Study on relationships between climate-related covariates and pipe bursts using evolutionary-based modelling

Daniele Laucelli, Balvant Rajani, Yehuda Kleiner and Orazio Giustolisi

ABSTRACT

Researchers extensively studied external loads since they are widely recognized as significant contributors to water pipe failures. Physical phenomena that affect pipe bursts, such as pipe-environment interactions, are very complex and only partially understood. This paper analyses the possible link between pipe bursts and climate-related factors. Many water utilities observed consistent occurrence of peaks in pipe bursts in some periods of the year, during winter or summer. The paper investigates the relationships between climate data (i.e., temperature and precipitation-related covariates) and pipe bursts recorded during a 24-year period in Scarborough (Ontario, Canada). The Evolutionary Polynomial Regression modelling paradigm is used here. This approach is broader than statistical modelling, implementing a multi-modelling approach, where a multi-objective genetic algorithm is used to get optimal models in terms of parsimony of mathematical expressions vs. fitting to data. The analyses yielded interesting results, in particular for cold seasons, where the discerned models show good accuracy and the most influential explanatory variables are clearly identified. The models discerned for warm seasons show lower accuracy, possibly implying that the overall phenomena that underlay the generation of pipe bursts during warm seasons cannot be thoroughly explained by the available climate-related covariates.

Key words | data-mining, Evolutionary Polynomial Regression, impact of climate on water main bursts, knowledge discovery, pipe burst modelling

Daniele Laucelli (corresponding author)
Orazio Giustolisi
Department of Civil Engineering and Architecture,
Technical University of Bari,
Via E. Orabona 4,
70125 – Bari,
Italy
E-mail: *d.laucelli@poliba.it*

Balvant Rajani
Rajani Consultants Inc.,
2024 Glenfurn Ave, Ottawa,
ON, K1J 6G8,
Canada

Yehuda Kleiner
National Research Council Canada,
1200 Montreal Rd, Ottawa,
ON, K1A 0R6,
Canada

NOMENCLATURE

ADD	daily variation of temperature (°K) within time step TS_i;
AGE_t	average age (years) of the pipes in year t ($t = 1962, \ldots, 1983$);
ASC	average snow cover (cm) within time step TS_i;
AT	average of mean daily temperatures (°K) within time step TS_i;
AmT	average of minimum daily temperatures (°K) within time step TS_i;
AMT	average of maximum daily temperatures (°K) within time step TS_i;
BR	pipes burst rate (m/km);
C	number of data subsets for MSC-EPR application;
CoD	coefficient of determination;
CTL_t	total length of pipes (km) at the start of year t;
FZI	freezing index (degree-days);
$g_{i,t}$	age of individual pipe i (belonging to the considered pipe cohort) at year t;
L_i	length of individual pipe i;
MTI	maximum temperature increase (°K) of mean daily temperatures within TS_i;
MTD	maximum temperature decrease (°K) of mean daily temperatures within TS_i;
m	number of days in time step TS_i;
NBE_{t-1}	number of bursts in the previous year;
N_i	number of time steps in the data subset i;

doi: 10.2166/hydro.2013.082

n_t total number of individual pipes at year t;
PBR_{t-1} burst rate of the previous year (events/km);
SG^j (TS_i) snow on the ground (cm) for day j in TS_i;
SSE_i sum of squared errors in the data subset i;
TDG decrease in temperature gradient within time step TS_i;
TIG increase in temperature gradient within time step TS_i;
$TR^j(TS_i)$ total rain (mm) measured for day j in TS_i;
TRN total rain (mm) within time step TS_i;
TS_i time step in the data subset i (days);
T^j_{mean} (TS_i) mean daily temperature (°K) of j in TS_i;
T^j_{min} (TS_i) minimum daily temperature (°K) of day j in TS_i;
T^j_{max} (TS_i) maximum daily temperature (°K) of day j in TS_i;
YLP_{t-1} length of pipes (km) installed in the previous year;
$y_{i,j}$ observed burst rates in time step j of subset i;
$\hat{y}_{i,j}$ predicted burst rates in time step j of subset i;
$\overline{y_i}$ average observed burst rates in the data subset i;
θ air temperature threshold (°K) for evaluation of FZI.

Operators and acronyms

AC Asbestos cement;
ANN Artificial neural networks;
CI Cast iron;
DI Ductile iron;
EPR Evolutionary Polynomial Regression;
MCS-EPR Multi-case study EPR;
MOGA-EPR Multi-objective genetic algorithm EPR;
PVC Polyvinylchloride;
WDN Water distribution network.

INTRODUCTION

The buried pipelines within the water distribution network (WDN) are affected by various mechanisms of deterioration related to material type, the surrounding soil type (e.g., backfill), and external (e.g., earth, traffic and accidental loads) and internal stresses (e.g., variations of pressure). This paper addresses a particular issue that affects WDNs, i.e., the possible impact of climatic and climate-related factors on pipe (water main) bursts. Earlier works on this issue include Newport (1981), Bahmanyar & Edil (1983), Lochbaum (1993), Habibian (1994), Sacluti *et al.* (1999), Skipworth *et al.* (2000), Kleiner & Rajani (2004), Ahn *et al.* (2005), Kleiner & Rajani (2009) and Rajani *et al.* (2012).

Newport (1981) and Walski & Pelliccia (1982) suggested that pipe burst rates (BR) might be correlated to the maximum frost penetration in a given year. Since frost penetration is generally not measured (and can be highly variable) it is practical to instead use some surrogate measures, e.g., the air temperature of the coldest month (Walski & Pelliccia 1982), or freezing index (FZI, expressed in degree-days) (Newport 1981; Lochbaum 1993; Kleiner & Rajani 2002). While observed increase in BR during winter can be related to the increase in earth loads due to frost action, Newport (1981) observed that increased BR in the United Kingdom coincided with very dry summers. One possible explanation for this increase in BR was likely due to the increase in shear (frictional) stress exerted on the pipe by (highly expansive) soil shrinkage in a dry period.

Soil moisture, which affects both frost penetration and soil shrinkage, is therefore an important factor influencing BR of buried pipes. Baracos *et al.* (1955) reported that water main bursts in Winnipeg (Canada) occurred largely between September and January and peaked especially under dry soil conditions after a hot summer or just prior to spring thaw. These findings were confirmed by Morris (1967), Clark (1971) and Hudak *et al.* (1998), who related them to the volumetric swelling and shrinkage of clays. The influence of soil moisture on frost penetration (in terms of depth of the frost front) can be estimated using the Berggren function (Aldrich 1956), which considers thermal properties of the unfrozen and frozen soils, FZI, and latent heat of the frozen soil. The Berggren function shows that dry soil (expected after an extreme dry season) will lead to deeper frost penetration, even if the thermal conductivity and FZI remain unchanged (Rajani & Zhan 1996). Kleiner & Rajani (2004, 2009) implemented both the FZI and the soil moisture (in terms of cumulative annual rain deficit and snapshot rain deficit) in models to predict annual pipe bursts. In particular, the cumulative annual rain deficit was considered as a surrogate for average annual soil moisture, while the snapshot rain deficit was

considered as representative of winter soil moisture, appropriate for cold regions as the frost penetration depends on soil moisture content at the start of winter.

Some studies investigated simply the relationship between air temperature and mains bursts. Bahmanyar & Edil (1983) observed that the number of bursts increased markedly when the temperature dropped suddenly and this drop was sustained for several days (cold snap). Others tried to consider the influence of water temperature (Habibian 1994) as well as a combination of water and soil temperatures (Ahn *et al*. 2005) or water and air temperatures, rainfall and operating pressure (Sacluti *et al*. 1999), or trained artificial neural networks (ANNs) to predict variations in burst rate from recent rainfall and temperature data (Skipworth *et al*. 2000). Ahn *et al*. (2005) confirmed the observation of Bahmanyar & Edil (1983) that the number of pipe bursts increased when the (water and soil) temperatures changed in spring and autumn.

Goodchild *et al*. (2009) explored a wide number of climatic covariates (daily minimum grass temperature, minimum temperature, rainfall, solar radiation, wind run, evapotranspiration, actual evaporation, water content in the topsoil, and soil moisture deficit) trying to explain the timing of observed pipe bursts in the UK. They identified actual evaporation, daily rainfall, minimum grass temperature, and soil moisture deficit to be significant covariates to predict bursts observed in cast iron (CI) and asbestos cement (AC) pipes buried in loam and clay.

Rajani *et al*. (2012) used data from water utilities in the USA and Canada to investigate the correlation among pipe bursts and some temperature covariates. These covariates included mean temperatures, temperature changes, intensities of mean temperature changes, duration and severity of extreme temperatures (by means of the *FZI*), minimally interrupted duration and severity of extreme temperatures, and normalized duration and severity of extreme temperatures. All these covariates were examined in conjunction with a non-homogeneous Poisson-based pipe deterioration model. They examined the impact of temperature changes on observed pipe *BR* for three pipe materials, namely, CI, ductile iron (DI) and galvanized steel, while trying to identify the best time step size for data aggregation in the model they used (the appropriate time step for analysis was identified as 30 days). From their analyses, Rajani *et al*. (2012) found that, especially for CI pipes, air temperature-based covariates could explain much of the occurrence of pipe bursts even if the same covariates were found not to be significant for all pipe groups. In particular, when analyses were conducted on data where only air temperatures were available (which is usual for most water utilities), three covariates, namely, average mean air temperature, maximum air temperature increase and decrease and air temperature increase and decrease intensities over a specific period of days were found to be the most consistently significant covariates. Additionally, they found signs for some of the coefficients of covariates to be contrary to intuitive expectations, e.g., decrease in the intensity of air temperature change was associated with an increase in predicted bursts. They therefore argued that it was quite plausible that some of these covariates interact together in a manner that is different from when each covariate is considered on its own.

Finally, a note about terminology is warranted here. In recording pipe burst, most water utilities essentially record a 'repair event', which sometimes is described as a break or a leak. In North America the term 'burst' is often understood to mean a pressure-related event, where a pipe burst due to stress is manifested by high internal water pressure (most likely manifested in a longitudinal split). In the UK and parts of Europe the term 'burst' is often used as a generic term for a repair event, as recorded by the water utility. In this paper the authors have adopted the European term 'burst'.

RELATIONSHIP TO EARLIER RESEARCH

The work conducted by Rajani *et al*. (2012) was used as a starting point for the research reported in this paper. Bursts and air temperature data for Scarborough (Ontario, Canada) are analysed. The analyses focused on failures in 150 mm diameter CI pipes (the largest cohort of pipes in the Scarborough WDN) recorded in the period 1962–1985. This research differs from and adds to the work conducted by Rajani *et al*. (2012) as follows:

- Analyses considered here separate the burst events recorded in the cold and warm seasons, as dependent

on two partially different sets of temperature covariates; this ensures that the climate impact of external loads on buried pipes are considered to act differently (if any) during summer and winter seasons.

- While some of the covariates used here overlap with those used by Rajani *et al*. (2012), new covariates (both temperature and precipitation-related) are examined here.
- Evolutionary Polynomial Regression (EPR) paradigm is used (Giustolisi & Savic 2006, 2009) in this research while Rajani *et al*. (2012) used a multi-covariate non-homogeneous Poisson modelling approach. The EPR searches for a broad class of polynomial expressions for the best fit, where the 'best' is identified based on both accuracy and parsimony (a detailed description is provided later in the paper). While Rajani *et al*. (2012) analysed separately vertically pit cast and spun-cast CI mains, in this research all CI mains were aggregated together. Partitioning data into homogeneous cohorts usually involves a trade-off between data set specificity and size. Pit and spun CI pipes were lumped together to obtain data sets of reasonable size because the analyses described in this paper involved cohort partitioning into two seasons as well as into single-year burst records (described later as the multi-case analysis of EPR or MCS-EPR).

The rest of this article is organized as follows. Available data are described in detail in the next section, followed by an outline of the multi-case strategy of EPR (MCS-EPR) (Berardi & Kapelan 2007; Giustolisi & Berardi 2007) modelling paradigm used here. Next, the application of the model to the available data is described and the results obtained are presented and discussed.

DESCRIPTION OF THE WDN OF SCARBOROUGH (ONTARIO, CANADA)

The data set obtained for the water pipe inventory of Scarborough comprises 6,879 pipes, with a diameter range of 32–1500 mm, laid between 1905 and 2000. Pipe burst records encompassed the period 1962 to 2003. Table 1 summarizes diameters, materials and length of the pipes in the data set. Clearly, 150 mm CI mains comprise the largest cohort. The predominant surficial soil type in Scarborough is clayey-silt to silty tills (glacial deposits represented in Figure 1 by the grey raster numbered with 3), which are typically low or non-expansive soils but susceptible to frost heave. While older pipes were buried in native soil, imported backfill was probably used for more recently installed pipes (mid to late 1960s). Figure 2 illustrates the histories of the total length of the pipe inventory and the associated number of recorded bursts.

It is noteworthy that a hotspot cathodic protection programme was initiated in 1984, whereby (anodes were opportunistically installed whenever a pipe was exposed for repair. Also, a retrofit cathodic protection programme

Table 1 | Length (km) of pipes by diameters and materials of Scarborough network (up to 2000)

Nominal diameter (mm)		CI	DI	AC	PVC	Other	Total length (km)	%
<100		0.21	0.00	0.00	0.00	0.99	1.20	0.10
100		1.44	0.06	0.00	0.00	0.00	1.50	0.13
150		**450.32**	146.23	35.28	23.14	0.00	654.97	56.69
200		79.40	76.62	20.19	11.68	0.68	188.57	16.32
225		0.01	0.00	0.00	0.00	0.00	0.01	0.00
250		8.19	1.64	0.00	0.00	0.00	9.83	0.85
300		111.52	80.14	32.17	18.12	0.08	242.02	20.95
>300		27.99	22.69	6.65	0.00	0.02	57.34	4.96
Total length	(km)	679.07	327.38	94.30	52.95	1.76	1155.45	
	%	58.77	28.33	8.16	4.58	0.15		

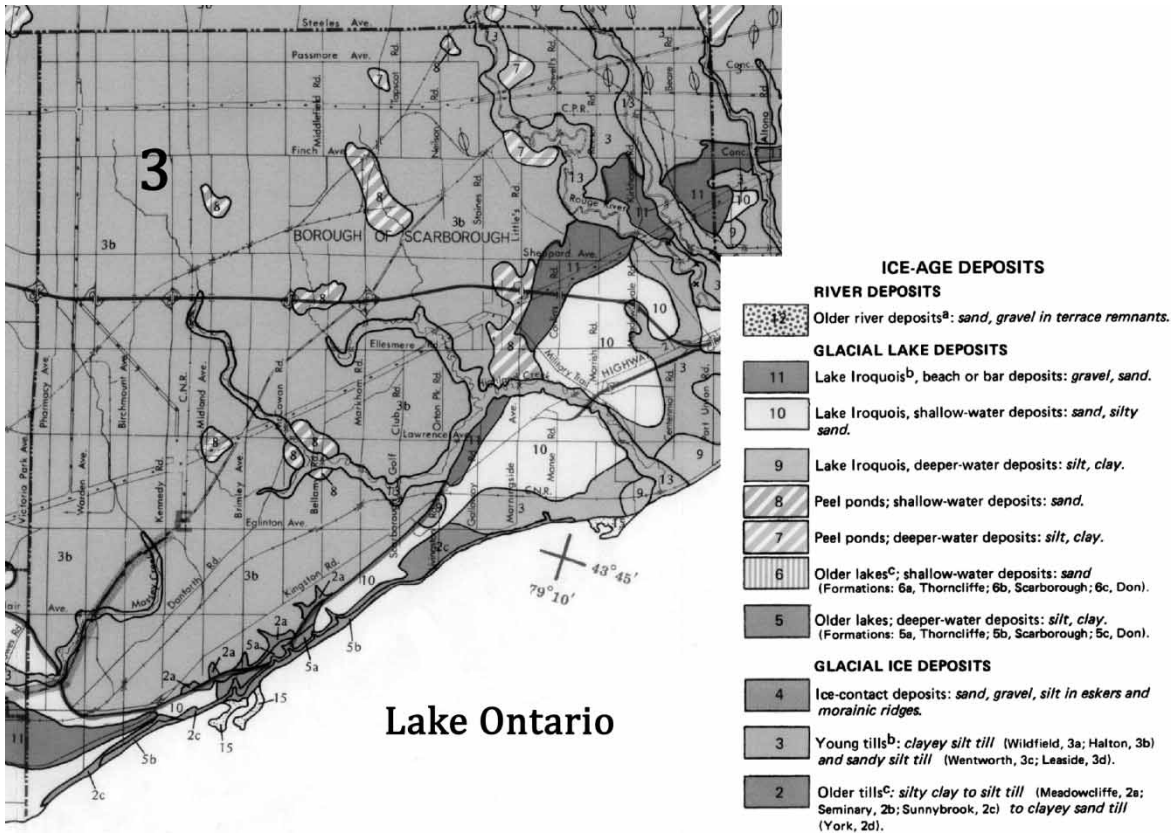

Figure 1 | Geology of the Scarborough area (scale about 1:100000). *Source:* Ontario Geological Survey document, Province of Ontario's Ministry of Northern Development and Mines (MNDM), http://www.geologyontario.mndm.gov.on.ca/mndmfiles/pub/data/imaging/P2204/p2204.pdf (May 2013).

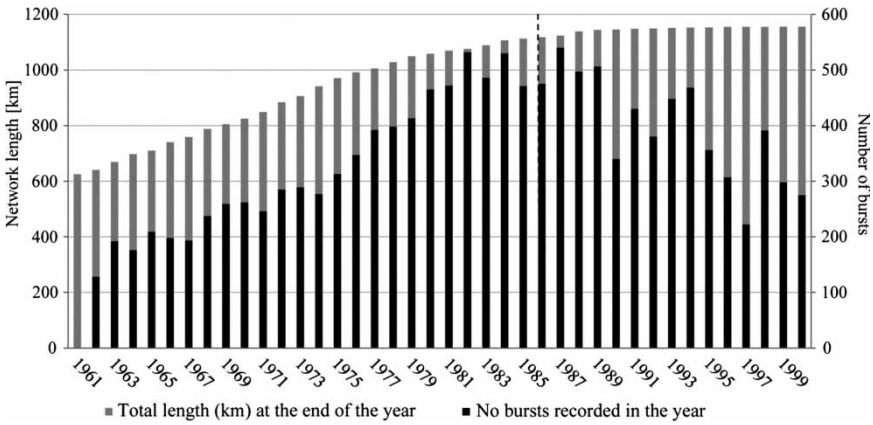

Figure 2 | Evolution of Scarborough network and related burst events.

was started in 1986, whereby individual cohorts of pipes were retrofitted systematically, according to identified need and available budget. These two programmes plausibly explain the down-trend of bursts after 1986; for this reason, 1986 has been considered as the threshold year to neglect the influence of cathodic protection on the occurrence of bursts in the modelling procedure here reported.

From the available database, only 150 mm diameter CI pipes are considered, which represents the largest relatively homogeneous cohort in the Scarborough pipe inventory (see Table 1). Table 2 summarizes all the relevant details about the data used in this research.

Note that the time period investigated (1962–1985) was deemed long enough to draw significant conclusions, while avoiding the issue of pipe cathodic protection. The first 22 years of data were used to identify most suitable models and the last 2 years were used for model validations.

PHYSICAL CONSIDERATIONS ON BURST GENERATION FOR CAST IRON

The various mechanisms that culminate in the ultimate failure of buried CI pipes have been described in detail by others, e.g., Rajani *et al.* (1996), Makar (2000), to name a few. In general, grey CI pipes fail because of one factor or a combination of factors that may include external loading, internal pressure, manufacturing flaws and corrosion damage. Most CI failures can be classified as circumferential cracking, longitudinal cracking, bell splitting (longitudinal break starting at the bell) or corrosion through-hole. The dominant failure modes of small-diameter CI pipes are related to external forces, mainly leading to breaks of the circular type, due to the brittle nature of CI, and climate-related factors such as frost penetration (vertical load causing bending moments) and soil shrinkage (frictional forces between external pipe surface and adhering soil) can make significant contributions to these forces.

Frost penetration can occur where the temperature drops below freezing during a sustained period of time. Water utilities typically endeavour to lay their water mains below frost depth (e.g., 1.5 m in Scarborough, 2.4 m in Ottawa). Frost penetration depth depends on the magnitude and duration of availability of ground moisture, ground temperature gradient (which in turn depends on ground temperature, snow cover, pavement type, ground cover, etc.) and soil properties (grain size distribution, compactness, etc.). Its precise depth is therefore difficult to estimate directly, and therefore a surrogate measure such as *FZI* is often used, representing the duration and severity of extreme air temperatures within a certain time interval, e.g., Kleiner & Rajani (2002).

Soil shrinkage is related to soil moisture, which leads expansive soils to expand when wet and shrink when dry. Expansion exerts vertical (typically crushing) loads on the pipe, while shrinkage induces shear stress. Gould *et al.* (2011) analysed failures of buried pipes in non-freezing environments where they identified a peak in circumferential failure rates during mid to late summer (longitudinal failure rate remained constant). They linked this elevated failure rate to differential soil movement as the result of shrinkage in expansive soils.

Table 2 | Main data features of the pipe cohort analysed

Pipe material	CI
Pipe diameter (mm)	150
Installation years	1921–1978
Bursts recording years	1962–1985
Cumulative length of failed pipes (km)	1,239.78
Number of recorded bursts	5177

DEFINITION OF CLIMATE-RELATED COVARIATES

Daily climate data for Scarborough were obtained from Environment Canada, and consisted of maximum, minimum and mean air temperatures (°C), total rain (mm) and snow on the ground (cm). Following Rajani *et al.* (2012), these data were pre-processed to obtain a number of candidate explanatory variables (or covariates) for the model. To avoid negative values, all temperatures were converted to degrees Kelvin (°K). Following Rajani *et al.* (2012), three time steps, namely, $TS = 5$, 15 and 30 days (non-overlapping) were examined. Longer time steps (90 days or more) result in fewer data points and require careful selection of a starting point in order to not miss seasonal temperature variations. Short time steps (up to 30 days) make the analysis insensitive to the starting point of the analysis period, which is especially important in short data sets.

The following covariates were used (note that each covariate is computed for each examined time step TS_i).

Daily temperatures

Average (over a time step) of mean daily temperatures (AT), minimum daily temperatures (AmT) and maximum daily temperatures (AMT):

$$AT(TS_i) = \frac{\sum_{j=1}^{m} T^j_{\text{mean}}(TS_i)}{m} \tag{1}$$

$$AmT(TS_i) = \frac{\sum_{j=1}^{m} T^j_{\text{min}}(TS_i)}{m} \tag{2}$$

$$AMT(TS_i) = \frac{\sum_{j=1}^{m} T^j_{\text{max}}(TS_i)}{m} \tag{3}$$

where m = number of days in time step i, and $T^j{}_{\text{mean}}(TS_i)$ = mean daily temperature of day j in TS_i; $T^j{}_{\text{min}}\ (TS_i)$ = minimum daily temperature of day j in TS_i; $T^j{}_{\text{max}}\ (TS_i)$ = maximum daily temperature of day j in TS_i.

Air temperature changes

Maximum temperature increase (MTI) and maximum temperature decrease (MTD) of the mean daily temperatures within time step i:

$$MTI(TS_i) = \max\left\{T^k_{\text{mean}} - T^j_{\text{mean}}\right\}(TS_i) \quad \forall j < k$$

where $j, k = (1, 2, \ldots, m)$ with m = number of days in TS_i (4)

$$MTD(TS_i) = \max\left\{T^j_{\text{mean}} - T^k_{\text{mean}}\right\}(TS_i) \quad \forall j < k$$

where $j, k = (1, 2, \ldots, m)$ with m = number of days in TS_i (5)

Intensities of air temperature changes

Quantifies the rate at which temperature changes over a consecutive number of days within a time step. Increase in temperature gradient (TIG) and decrease in temperature gradient (TDG) are given by

$$TIG(TS_i) = \max\left\{\frac{T^k_{\text{mean}} - T^j_{\text{mean}}}{k - j}\right\}(TS_i) \quad \forall j < k \tag{6}$$

$$TDG(TS_i) = \max\left\{\frac{\left(T^j_{\text{mean}} - T^k_{\text{mean}}\right)}{k - j}\right\}(TS_i) \quad \forall j < k$$

where $j, k = (1, 2, \ldots, m)$ (7)

Freezing index

Quantifies the duration and severity of extreme air temperatures within a time step which is expressed in degree-days, and defined as the cumulative minimum daily temperature below a specified air temperature threshold. The 'FZI' for time step i is computed as

$$FZI(TS_i) = \sum_{j=1}^{m}\left(\theta - T^j_{\text{min}}\right) \quad \forall T^j_{\text{min}} < \theta \tag{8}$$

where θ is the air temperature threshold, taken as 273.15 °K (i.e., 0 °C) in this research.

Daily variation of temperature

Average (over TS_i) of the maximum daily temperature minus minimum daily temperature

$$ADD(TS_i) = \sum_{j=1}^{m}\frac{\left(T^j_{\text{max}} - T^j_{\text{min}}\right)}{m} \tag{9}$$

Snow cover

Quantifies the effect of snow insulation and its consequence on frost penetration

$$ASC(TS_i) = \sum_{j=1}^{m}\frac{SG^j(TS_i)}{m} \tag{10}$$

where $SG^j(TS_i)$ = observed depth of snow cover on the ground on day j in TS_i.

Total rain

Indirectly estimates the soil moisture in the time step TS_i

$$TRN(TS_i) = \sum_{j=1}^{m} TR^j(TS_i) \tag{11}$$

where $TR^j(TS_i)$ = total rain for day j in TS_i.

While some of the covariates such as *AT*, *MTI*, *MTD*, *FZI*, *TIG* and *TDG*, as defined above were previously used by Rajani *et al*. (2012), the last three covariates are new. Rajani *et al*. (2012) applied the non-homogeneous Poisson-based pipe deterioration model using the first six covariates not distinguishing between warm and cold seasons. That approach could have masked possible variations between seasons in how covariates impact pipe bursts. Bursts occurring in warm and cold seasons were accounted with separate covariates in this study. Warm and cold seasons were distinguished by specifying arbitrary threshold *FZI* values of 0, 15 and 30 (degree-days) for time step sizes of 5, 15 and 30 days, respectively. For example, a month (or a 30-day period) was considered cold if its corresponding *FZI* was greater than 30. This working hypothesis was defined to roughly correspond to freezing vs. non-freezing periods in Ontario, Canada. The sensitivity of the models to these threshold values was not examined in this study.

Finally, it should be noted that the dependent variable here is the burst rate *BR* in units of m/km (length of the individual pipe on which a burst occurred is normalized by the total length of the pipe cohort). This enumeration of the dependent variable differs from Rajani *et al*. (2012) who used the more common enumeration of number of breaks/km. The m/km enumeration used here explicitly considers the length of the individual pipe affected (may be important at the decision making stage), even if it may mask possible occurrences of two or more breaks on an individual pipe.

EPR PARADIGM

Evolutionary Polynomial Regression (EPR) is a hybrid modelling technique that allows the exploration of polynomial models, where candidate covariates are evaluated on the basis of accuracy and parsimony. The technique helps identify the most important input covariates for the phenomena under study.

Typically, a pseudo-polynomial expression is used, where each term comprises a combination of the candidate inputs (covariates) and each covariate gets its own power (exponent) value. Each polynomial term is multiplied by a constant coefficient(s) to be determined during the search, and can include user-selected functions among a set of possible alternatives (which includes no function selection). During the evolutionary (genetic algorithm (GA)-based) search, the candidate power values are selected from a user-defined set of candidate values, which usually include a zero value as well (a covariate raised to the power of zero is actually excluded from the model – Giustolisi & Savic 2006). At each generation, all the candidate models have different number of terms and combination of inputs, and the constant coefficients are regressively determined using the available training set, and then the candidate models are selected based on a multi-objective scheme. The EPR evaluates a candidate model based on three criteria, namely (a) model accuracy (maximization of fitness to data), (b) parsimony of covariates (minimizing the number of explanatory variables included in final model expressions) and (c) parsimony of mathematical equation (minimization of the number of polynomial terms). EPR uses a multi-objective genetic algorithm (MOGA) to find candidate models and rank them by these three criteria (Giustolisi & Savic 2009; Laucelli & Giustolisi 2011). Once the models are obtained, their symbolic nature allows their validation in the light of the physical knowledge on the phenomena in hand.

All these features make the EPR modelling paradigm substantially different from purely regressive methods (e.g., ANN) where accuracy of model predictions is the only criterion to drive model selection and final mathematical expressions can be rarely validated from a physical perspective, as proved, among others, by the work of Savic *et al*. (2009).

MULTI-CASE STRATEGY FOR EPR

While the accuracy criterion for model evaluation is intuitively understood, the role of the parsimony criteria is to prevent over-fitting of model to data, and thus endeavour to capture underlying general phenomena without replicating noise in data. In some cases there is a need to model a certain phenomenon but the available data refer to different realizations of this phenomenon under various conditions/experiments. This can make it more difficult to identify the pattern among variables describing the underlying (i.e., main) phenomenon (Berardi & Kapelan 2007).

The exploration of the available data set by the multi-case strategy of EPR requires its partition into subsets, each representing a peculiar realization/experiment of the same phenomenon. Thus, the MCS-EPR simultaneously identifies the best pattern among significant explanatory variables to describe the same phenomenon in all data partitions, while neutralizing possible impacts of non-stationary conditions (Berardi *et al.* 2008). The MCS-EPR also makes use of the MOGA optimization scheme, as described above, where each candidate model structure is performed on each considered data partition (Berardi *et al.* 2008).

In this study the EPR paradigm was applied in two stages. In the first stage the 22-year training data set was partitioned into 22 1-year subsets, thus neglecting the effect of pipe aging. In the second stage, the effects of aging, number of bursts in the previous year, cumulative length of installed pipes and other relevant covariates were included. Each subset is subsequently divided into time steps, which have been discerned as belonging to the cold or warm seasons' sub-subsets.

The model accuracy (or coefficient of determination) is computed by

$$CoD = 1 - \frac{\sum_{i=1}^{C}\sum_{j=1}^{N_i}(\hat{y}_{i,j} - y_{i,j})^2}{\sum_{i=1}^{C}\sum_{j=1}^{N_i}(y_{i,j} - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^{C} N_i SSE_i}{\sum_{i=1}^{C}\sum_{j=1}^{N_i}(y_{i,j} - \bar{y}_i)^2} \quad (12)$$

where C is the number of data subsets, N_i is the number of time steps in data subset i, $y_{i,j}$ and $\hat{y}_{i,j}$ are the observed and predicted BR, respectively, in time step j of subset i, $\overline{y_i}$ is the average observed BR in subset i; and SSE_i is the sum of squared errors for data subset i (Savic *et al.* 2009). The candidate solutions/models are examined and ranked for accuracy and parsimony using the Pareto dominance criterion. In addition, the recurrent presence of certain covariates in several non-dominated models indicates the robustness of these covariates as potential explanatory variables of the phenomenon.

In this research, possible values for the power values are [−2, −1.5, −1, −0.5, 0, 0.5, 1, 1.5, 2], thus exploring well known direct and inverse possible relationships, e.g., linear, quadratic, inverse linear, square root, etc., for each term involved in the models. Such a modelling approach was mainly aimed towards finding more general formulations. The number of polynomial terms was set to $m = 1$ plus the bias a_0, in order to find compact expressions that are easier to interpret.

The MOGA process was set to 900 generations. The objective functions to be optimized are: (1) maximization of the model accuracy as measured by the means of the *CoD*; (2) minimization of the number of covariates, or explanatory variables. The optimal solutions/models returned on the Pareto front are shown in Tables 3 and 4. These optimal solutions obtained from the application of MCS-EPR are a trade-off between accuracy and parsimony.

COLD SEASON RESULTS

MSC-EPR was run for the three time steps *TS* of 5, 15 and 30 days. Table 3 provides the relevant information for the best candidate models obtained for each time step in the cold season. Note that the power values provided in Table 3 determine the relationship between each covariate and the number of bursts (e.g., direct or inverse, linear or not, etc.). Blank cells indicate power value of 0.

For $TS = 5$, the most significant explanatory variables are *FZI* and *ADD*, as they were selected by the MCS-EPR in almost all the candidate models. *FZI* has a linear relationship with the burst rate and this linear relationship is positive as expected. Daily temperature variation (*ADD*) on the other hand is inversely proportional; which means that the bursts rate gets higher as the temperature difference between daily maximum high and low temperatures get smaller. This occurrence might be explained by the fact that frost penetration is sustained during cold spells when temperatures do not change significantly. The covariate *AmT* also indicates an explanatory power of some significance, as it is selected in three out of five candidate models (inverse dependency).

The most significant explanatory variables for $TS = 15$ and 30 are also *FZI* and *ADD*. The *ADD* covariate has a non-zero power value, indicating more influence on the burst rate than observed in the model with *TS* of 5 days. In addition, the *TDG* covariate (temperature decrease intensity) emerges as having a substantial explanatory power, with a direct dependency of square root on the burst rate. This dependency corresponds to reported observations of increased pipe failures following sudden decreases in temperatures.

Table 3 | Power values of explanatory variables in the best models obtained for cold seasons

	TS = 5 days					*TS* = 15 days					*TS* = 30 days					
	Model number					Model number					Model number					
Inputs	**#1**	**#2**	**#3**	**#4**	**#5**	**#1**	**#2**	**#3**	**#4**	**#5**	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
AT				−1.5	−2					2						2
AmT			−2	−2	−2				2	1.5						−1.5
AMT					0.5					−1.5					2	2
MTI														−0.5	−0.5	−0.5
MTD																
TIG																
TDG								0.5	0.5	0.5			**0.5**	0.5	0.5	0.5
ADD		−1	−1	−1	−1		−1.5	2	−1.5	−1.5		−1.5	**−2**	−1.5	1.5	−1.5
FZI	1	1	1	1	1	1	1	1	1	1	1	1	**1**	1	1	1
ASC																
CoD	0.37	0.38	0.39	0.39	0.39	0.58	0.62	0.63	0.63	0.63	0.70	0.74	0.76	0.76	0.76	0.76

Table 4 | Power values of explanatory variables in the best models obtained for warm seasons

	TS = 5 days						*TS* = 15 days						*TS* = 30 days				
	Model number						Model number						Model number				
Inputs	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#1**	**#2**	**#3**	**#4**	**#5**
AT					−2	−1				−2	−2	−2					
AmT						−2						−0.5					
AMT			−2	−2	−2	−2			−2	−2	−2	−2					−2
MTI		2									0.5	0.5					
MTD		2	2	2	2	2	2	0.5	0.5	0.5	0.5	0.5	2	2	**2**	2	2
TIG				−0.5	−0.5	−2										1	1
TDG																	
ADD	−2			−1.5	−1.5	−1.5		−2	−2	−2	−2	−2		−2	**−2**	−2	−2
TRN															**−0.5**	−1	−1
COD	0.07	0.05	0.09	0.08	0.09	0.08	0.18	0.19	0.20	0.20	0.21	0.21	0.38	0.42	0.45	0.47	0.47

Table 3 (and Table 4 for that matter) shows that model accuracy, as represented by *CoD* tends to be lower for shorter time steps. This outcome can be explained as follows. While the observations $y_{i,j}$ are integers (number of bursts), the modelled values $\hat{y}_{i,j}$ are real numbers. In shorter time steps integer values $y_{i,j}$ are rather small and many of them (25% in the cold season) equal zero. The model candidates cannot yield $\hat{y}_{i,j} = 0$. As a result, *SSE* is expected to be high relative to the average number of observed bursts. As the time step size becomes longer, the number of observed pipe bursts $y_{i,j}$ becomes larger (with fewer zero) resulting in lower *SSE* relative to the average number of observed bursts (Rajani *et al*. 2012). In the limit of the present study, a time step $TS = 30$ days can be the best choice to have candidate models with good performance.

WARM SEASON RESULTS

Table 4 provides the results for the warm season. It is noted that accuracy is generally lower in the warm than in the cold seasons. This occurrence could be explained by the fact that soils in Scarborough are generally non-expansive and therefore wet-dry conditions during the warm season are not expected to cause increasing loads on the buried pipes (Kleiner & Rajani 2001; Gould *et al*. 2011).

Among the covariates that correspond to the warm season, the most significant explanatory variable appears to be the *MTD*, followed by *ADD*. The significance of the *MTD* means that in warm seasons, a strong temperature decrease can cause an increase in the pipe burst rate.

MODEL SELECTION AND FORECASTING CAPACITY

From the results provided in Tables 3 and 4, it was judged that the 30-day time step was the most appropriate, a conclusion also reached by Rajani *et al*. (2012). Additionally, two models were selected to represent the causal relationships between climate covariates and *BR*, one for each season, warm and cold. The selected model for the cold season is (bolded column in Table 3)

$$BR = a_1 \frac{FZI\sqrt{TDG}}{ADD^2} + a_0 \tag{13}$$

while the selected model for warm season is (bolded column in Table 4)

$$BR = a_1 \frac{MTD^2}{ADD^2\sqrt{TRN}} + a_0 \tag{14}$$

Each model has 22 realizations, with 22 sets of constant coefficients a_1 and a_0. Figures 3 and 4 show the values and trends of constant coefficients a_1 and a_0 respectively, for both models in Equations (13) and (14).

The values of both constant coefficients generally increase over time (with the exception of a_0 for the warm season, which displays no clear trend). This trend suggests that there is another mechanism(s) at work that influences pipe burst frequency, probably related to pipe aging and possibly also to the increase in total length of pipes installed over the years. In the second stage therefore, constant coefficients a_1 and a_0 are modelled using MOGA-EPR in conjunction with a new set of explanatory variables.

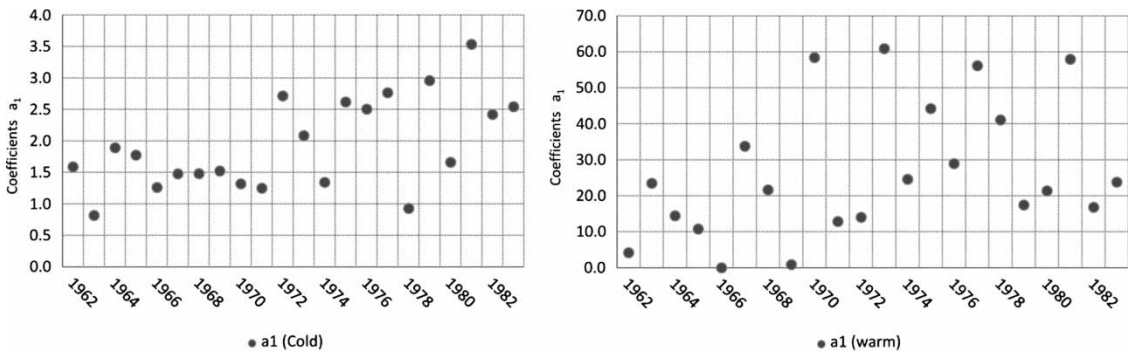

Figure 3 | Constant coefficients a_1 in Equation (13) (cold season) and Equation (14) (warm season) – $TS = 30$ days.

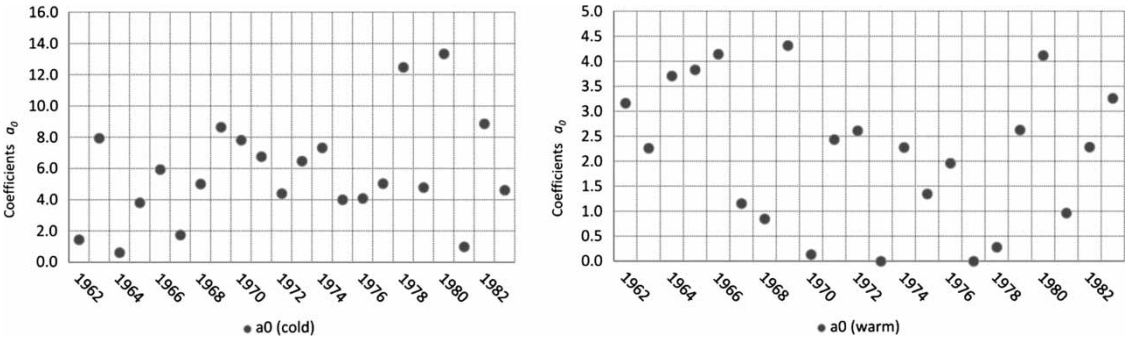

Figure 4 | Constant coefficients a_0 in Equation (13) (cold season) and Equation (14) (warm season) – $TS = 30$ days.

The explanatory variable AGE_t is the average age of the pipes in year t ($t = 1962, \ldots, 1983$) and is given by

$$AGE_t = \frac{\sum_{i=1}^{n_t} g_{i,t} \cdot L_i}{\sum_{i=1}^{n_t} L_i} \tag{15}$$

where $g_{i,t}$ is the age of individual pipe i (belonging to the considered pipe cohort) at year t; n_t is the total number of individual pipes at year t (some pipes have been installed in Scarborough during period 1962–1983) and L_i, is the length of individual pipe i.

Additional explanatory variables include burst rate of the previous year (PBR_{t-1}) in units of events/km, the number of bursts in the previous year (NBE_{t-1}), the total length of the pipes at the start of year t (CTL_t) and the length of pipes laid in the previous year (YLP_{t-1}). Table 5 provides the results of this second stage process. Note that the same parameters (power values, 900 generations, etc.) were used in the second stage MOGA-EPR as in the first stage, with the exception of $m = 2$ plus the bias a_0, which has been assumed in order to better exploit the few available data points for evaluation of coefficients.

The following models have been selected by the authors as adequate trade-offs between accuracy and parsimony. The coefficients a_1 and a_0 in Equation (13) for the cold season are

$$\begin{aligned} a_1\{\text{cold}\} &= 2.6 \cdot 10^{-7} \cdot YLP^{3/2} + 0.093AGE \\ a_0\{\text{cold}\} &= 2.05 \cdot 10^{-6} \cdot CTL^2 \cdot \sqrt{NBE} \end{aligned} \tag{16}$$

and the coefficients a_1 and a_0 in Equation (14) for the warm season are

$$\begin{aligned} a_1\{\text{warm}\} &= 0.007 \cdot NBE^{3/2} + 1.44 \cdot 10^{-5} TL^2 \\ a_0\{\text{warm}\} &= 0.047 \cdot PBR \cdot \sqrt{YLP} + 1.348 \end{aligned} \tag{17}$$

MODEL VALIDATIONS

The general models were tested on the years 1984 and 1985, for which data were available at a monthly resolution. The constant coefficients in Equations (13) and (14) for years 1984 and 1985 were determined using Equations (16) and (17), respectively. Subsequently, the burst rate for each month in the year was forecasted using Equation (13) for months during the cold season ($FZI > 30$) and Equation (14) for months during the warm season ($FZI \leq 30$). Figure 5 compares the predicted against observed burst rate values. The models appear to be able to predict observed values fairly well ($CoD = 0.79$).

CONCLUDING COMMENTS

The paper investigated the possible influence of climate-related factors on pipe burst frequency, using a data set from Scarborough (Ontario, Canada). Data comprising 22 years of pipe burst records were used to build predictive models and data comprising pipe bursts in the subsequent 2 years (holdout sample) were used to validate these models.

The investigation was carried out using the EPR modelling paradigm, which implements a multi-objective

Table 5 | Constant coefficients obtained for the best models to quantify a_1 and a_0, Equations (13) and (14)

Output variables				Input variables				
Cold season		Warm season		AGE (years)	NBE (–)	PBR (burst/km)	CTL (km)	YLP (m)
a_1	a_0	a_1	a_0					
1.591	1.450	4.184	3.167	10.60	99	0.250	394.55	17,018
0.820	7.943	23.489	2.262	11.33	114	0.282	404.77	10,221
1.892	0.627	14.458	3.711	12.18	169	0.412	410.24	5475
1.777	3.812	10.794	3.833	12.74	151	0.355	425.65	15,401
1.263	5.937	0.000	4.143	13.64	187	0.436	428.95	3306
1.478	1.748	33.788	1.157	14.43	171	0.392	435.95	7000
1.483	5.019	21.666	0.849	15.34	160	0.365	438.50	2547
1.525	8.654	0.921	4.319	16.32	188	0.428	439.18	686
1.321	7.828	58.440	0.135	17.31	224	0.510	439.49	309
1.250	6.773	12.880	2.437	18.24	228	0.517	441.16	1671
2.717	4.401	14.020	2.612	19.15	203	0.458	443.25	2088
2.087	6.476	60.932	0.000	20.13	251	0.566	443.82	568
1.344	7.330	24.595	2.280	21.11	231	0.520	444.35	534
2.619	4.007	44.236	1.348	22.09	221	0.497	444.74	390
2.507	4.098	28.969	1.962	23.05	249	0.559	445.40	654
2.767	5.036	56.213	0.000	24.05	255	0.573	445.40	0
0.929	12.481	41.090	0.281	25.00	266	0.596	446.35	948
2.961	4.787	17.441	2.627	25.98	269	0.602	446.79	444
1.663	13.351	21.385	4.120	26.96	293	0.655	447.15	364
3.537	0.992	57.996	0.966	27.95	299	0.668	447.35	200
2.421	8.876	16.828	2.288	28.95	263	0.588	447.35	0
2.544	4.620	23.824	3.264	29.93	277	0.619	447.52	167

genetic search algorithm, where the objective functions are accuracy (*CoD*) and parsimony (number of covariates). The investigation was conducted in two stages. In the first stage, the MCS-EPR was used to discern a common model structure explaining the influences of climate-related covariates on pipe bursts. Three time steps were examined for suitability of these covariates, and the 30-day time step was found to be the most appropriate. In the second stage MOGA-EPR was used to incorporate covariates related to pipe aging and pipe inventory changes over time. Separate models were established to predict *BR* in the cold and warm seasons.

Direct comparison between the results obtained here and those obtained by Rajani *et al.* (2012) is rather difficult due to various differences in methodological details (e.g., separation of seasons, aggregation of all vintages of CI pipes, the manner with which a solution quality is evaluated, the manner in which pipe aging was accounted for, etc.). However, some observations can be made. The data analysed in this paper correspond to three groups of data analysed in Rajani *et al.* (2012), namely groups CS1, CS2 and CS3. Rajani *et al.* (2012) found that the most significant covariates in these three groups were mean air temperature, which was found significant both with the likelihood ratio measure as well as with the adjusted coefficient of determination. The second most significant covariate for the three groups was the maximum air increase (the maximum difference between maximum and

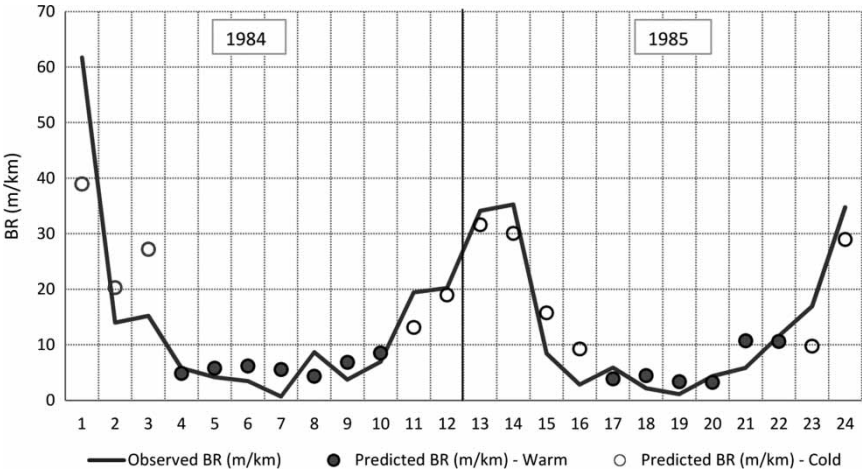

Figure 5 | Pipe burst rate predicted and observed values for years 1984 and 1985.

minimum daily temperature in a time step where the change is from low to high temperature), followed by the maximum air decrease. In this paper, the most significant covariate was found to be *FZI*, followed by the maximum temperature difference (*ADD*) in a time step. It can be argued that the *ADD* covariate in this paper is an aggregate of the maximum air increase and the maximum air decrease covariate in Rajani *et al.* (2012). Therefore, there appears to be at least a partial agreement between the two approaches.

The model for the cold season generally performed better than the warm season model. This is expected in Scarborough, which is subject to relatively cold winters. The most significant explanatory variables were freezing index (*FZI* – a surrogate for winter severity and potential frost loads) and the average daily temperature difference (*ADD* – difference between maximal and minimal daily temperatures) which represents the stability of a cold spell (smaller values exacerbate frost penetration).

The most significant covariates for the warm season model were found to be the *MTD* in a time step, *ADD* and the total rain. However, accuracy of this model was significantly lower, possibly implying that the overall phenomenon that underlies the generation of pipe bursts during warm seasons cannot be completely explained by the available climate-related data. Nonetheless, even with this deficiency, since *BR* in Scarborough are typically much lower in summer the validation results for the holdout sample (years 1984–1985) demonstrated a rather high accuracy level.

REFERENCES

Ahn, J. C., Lee, S. W., Lee, G. S. & Koo, J. Y. 2005 Predicting water pipe breaks using neural network. *Water Science and Technology: Water Supply* **5** (3–4), 159–172.

Aldrich, H. P. 1956 Frost penetration below highway and airfield pavements. *Highway Research Board* **36** (145), 125–144.

Bahmanyar, G. H. & Edil, T. B. 1983 Cold weather effects on underground pipeline failures. In: *Proceedings of Conference on Pipelines in Adverse Environments II* (M. B. Pickell, ed.). ASCE, Reston, Virginia, USA, pp. 579–593.

Baracos, A., Hurst, W. D. & Legget, R. F. 1955 Effects of physical environment on cast iron pipe. *Journal – American Water Works Association* **47** (12), 1195–1206.

Berardi, L. & Kapelan, Z. 2007 Multi-Case EPR strategy for the development of sewer failure performance indicators. In: *Proceedings of the World Environmental & Water Resources Congress*. ASCE, Reston, Virginia, USA, pp. 1–12.

Berardi, L., Kapelan, Z., Giustolisi, O. & Savic, D. 2008 Development of pipe deterioration models for water distribution systems using EPR. *Journal of Hydroinformatics* **10** (2), 113–126.

Clark, C. M. 1971 Expansive-soil effect on buried pipe. *Journal – American Water Works Association* **63**, 424–427.

Giustolisi, O. & Savic, D. A. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics* **8** (3), 207–222.

Giustolisi, O. & Berardi, L. 2007 Pipe level burst prediction using EPR and MCS-EPR. In: *Proceedings of Computer and Control in Water Industry (CCWI) – Water Management Challenges*

in Global Changes (B. Ulaniki, K. Vairavamoorthy, D. Butler, P. L. M. Bounds & F. A. Mermon, eds). Taylor & Francis Group, London, UK, pp. 39–46.

Giustolisi, O. & Savic, D. A. 2009 Advances in data-driven analyses and modelling using EPR-MOGA. *Journal of Hydroinformatics* **11** (3), 225–236.

Goodchild, C. W., Rowson, T. C. & Engelhardt, M. O. 2009 Making the earth move: modelling the impact of climate change on water pipeline serviceability. In: *Proceedings of Computing and Control in the Water Industry: Integrating Water Systems* (J. Boxall & C. Maksimovic, eds). University of Sheffield Press, Sheffield, UK, pp. 807–811.

Gould, S. J. F., Boulaire, F. A., Burn, S., Zhao, X. L. & Kodikara, J. K. 2011 Seasonal factors influencing the failure of buried water reticulation pipes. *Water Science and Technology* **63** (11), 2692–2699.

Habibian, A. 1994 Effect of temperature changes on water main break. *Journal of Transportation Engineering* **120** (2), 312–321.

Hudak, P., Sadler, B. & Hunter, B. 1998 Analyzing underground water-pipe breaks in residual soils. *Water Engineering and Management* **145** (12), 15–20.

Kleiner, Y. & Rajani, B. B. 2001 Comprehensive review of structural deterioration of water mains: physically-based models. *Urban Water* **3** (3), 151–164.

Kleiner, Y. & Rajani, B. B. 2002 Forecasting variations and trends in water main breaks. *Journal of Infrastructure Systems* **8** (4), 122–131.

Kleiner, Y. & Rajani, B. B. 2004 Quantifying effectiveness of cathodic protection in water mains: theory. *Journal of Infrastructure Systems* **10** (2), 43–51.

Kleiner, Y. & Rajani, B. B. 2009 I-WARP: individual water main renewal planner. In: *Proceedings of Computing and Control in the Water Industry: Integrating Water Systems* (J. Boxall & C. Maksimovic, eds). University of Sheffield Press, Sheffield, UK, pp. 639–644.

Laucelli, D. & Giustolisi, O. 2011 Scour depth modelling by a multi-objective evolutionary paradigm. *Environmental Modelling and Software* **26** (4), 498–509.

Lochbaum, B. S. 1993 PSE&G develops models to predict main breaks. *Pipeline and Gas Journal* **20** (9), 20–27.

Makar, J. M. 2000 A preliminary analysis of failures in grey cast iron water pipes. *Engineering Failure Analysis* **7**, 43–53.

Morris, R. E. 1967 Principal causes and remedies of water main breaks. *Journal – American Water Works Association* **54**, 782–798.

Newport, R. 1981 Factors influencing the occurrence of bursts in iron water mains. *Water Supply and Management* **3**, 274–278.

Rajani, B. & Zhan, C. 1996 On the estimation of frost load. *Canadian Geotechnical Journal* **33** (4), 629–641.

Rajani, B., Zhan, C. & Kuraoka, S. 1996 Pipe soil interaction analysis of jointed water mains. *Canadian Geotechnical Journal* **33** (3), 393–404.

Rajani, B., Kleiner, Y. & Sink, J. E. 2012 Exploration of the relationship between water main breaks and temperature covariates. *Urban Water Journal* **9** (2), 67–84.

Sacluti, F., Stanley, S. J. & Zhang, Q. 1999 Use of artificial neural networks to predict water distribution pipe breaks. In: *Proceedings of the 51st Annual Conference of the Western Canada Water and Wastewater*, Saskatoon, Saskatchewan, p. 12.

Savic, D. A., Giustolisi, O. & Laucelli, D. 2009 Asset performance analysis using multi-utility data and multi-objective data mining. *Journal of Hydroinformatics* **11** (3–4), 211–224.

Skipworth, P. J., Saul, A. & Engelhardt, M. O. 2000 Distribution network behaviour – extracting knowledge from data. In: Proceedings of International Symposium CWS 2000, University of Exeter, Exeter, UK.

Walski, T. M. & Pelliccia, A. 1982 Economic analysis of water main breaks. *Journal – American Water Works Association* **74** (3), 140–147.

First received 26 June 2013; accepted in revised form 4 November 2013. Available online 17 December 2013