World Research Journal of Civil Engineering
ISSN: 2277-5986 & E-ISSN: 2277-5994, Volume 4, Issue 1, 2015, pp.-46-52.
Available online at http://bioinfopublication.org/journal.php?opt=azjou&jouid=BPJ0000006

HOTSPOT ANALYSIS OF WATER PIPE BREAKS USING KERNEL DENSITY ESTIMATION

POTYRALLA M.* AND ZAWADZKI J.
Faculty of Environmental Engineering, Warsaw University of Technology, 20 Nowowiejska Street, 00-653 Warsaw
*Corresponding Author: Email-maciej.potyralla@gmail.com

Received: August 19, 2015; Revised: October 22, 2015; Accepted: October 22, 2015

Abstract- Water supply system is one of the strategic facilities of city infrastructure. It is one of the most important and most costly elements of the water system (60 to 90% of the cost of the entire system), while the network plays a very important role because of the need for reliability of the water supply to consumers. Inefficiency of water supply network deteriorates the quality of life of residents and leads to loss of water supply security. The article presents the analysis of the reliability of the water supply system in the city of Warsaw between 2010 and 2012. Non-parametric methods find increasing number of applications in the area of data analysis and data exploration. In this paper the most popular tool used among those methods was presented – kernel estimators. It became an effective tool to predict high-risk zones within point patterns of pipe burst events by producing a smooth, continuous surface that defines the level of risk for examined area. In addition, kernel density estimation represents a powerful way to conduct hotspot analysis and easily visualize trends over large areas for decision-makers. Practical estimation aspects were presented and example of applications was used in analysis of the water supply network reliability data.

Keywords- failure frequency per unit, kernel density estimation, reliability, water supply system

Citation: Potyralla M. and Zawadzki J., (2015) Hotspot Analysis of Water Pipe Breaks Using Kernel Density Estimation. World Research Journal of Civil Engineering, ISSN: 2277-5986 & E-ISSN: 2277-5994, Volume 4, Issue 1, pp.- 46-52.

Introduction

The range of the Warsaw water supply system covers approximately 80% of areas within the administrative boundaries of the city. The system satisfies the demand for water of 95.5% of residents. The primary water sources for the city are the following: the Central Water Main, Water Main for Praga City District and the Northern Water Main. All the aforementioned sources feed water to a common water supply network via water distribution pumping stations at water supply stations. The International Standard gives the definition of DN (nominal size) when applied to components of a pipework system, as specified in those standards which use the DN designation system. The term Diameter Nominal refers to the internal diameter of a pipe, in which sizes are measured in millimeters. The water supply network within the system is formed by the following: water mains - DN 300 - 1400 mm pipelines, DN 80 - 250 mm distribution pipelines, water supply connecting pipelines with DN below 80 mm.

The total network length in the area of Warsaw comes to 3076.36 km, including [1]:

- transmission pipelines and water mains - 392.41 km,
- distribution pipelines - 1936.3 km,
- water supply connecting pipelines - 747.65 km.

The water mains are basically connected into a ring configuration. Only some of them are branched water mains and these are mainly water mains feeding water to Włochy, Ursus and Rembertów city districts as well as suburban areas around Warsaw.

Majority of water losses is generated due to leakage caused by poor condition of the water supply system. For water undertakers, losses at water supply networks are most serious. The most frequent root cause for water losses at external water supply networks are as follows:

- leakage from leaky pipelines and plumbing appurtenances,
- water supply pipeline failures,
- hydrant leakages,
- material flaws and defective placement of water supply pipelines,
- incorrect operation,
- Water stealing.

The following determine water supply network leakage volume:

- condition of pipelines and fittings,
- pipeline diameter and age,
- pipeline material,
- density of the mains water connections,
- pipeline density within the network,
- soil conditions,
- water pressure within the network and daily pressure fluctuations [2].

Operating reliability of the water supply system can be expressed as its function to complete the system's tasks related to delivering water in adequate quantity, according to a required quality and at a correct pressure. The water supply systems are to be selected and adapted to customer requirements, so as to provide operating reliability of the systems.

Methods of the analysis

The main reasons for leakage in the water supply system are improper condition and excessive pressure in the network. Both domestic and foreign data confirm that the control of pressure level can significantly contribute to reducing the number of failures and water losses occurring in water supply network. Pressure control has a very important meaning and huge impact on the operation, working network conditions and its service life. Each city water network is subjected to external impacts, such as irregular water consumption, variable pressure (including water hammer) and uncontrolled leaks. Such modern approach to automatic pressure regulation and water losses reduction is mostly solved by pressure control valves (PC Valves).

As a measure of the water supply network condition one can assume its failure frequency expressed as a number of failures per kilometer per annum. The value of this coefficient shall not exceed 0.2 [failure/km x year] [3]. In the period from 1992 to 1996 the average value of this coefficient for Warsaw came to 0.62 [failure/km x year] [4]. Many water supply systems in Europe and in the rest of the world feature the failure frequency coefficient approximated to a value within the range of 0.10 - 0.30 [failure/km x year].

Reliability

The water supply system failure frequency expresses a ratio of the number of failures occurring in a given time to length of this time and the number of analyzed elements. The formula used to evaluate this coefficient based on operation data for water supply pipelines can be expressed as follows [5] [Eq-1]:

$$\mu(t) = \frac{n(t, t+\Delta t)}{n(t)\Delta t} \qquad [1]$$

where:
n(t, t+Δt) – number of facilities renovated by the time (t, t+Δt),
n(t) – number of facilities renovated by the time t,
Δt– time periods into which the analyzed renovation periods are divided.

Hotspot identification

Hotspot analysis is a spatial analysis and mapping technique for selecting a set of high-risk zones. These areas are depicted as points in a map and refer to locations of events or objects. A hotspot can be defined as an area that has higher concentration of events compared to the expected number given a random distribution of events [6-7]. The technique has been shown to have advantages over other methods of aggregating point data, such as spatial clustering, choropleth mapping, and quadrat mapping [8]. It became an effective tool to predict high-risk zones within point patterns of pipe burst events by producing a smooth, continuous surface that defines the level of risk for examined area. In this paper the most popular tool used among those methods was presented – kernel estimators.

Kernel Density Estimator

Kernel Density Estimator is a nonparametric estimator type, intended to determine probability density function of random variable based on obtained sample, i.e. the value which the analyzed variable assumed during measurements performed so far [9]. It has the desirable qualities of directly producing a density estimate, and being influenced by effects of grid size and placement. Furthermore, because it is nonparametric, it has the potential to accurately estimate densities of any shape, provided that the level of smoothing is selected appropriately [10]. Kernel density estimator of probability density function for random variable X is defined as [11] [Eq-2]:

$$\widehat{f}(x) = \frac{1}{mh^n} + \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right) \qquad [2]$$

where n is the number of point observations, h is the bandwidth, K is the kernel, x is a vector of coordinates that represent the location where the function is being estimated, and xi are vectors of coordinates that represent each point observation [12].
Zero-symmetric function with a weak global maximum at this point fulfills following condition [Eq-3]:

$$K : R^n \rightarrow [0, \infty) \qquad [3]$$

considering the following dependency [Eq-4]:

$$\int_{R^n} K(x)dx = 1 \qquad [4]$$

and is termed as kernel while the h positive coefficient is referred to as a smoothing parameter [13].
The most important kernel function types are: Gaussian kernel, triangular kernel, quartic kernel and Epanechnikov's kernel [14].

Analyzed network profile

The material structure of the water supply network in Warsaw has been changing systematically since 1990. Last studies show that despite intensive extension and replacement of old pipelines into pipes made of new materials, the material structure of the analyzed water supply networks still features large quantities of grey cast iron (56.54%), ductile cast iron (21.45%) and steel (12.35%). The material structure includes a small share of plastic pipes which comprise approximately 5.01% (PVC) and 0.22% (PE) of the length of analyzed water supply networks, respectively. The asbestos cement pipelines (4.43%) have been undergoing systematical replacement into pipes made of new materials such as PE and ductile cast iron. In general, one may conclude that the water supply networks are constructed mainly using grey cast iron, steel, PVC and PE pipes - the pipelines made of these materials comprise 74% of length of the analyzed networks [15].
The highest number of failures entails cast iron and comes to 59%. The lowest failure percentage share has been identified for pipes made of unplasticized polyvinyl chloride (PVC) and polyethylene (PE) and comes to 1% of failure rate for either pipe type.
During the analyzed period, the largest number of failures was found at DN 80 - 250 mm distribution pipelines, which represent 66% of all failures. Mains water connections with diameter below DN 80 mm comprise 28% while the fewest failures were noted at DN 300 – 1400 mm mains (only 6% during analyzed period).
The completed failure frequency analysis for the water supply network proved explicit impact of soil and mains water temperature, caused by air temperature changes, on increase of breakage intensity in the winter-to-spring (December to March) period. During spring-to-summer (March to June) period, in turn, one can find the pipeline failure frequency to be lower. Based on figure [Fig–1] it is possible to draw a conclusion that from July to November, the failure rate at the water supply network is kept at a similar level, from 300 to 350 failures per month. Average air temperatures in the city of Warsaw during winter period (December to February) is higher in 2011 than in 2010 and 2012 [Fig-2].

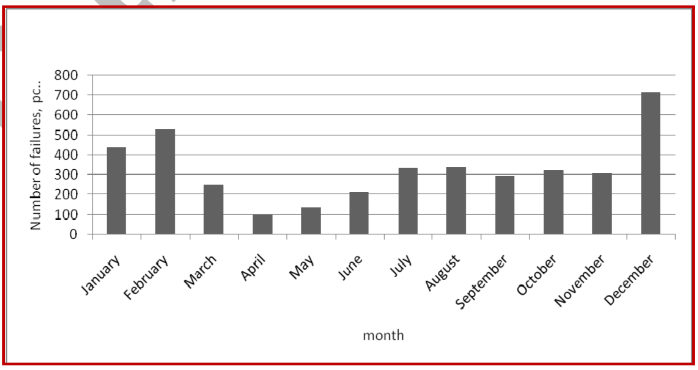


Fig-1 Total failures of the water supply network in each month

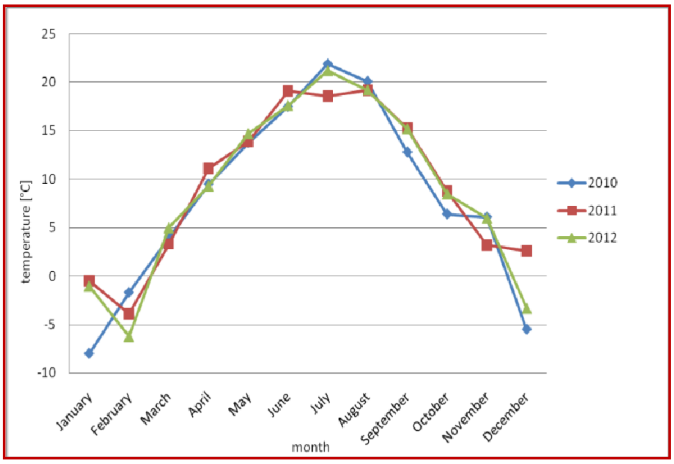


Fig- 2 Average air temperature in the city of Warsaw in 2010, 2011 and 2012

The most frequent method for repairing failures at the water supply network was application of a triple repair coupling [Fig-3]. The number of times this method was used came to 1931 and it was used mainly for damages related to lateral breakage at water supply pipelines. Repairs involving inserting a pipe section were made 654 times and were mainly related to removing a longitudinal breakage and a pipe hub (bell) breakage.

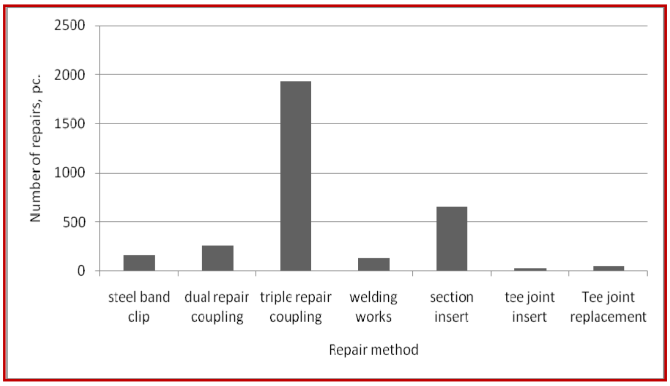


Fig-3 Pipe failure repair method

During the analyzed period, the largest number of failures occurred at distribution pipelines featuring diameter of 100 mm and 150 mm and came to 28% and 25% respectively [Fig-4]. The graph shows that 50 mm and 200 mm diameter pipelines also ranked high in terms of failure rate - 11% for either diameter. Failures at water mains with diameter over 300 mm showed little influence in the analyzed period.

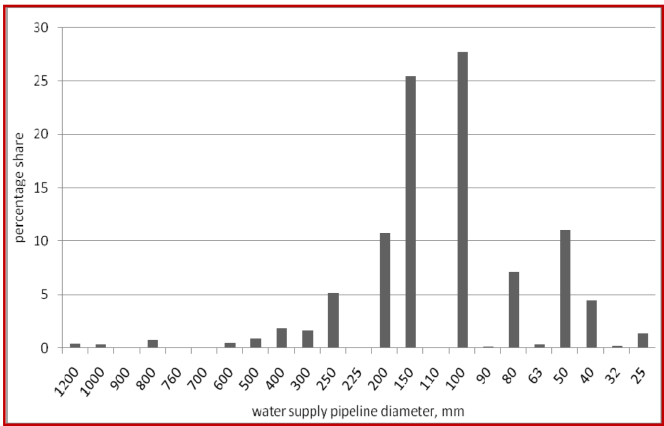


Fig-4 Percentage share of diameters of water supply network

Based on the analysis of the water supply network failure distribution [Fig-5] it follows that during the analyzed period the pipelines were most frequently affected by lateral breakage (1499 failures) and steel point corrosion (1408 failures) Longitudinal breakage has also a considerable share in failure rates. These three types of damage play a crucial role in all incidents at the water supply network.

The completed analysis of the water supply network failure frequency proved that the highest failure rate depending on the pavement type pertains to soft landscaped areas, i.e. street lawns [Fig–6]. During the analyzed period, 1058 failures were noted in these locations. Failures occurring at bituminous roadways ranked second - 760 incidents. When it comes to walkway pavements, in turn, the largest number of failures (741) occurs at paver walkways. On the other hand, the least frequent failures were noted at cobblestone roadways (11 failures), granite paver walkways (16 failures) and at roadway shoulders (10 failures).

The year of 2011 saw the lowest failure frequency per unit at the distribution network pipelines and at mains water connections. This came to 0.3656 [failure/km x year] and 0.3317 [failure/km x year] respectively. The highest water supply network failure frequency was noted in 2012. The failure frequency figure for the distribution network came to 0.5280 [failure/km x year] while for water mains connections this value came to 0.5216 [failure/km x year]. During the analyzed period the water mains showed similar values from 0.1962 [failure/km x year] in 2010 to 0.2192 [failure/km x year] in 2012 [Fig–7].

For the purpose of the density analysis, a non-parametric kernel estimation method was employed. When selecting the kernel form as well as the h smoothing parameter, the Mean Integrated Square Error (MISE) minimizing criterion was used. The following assumptions were made to determine the MISE values:

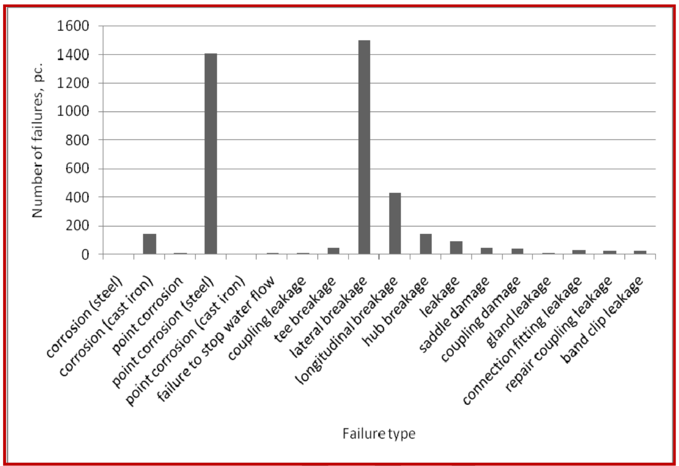


Fig-5 Distribution of the water network failures

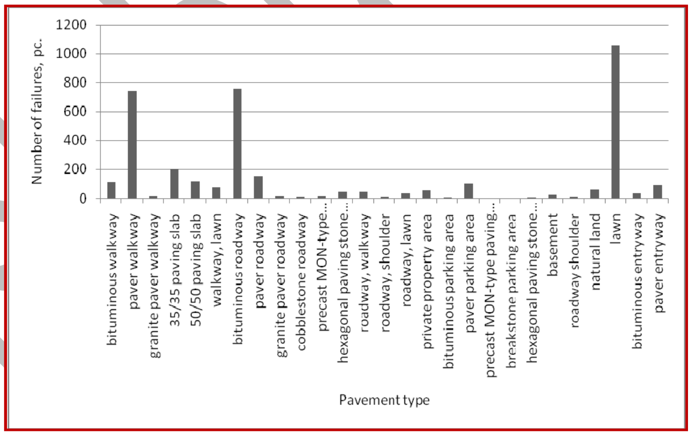


Fig-6 Total failures of the water supply network depending on the type of ground/pavement type

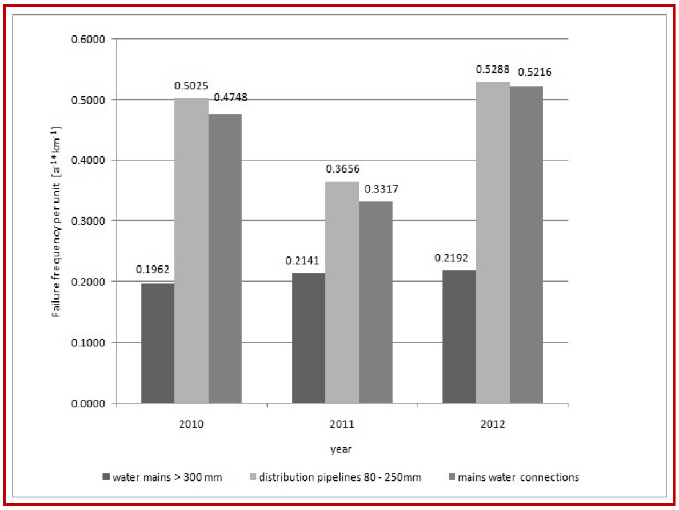


Fig-7 Failure frequency per unit of water supply system in particular years depending on the pipeline type

For the purpose of the water supply network failure density analysis, a non-parametric kernel estimation method was employed. For selection of the kernel form as well as the h smoothing parameter, the Mean Integrated Square Error (MISE) minimizing criterion was used. The following assumptions were made to determine the MISE values:

- the f density features the second derivative that is square integrable,
- the K function fulfills the following condition [Eq-5]:

$$\int_{-\infty}^{\infty} x^2 K(x)dx < \infty \qquad [5]$$

The following dependencies were defined [Eqs-6-8]:

$$U(K) = \int_{IR^n} x^2 K(x)dx \qquad [6]$$

$$W(K) = \int_{IR^n} K(x)^2 dx \qquad [7]$$

$$Z(f) = \int_{IR^n} \left[\nabla^2 f(x)\right]^2 dx \qquad 8]$$

Where [Eq-9]:

$$\nabla^2 f(x) = \sum_{i=1}^{n} \frac{\partial^2 f(x)}{\partial x_i^2} \qquad [9]$$

the bias value necessary to calculate the MISE value is equal to [Eq-10]:

$$E\left(\hat{f}(x)\right) - f(x) = \frac{1}{2} h^2 U(K) \nabla^2 f(x) + o\left(h^2\right) \qquad [10]$$

while the variance value comes to [Eq-11]:

$$V\left(\hat{f}(x)\right) = \frac{1}{mh^n} W(K) f(x) + o\left(\frac{1}{mh^n}\right) \qquad [11]$$

the o(h2) function fulfills the following condition [Eq-12]:

$$\lim_{h \to 0} \frac{o\left(h^2\right)}{h^2} = 0 \qquad [12]$$

whereas $o\left(\frac{1}{mh^n}\right)$ assumes the form of [Eq-13]:

$$\lim_{h \to 0} \frac{o\left(\frac{1}{mh^n}\right)}{\frac{1}{mh^n}} = 0 \qquad [13]$$

Upon completing the analysis of the aforesaid equations it was concluded that in order to obtain the estimator bias tendency towards zero with an increase of the size of the random sample, the h smoothing parameter should be selected by using the following condition [Eq-14]:

$$\lim_{m \to \infty} h = 0 \qquad [14]$$

In case the estimator variance shows a tendency towards zero, the smoothing parameter is to be selected considering the following dependency [Eq-15]:

$$\lim_{m \to \infty} \frac{1}{mh^n} = 0 \qquad [15]$$

The kernel estimator is asymptotically unbiased at any given x point while variance decreases towards zero with a simultaneous increase of the random sample size.

The Mean Integrated Square Error (MISE) assumes the form of [Eq-16]:

$$MISE = \frac{1}{4} h^4 U(K)^2 Z(f) + \frac{1}{mh^n} W(K) + o\left(h^4\right) + o\left(\frac{1}{mh^n}\right) \qquad [16]$$

Upon considering the above-mentioned conditions, the MISE value comes to [Eq-17]:

$$MISE = \frac{1}{4} h^4 U(K)^2 Z(f) + \frac{1}{mh} W(K) \qquad [17]$$

Reduction of the h smoothing parameter so as to reduce the first summand (bias) causes an increase of the second summand (variance). This dependence is the other way round when the h parameter is increased. Selection of the right values enables minimization of the Mean Integrated Square Error (MISE) value. The minimum value occurs for [Eq-18]:

$$h_0 = \sqrt[n+4]{\frac{nW(K)}{U(K)^2 Z(f) m}} \qquad [18]$$

and comes to [Eq-19]:

$$MISE = d(n) \sqrt[n+4]{\frac{U(K)^{2n} W(K)^4 Z(f)^n}{m^4}} \qquad [19]$$

The above formula enables selection of the right kernel. The mean square functional is proportional to the expression of $\sqrt[n+4]{U(K)^{2n} W(K)^4}$, which is the lowest in case of Epanechnikov kernel.

A raster was formed within which density of occurrence of the analyzed points in a cell was assumed as the attribute of each cell. On that basis, square areas with a specific size were formed. The density values were summed up for all the incidents, by giving estimated number of points for a given cell. The process was repeated for each raster cell. The areas combined in this manner provided a generalized result of the density analysis transformed into a raster image indicating color-coded areas of larger input data clusters [Fig–8].

The obtained results were presented on a vector layer covering the area of the capital city of Warsaw. The map presents density of given clusters - the more intense the color, the larger cluster density. The hotspot method enables identification of clusters featuring high and low values of the analyzed parameter. Based on the analyzed data it follows that in 2010 the largest number of failures was noted at the water supply network in the area of east part of Wola and west Mokotów districts.

In 2011 the largest number of failures was noted at the water supply network in the area of and Śródmieście Północne (north downtown area) and west Mokotów districts.

In 2012 the largest number of failures was noted at the water supply network in the area of west Mokotów district.

The obtained results, including geographical co-ordinates, were presented on a 3D graph [Fig-9]. It presents the analysis of the historical reliability of the water supply system in the city of Warsaw between 2010 and 2012. The calculated density distribution features one, very distinct mode in the middle part of the analyzed area. This mode was marked in red. The largest figures were noted for the value range from 20.96 to 21.50 degree. Blue color was used to identify the areas with low density of the water supply network failure occurrence.

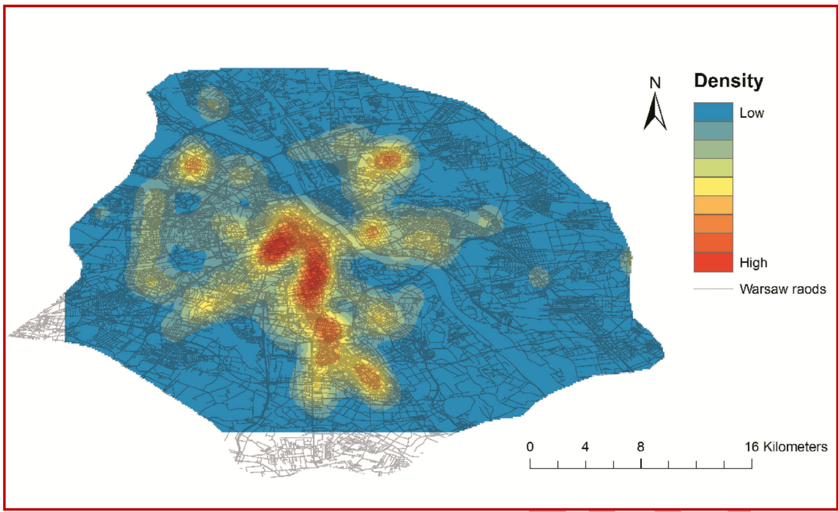
Results and discussion

The obtained data was analyzed and then used to calculate reliability and kernel density estimator for the water supply network. This made it possible to formulate the following results.

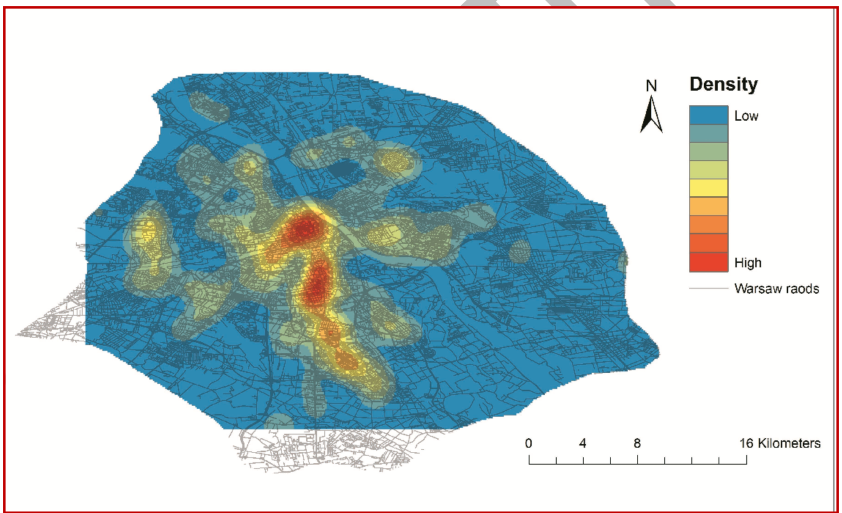
The analyses regarding water supply network failure frequency demonstrated an influence of air and soil temperatures on increase of the failure rate in winter-to-spring (December to March) season. Cast iron pipes cracks due to pressure movement in the ground at a low air temperatures. Air temperature at or below freezing causes the ground above a pipe to freeze increasing external stress on a pipe. In spring-to-summer (March to June) the failure rate is lower. Failure frequency per unit of water supply system in 2011 for water mains, distribution pipelines and connecting pipelines is much lower than in other years. It is connected with higher average air temperatures during winter season.

The most relevant coefficient for evaluation of the water supply pipeline condition is the failure frequency per unit. The foregoing results from a lack of required data enabling performance of calculations.

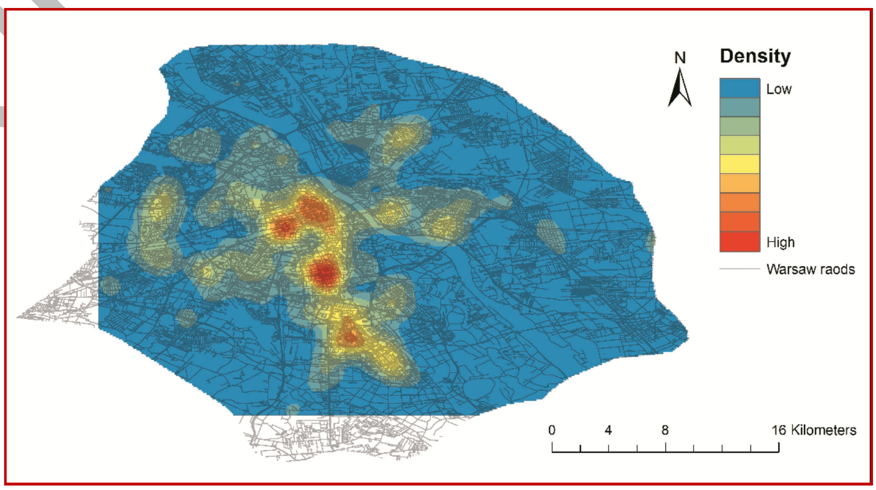
The coefficient expressing water supply network failure frequency per unit for the period



a)



b)



c)

Fig- 8 Density map for occurrence of water supply pipeline failures in a) 2010, b) 2011 and c) 2012

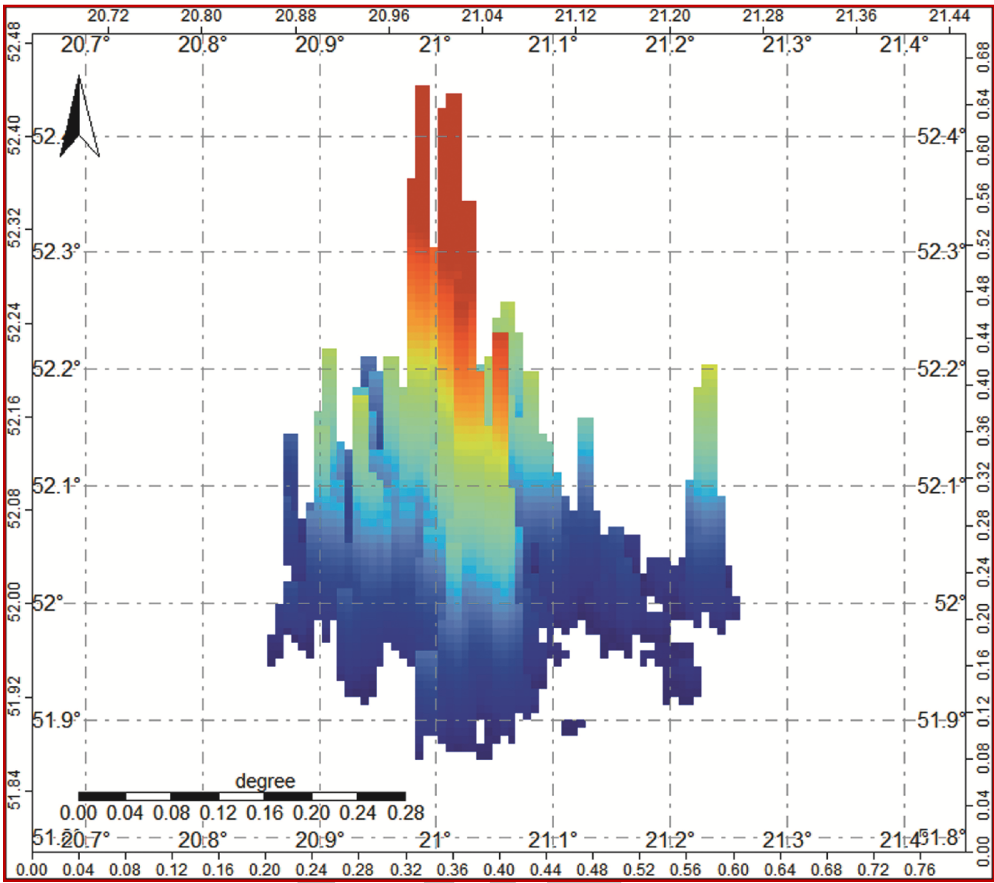


Fig- 9 The estimation density result depending on the failure location

2010 - 2012 came on average to 0.37 [failure/km x year] and represented a considerable improvement against the average value of 0.62 [failure/km x year] noted in Warsaw in the period 1992 - 1996. The total network length in the area of Warsaw increased from 2561,3 km in 1996 to 3076.36 km in 2012.

The coefficient expressing water supply network failure frequency per unit for the period 2010 - 2012 came on average to 0.37 [failure/km x year] and exceeded the permissible values within the range 0.10 - 0.30 [failure/km x year] for many water supply systems in Europe.

Plastic pipes made of PVC and PE feature the lowest failure frequency when compared to pipes made of steel or cast iron.

In terms of pipe diameters, the failure frequency analysis demonstrated the highest failure frequency in 100 mm diameter pipes in the analyzed period. The most positive results for the period 2010 - 2012 were found in 90 mm and 110 mm diameter pipes, featuring very low coefficient of average failure frequency per unit.

Estimation results indicate high-risk zones. The results from prediction map for occurrence of water supply pipeline failures shows that the high failure range in 2012 was seriously reduced compared to previous years. This method allows to define areas for further network monitoring for example high pressure level inspections or pipe replacement.

Conclusions

The goal of the present study was to spatialize pipe breaks occurrence data as an input for pipe breaks risk modeling by using a kernel approach to interpolate historic pipe breaks observations. The analysis has been applied in city districts area in Warsaw to prove that such a technique can be applied in the field of water supply system. The kernel density estimation method is particularly useful as it enables precise identification of the location, spatial range and intensity of areas showing diverse variable values (hotspot). This method can be used in water pipe breaks research to target areas of high concentrations of events for control and to target areas at higher risk for prevention. Using the kernel density interpolation analysis in this study was useful for visualizing the community-wide effects of water supply suspension. The information of spatial sampling density and hotspot pattern could be useful for long-term monitoring and assessment covered by the study.

Acknowledgments

This work was financially supported by the Municipal Waterworks and Sewer Enterprise in the Capital City of Warsaw Joint Stock Company.

References

[1] MPWiK, (2012), Annual Report, Municipal Waterworks and Sewer Enterprise in the Capital City of Warsaw Joint Stock Company, Warsaw, 11-12.

[2] Siwoń Z., Cieżak J. and Cieżak W. (2004) *Environmental Pollution Control*, 4, 25-30.

[3] Denczew S. (2002) Reliability of the safety and reliability of water supply systems on the example of the City of Warsaw, Arkady, Warsaw, 1-18.

[4] Kwietniewski M. (2011) Unreliability of water supply and wastewater infrastructure in Poland based on field tests, XXV Conference on structural failures, 24-27 May 2011, Międzyzdroje, Poland, vol. 25, 127-140.

[5] Kwietniewski M. and Roman M. (1993) Reliability of water supply and sewerage system, Arkady, Warsaw, 21-27.

[6] Lin Y.-P., Chu H.-J., Wu C.-F., Chang T.-K. and Chen C.-Y. (2011) *International Journal of Environmental Research and Public Health*, 8(1), 75-88.

[7] Thakali L., Kwon J. T. and Fu L. (2015) *Journal of Modern Transportation*, 23(2), 93-106.

[8] Leitner M., Arden B. W. and Heukelbach J. (2009) Using kernel density interpolation to visualize the effects of mass treatment with ivermectin on helminth prevalence in rural northeast Brazil, the 24th International

Cartographic Conference, 15–21 November 2009, Santiago de Chile, Chile, 1-10.
[9] Łukasik S. (2008) *Technical Transactions*, 1(E), 3-13.
[10] Seaman D.E. and Powell R.A. (1996) *Ecology*, 77, 2075–2085.
[11] Manepalli U., Bham G. and Kandada S. (2011) Evaluation of hotspots identification using kernel density estimation (k) and getis-ord (gi*) on i-630, 3rd International Conference on Road Safety and Simulation, 14-16 September 2011, Indianapolis, USA, vol. 3, 1-17.
[12] de La Riva J., Pérez-Cabello F., Lana-Renault N. and Koutsias N. (2004) *Remote Sensing of Environment,* 92, 363–369.
[13] Silverman B. (1986) Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, Chapman and Hall, 1-22.
[14] Danese M., Lazzari, M. and Murgante, B. (2008) Kernel density estimation methods for a geostatistical approach in seismic risk analysis: the case study of Potenza Hilltop Town (Southern Italy). Gervasi, O., Murgante, B., Lagana, A., Taniar, D., Mun, Y., Gavrilova, M. (Eds.), Computational Science and Its Applications - ICCSA 2008, Part I, LNCS 5072. Springer, Verlag Berlin Heidelberg, 415–429.
[15] Kwietniewski M., Miszta-Kruk K. and Piotrowska A. (2011) *Environmental Engineering*, 1(Ś), 113-127.