance records for a larger area, which included Taichung City Branch.

The probabilities estimated from the empirical equation were thus much less than those that actually occurred in the study area. In this situation, the simulated numbers of failure events were significantly less than the actual observed failure events, which led to more spatial clusters being identified with statistical significance (low p value) using the empirical equation than by using the other two methods. Underestimated failure probability can cause inappropriate outcomes and can also overestimate the irregularity of the observed failure events of cluster areas. Theoretically speaking, the more comprehensive and correct is the failure probability, the more credible are the spatial clusters determined by using the OLCP. As a consequence, as the reliable failure probability for each pipeline section is difficult to obtain, this study suggest that the local average failure probability estimated from existing maintenance records is a viable alternative.

Conclusion

In this study, a spatial cluster detection method, OLCP, was applied to improve the WDN pipeline leakage detection analyses. In an OLCP analysis, the pipeline failure probability plays an important role, which is a gauge used to determine whether the observed failure events of cluster areas are statistically significant. Three different methods to estimate the failure probability were discussed in this study and the results indicate that an average rate of failure in either a local or global area is applicable to an OLCP analysis. This finding should simplify the probability estimation required for an OLCP analysis. If comprehensive maintenance or pipeline information is not available or if the derived empirical equation is not suitable for a particular set of records, then the analyst can use average failure probability instead when applying the OLCP. After the OLCP analysis, the statistical significances of failure events of spatial cluster areas provide priority information for further on-site leakage detection. The area required for on-site leakage detection should be reduced or narrowed down by using the proposed method. Additionally, the proposed method has the advantages of cost efficiency, flexibility and opportunity for an immediate response.

Acknowledgments The writers would like to thank the National Science Council of Taiwan, the Republic of China, for financially supporting this research under Contract No. NSC 101-2211-E-324-012. The writers would also like to express special thanks to Dr. Qiang Cai for his developing DMAP IV and assistance in helping the authors use the program in this study.

References

Alvisi S, Franchini M (2005) Rehabilitation scheduling of water distribution systems based on multi-objective genetic algorithms. In: Paper presented at the proceedings of the eight international conference on computing and control for the water industry, University of Exeter, UK

Best NG, Ickstadt K, Wolpert RL, Briggs DJ (2001) Chap 22 combining models of health and exposure data: the SAVIAH study. In: Elliott P, Wakefield J, Best N, Briggs D (eds) Spatial epidemiology: methods and applications. Oxford University Press, NYC

Charalambous B (2005) Experiences in DMA redesign at the Water Board of Lemesos, Cyprus. In: Paper presented at the leakage 2005 conference proceedings

Chuang L-W (2008) District metering area partition procedure and optimization model for water distribution network leakage detection. National Chiao-Tung University, Taiwan

Dandy GC, Engelhardt MO (2006) Multi-objective trade-offs between cost and reliability in the replacement of water mains. J Water Resour Plan Manag 132(2):79–88

Department of Geography, The University of Iowa (2007) DMAP IV. http://www.uiowa.edu/~gishlth/DMAP4

Farley M, Trow S (2003) Losses in water distribution networks: a practitioners' guide to assessment, monitoring and control. International water association

Kleiner Y, Adams BJ, Rogers JS (1998) Long-term planning methodology for water distribution system rehabilitation. Water Resour Res 34(8):2039–2051

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. Stat Med 14:799–810

Mondéjar-Jiménez J, Vargas-Vargas M, Segarra-Oña M, Peiró-Signes A (2013) Categorizing variables affecting the proactive environmental orientation of firms. Int J Environ Res 7(2):495–500

Openshaw S, Charlton M, Craft AW, Birch J (1988) Investigation of leukaemia clusters by use of a geographical analysis machine. The Lancet 331(8580):272–273

Rushton G, Lolonis P (1996) Exploratory spatial analysis of birth defect rates in an urban population. Stat Med 15:717–726

Smith GH (2001) Disease cluster detection methods: the impact of choice of shape on the power of statistical tests. http://www.

cobblestoneconcepts.com/ucgis2summer/smith/SMITH.HTM. 2013

Su YC, Mays LW, Duan N, Lansey KE (1987) Reliability based optimization model for water distribution systems. J Hydraul Eng (ASCE) 114(12):1539–1556

Walski TM, Pelliccia A (1982) Economic analysis of water main breaks. J Am Water Works Assoc 74(3):140–147

Yiannakoulias N (2009) Using population attributable risk to understand geographic disease clusters. Health Place 15(4):1142–1148. doi:10.1016/j.healthplace.2009.07.001