

A Multivariate Geospatial Data-driven Approach to Descriptive Modelling of Burst Behaviour in a Small Island Context

T. Mackey¹, S. Mounce², J. Boxall², A. Cashman¹

¹ Centre for Resource Management and Environmental Studies, University of the West Indies,
Cave Hill Campus, Barbados

² Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield,
Sheffield, United Kingdom

¹tara.mackey@mycavehill.uwi.edu

ABSTRACT

The efficient management of water resources in the Caribbean is challenging for several reasons: weak institutional and regulatory frameworks, limited financial and technical resources, ageing pipe assets, average unaccounted-for water along with leakage rates that are approximately 50%, and limited data resources. This paper presented a methodology that used geographic information systems (GIS) to geo-spatially model 8 factors that contribute to pipe bursts (i.e. material, diameter, length, installation era, pressure, connection density, soil type, and rainfall) and incorporated these factors as input variables for data-mining using Self Organising Maps (SOM). SOMs is a data-driven technique used to explore complex multi-dimensional data even though data may be sparse or limited. The combination of GIS and SOMs allowed for the efficient use of available data and promoted knowledge discovery. Key findings revealed that pipe bursts in Barbados are mainly driven by pressure, clay soils, high rainfall, pipe age, and small pipe diameters. Although there were data limitations, results showed that this methodology can allow utilities, especially in small island states, to narrow their focus on susceptible areas of the pipeline, reduce futile efforts, maximise resources, and implement a technical tool that may be used before making any major investment decision.

Keywords: Pipe bursts, data-driven modelling, leakage

1 INTRODUCTION

In the Caribbean, the institutional and regulatory frameworks in which water utilities operate are described as weak, confusing, and redundant [1]. Nonetheless, most water utilities in these small island states are tasked not only to be the sole distributors of potable water in their respective countries [2] but are also expected to be as comparably efficient and well-managed as other longstanding water service providers worldwide. Globally, key performance indicators such as unaccounted-for water (UFW) and leakage rates are used to assess the efficiency of water distribution systems. In Europe and the USA leakage rates range between 6% and 30%. However, in the Caribbean, UFW and leakage rates range between 40% and 67% [1-3]. Challenges such as (a) increasing deterioration of buried ageing assets, (b) limited access to financial capital, (c) high operational and energy costs, (d) limiting technical capacity, and (e) major gaps in data availability and database management have been identified as common problems that utilities have to contend with in the Caribbean [1-3]. Although most of these challenges are inherent to any water utility, the consequences and effects are acute in small island states.

The aim of this paper was to develop a geospatial data-driven methodology that can assist utilities, especially in small island states, in gaining insight into deterioration behaviour that eventually lead

to bursts and leaks. The use of geographic information systems (GIS) combined with the data mining approach of Self Organising maps (SOMs) proved to be an innovative and effective method to promote knowledge discovery of factors that influence pipe bursts in the case study country of Barbados.

2 BACKGROUND

When compared to other small island states the potable water situation in Barbados (the most eastern and densely populated country in the Caribbean) is most dire as it is a water scarce country. It is using virtually all of its renewable water resources and in recent years has experienced periods of drought. One of the several initiatives of the local water utility, the Barbados Water Authority (BWA), to tackle this was a mains replacement programme. A preliminary study by Mackey et. al [4] was carried out for which the primary objective was to reduce non-revenue water losses that stood between 43% and 48%. The study attempted to identify hotspots in the pipe network in 2 out of the 16 districts by weighting and geospatially modelling some of the factors that contribute to pipe deterioration using a geographical information system (GIS). The approach was based on the concept that pipe bursts are a function of a combination of factors. This study builds on the previous preliminary work [4] with the broadening of the scope to include the entire pipe network of Barbados, a revision and addition of new explanatory factors, and the application of a more robust set of data analysis tools.

A variety of environmental and engineering factors can influence the occurrence of water mains bursts and leakages [5]. This study geospatially modelled the following factors as layers in a GIS: (1) reported pipe burst incidents, (2) pipe material, (3) pipe diameter, (4) pipe length, (5) pipe installation era, (6) pressure, (7) connection density, (8) soil type, and (9) rainfall. Some of these factors such as pipe material and pressure are commonly found in the literature as variables considered in pipe deterioration studies. However, even for these common core variables, assumptions had to be made as there were major gaps in the data. This limitation in key data variables also prompted the novel geospatial modelling of variables such as connection density in attempt to fill data gaps, make efficient use of available data, and promote knowledge discovery. GIS was a convenient tool that allowed for the geo-spatial capture of all the input factors promoting a practical and innovative approach to modelling pipe burst occurrence.

An analysis technique was required that had the ability to explore across complex multi-dimensional space, facilitate expert interpretation through readily accessible visual presentation and be robust to sparse datasets. Given these requirements, SOMs - a form of Artificial Neural Network (ANN) - was selected. A SOM resembles the way biological brain maps spatially order their responses by modelling those self-organising and adaptive learning features of the brain [6]. In unsupervised learning (also referred to as self-organisation) the inputs are presented to an ANN which forms its own clustering of the training data without any a priori knowledge. SOMs possess the properties of both vector quantization and vector projection algorithms [7]. SOMs are noise tolerant; this property is highly desirable when sparse data are used and thus there are many potential applications in the water industry.

3 METHODS

3.1 Data Collection and Compilation

The case study area consisted of all land area in Barbados. Thus, data were collected on an island-wide scale and ultimately grouped by sub-district since all pipes in the pipe network were categorised by sub-district. The input variables used for this study are shown in Table 1. All data were cleaned, transformed, normalised, digitised, created as a GIS layer, analysed, grouped, and compiled into a tabular database.

Table 1 - A descriptive list of input variables used in GIS and SOMs analyses.

Variable	Description
Pipe bursts	Total recorded pipe bursts (2002-2012) fixed by pipe diameter size for each sub-district (851 sub-districts)
Pipe diameter	Asset data of pipe diameter (in inches)
Pipe material	Asset data of type of pipe material (8 categories)
Installation era	Asset data of pipe installation era (5 eras)
Pipe length	Asset data of pipe length (in meters)
Pressure	Modelled data of average pressure (in meters of head)
Connection density	GIS-derived data for estimating connection density (i.e. no. of connections along a pipe divided by length)
Soil Type	Geo-referenced GIS data of soil type (11 types)
Rainfall	GIS-derived interpolation of average rainfall (in millimetres) from 2000-2015

3.2 Data Limitations

Data limitations are inherent to any study, especially for asset data of water utilities which commonly contain major gaps and inconsistencies. Several data limitations were encountered in this study; however, the most limiting are listed below:

- Pipe burst data was not available at the pipe-level (i.e. for each individual pipe). Bursts were only recorded by diameter size for each sub-district;
- Considerable gaps had to be filled to determine pipe material and pipe installation era as majority of that information was not initially attributed;
- Historical pipe bursts solely captured reported bursts for eleven years, between 2002 and 2012. Dates (i.e. month or year) of historical bursts and repair records were unavailable. As a result, a temporal aspect could not be included in the analysis. Other signs of deterioration such as impaired water quality and background leaks were also unavailable.

3.3 Creating GIS Layers

All input variables were created as layers in GIS. An overview of the methodology used to create each layer is described, and the coding of categorical and non-ordinal variables for the SOMs analysis is also given.

Bursts and Sub-districts

Bursts were recorded by diameter size for each sub-district. Each pipe (18,618 pipes in total, see Figure 1) was identified as belonging to a sub-district, however, spatial boundaries were not defined for the sub-districts. The Minimum Bounding Geometry tool in ArcGIS (Version 10.5.0.6491; ESRI ArcMap) tool was used to create polygons around pipes belonging to each sub-district. Initial and

final polygons representing the sub-districts after data cleaning were 834 and 851, respectively (see Figure 1).

Asset Data: Pipe Diameter, Pipe Length, Pipe Material, and Installation Era

Data collected for diameter and length at the pipe-level did not contain any missing entries or inconsistencies. However, for other asset data such as pipe material and pipe installation era, approximately 77% and 95% of the data entries were missing, respectively. Primary and secondary sources such as formalised rules of the local water utility and technical reports were used to fill most of the gaps in the data.

The coding of pipe materials used for the SOMs analysis were as follows: 1 – PVC, 2 – HDPE, 3 – DIP, 4 – CIP, 5 – Galvanised Iron, 6 – Steel, 7 – Asbestos Cement, and 8 – Unknown; that is, moving from plastics, to iron, and to other and unknown materials. The coding of installation era used for the SOMs analysis were as follows: 1 – 1990 to Present, 2 – 1970 to 1996, 3 - 1964, 4 – 1860 to 1969, and 5 – Unknown.

Modelled Pressure Data

Pressure was extracted from a recently built hydraulic model from the BWA. The average modelled pressure of each pipe was calculated from the modelled values of pressure at adjacent nodes in the hydraulic model (see Figure1). It is important to note that there were a few discrepancies with the hydraulic model.

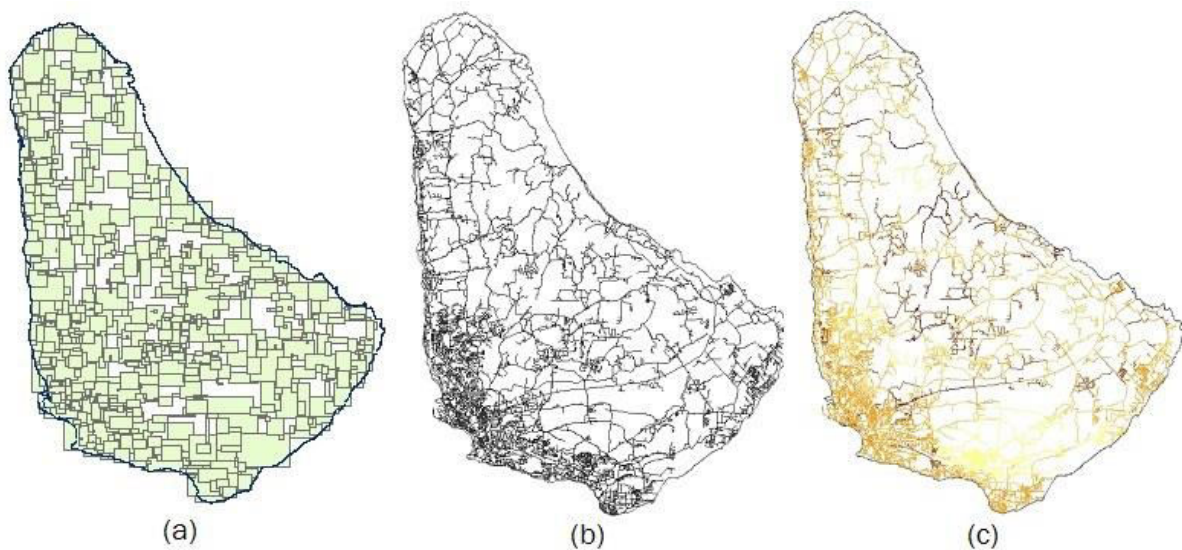


Figure 1 - GIS layers of (a) 851 polygons generated to delineate sub-districts; (b) entire pipe network of 18,618 pipes; (c) average modeled pressure, with higher values depicted by darker colours.

Connection Density

Connection density was depicted by dividing the number of connections along a pipe by the length of that pipe (see Figure 2). It was assumed that the geo-referenced meters that were nearest to a pipe represented a service connection of that pipe. This was accomplished using the Geoprocessing tool in ArcGIS.

Soil Type and Rainfall

A soil map of Barbados by Vernon and Carroll [8] was digitised in GIS to represent the 11 soil associations in Barbados (see Figure 2). The coding of soil types used for the SOMs analysis were numbered from 1 to 11 moving from coastal and sandy soils, to medium clay soils, and to the Scotland District clay soil. A GIS layer was created using the Interpolation tool of ArcGIS to depict average rainfall that was based on historical monthly rainfall data from 2000 to 2015 for 24 stations across the island (see Figure 2).

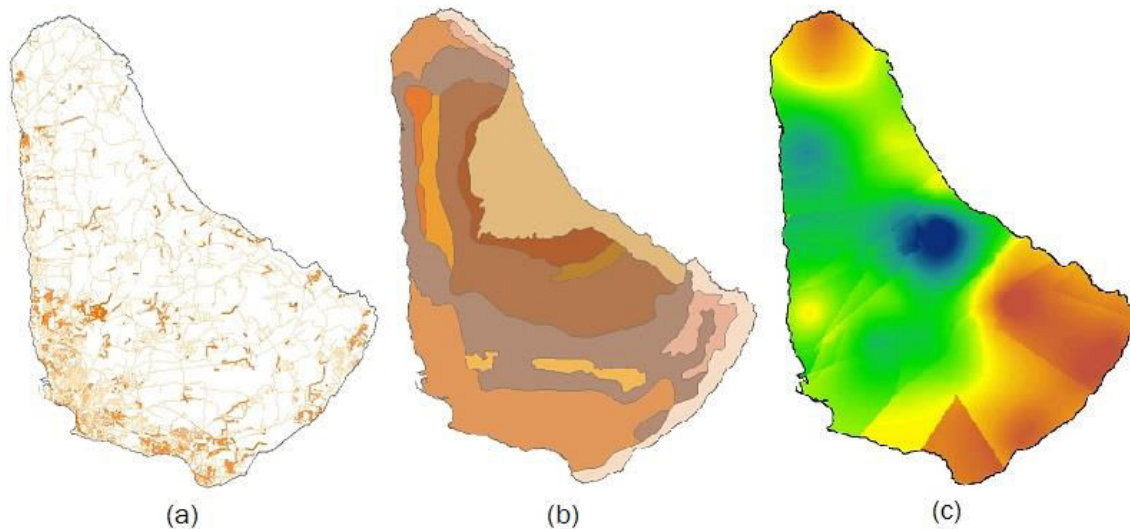


Figure 2 - GIS layers of (a) connection density, with higher values depicted by darker colours; (b) soil types; and (c) average rainfall, where wet and dry areas are depicted by blue and brown, respectively.

Final Database

The final database was formed by dissolving the attributes of the following GIS layers: pipe bursts, pipe diameter, pipe length, pipe material, installation era, connection density, and pressure by sub-district and diameter. The Zonal Calculations tool of ArcGIS was used to calculate the average value of rainfall and the dominant soil type for each sub-district. The data were joined by sub-district to obtain a final spatial and tabular database organising pipes by sub-district and diameter size, with all associated values of other input variables.

3.4 Application of SOMs

The SOMs for the assembled datasets were generated using the program MATLAB (Version 7.2.0.635; The Mathworks Inc.) using the SOM toolbox developed at the Helsinki University of Technology (available online at <http://www.cis.hut.fi/projects/somtoolbox>). The input layer consisted of a number of neurons corresponding to the number of variables used and the output layer consisted of a hexagonal Kohonen map whose size was optimally selected by the SOM toolbox. A batch training method was used with a Gaussian neighbourhood. The initial learning rate of 0.5 was used for the first rough phase of training corresponding to the creation of a ‘coarse’ mapping which is when the global order is imposed on the map. Later the learning rate is reduced to 0.05 for the second phase, in which the fine structure is added to the map.

A SOM can represent all input variables in the form of component planes. Each component plane is made up of colour-coded hexagons. The connections between the input and output layers are represented by each hexagonal cell. The colour coding of the cell refers to actual numerical values of the input variables for the particular component planes. The range of these values is shown as a scale to the right of the respective component plane. Low and high values correspond to shades of blue and red, respectively. Simple visual inspection of comparing component planes shows how relationships vary among all variables.

4 RESULTS AND DISCUSSION

4.1 Results of SOMs Analysis for all Sub-districts in Barbados

The results of the SOM analysis were generated to determine the relationships that exist between historical bursts (normalised by pipe length and grouped by diameter size) for each sub-district and each of the geo-spatially modelled input factors is shown in Figure 3. The cluster of high peak values in the bursts component plane corresponds to the high peak values of pressure, rainfall, and soil type (i.e. Scotland District clay). The exact alignment of peak values of bursts and pressure suggests that pressure is a dominant factor. Higher values of bursts also relate to CIP and DIP materials, and pipes that were installed between the eras of 1860 and 1964.

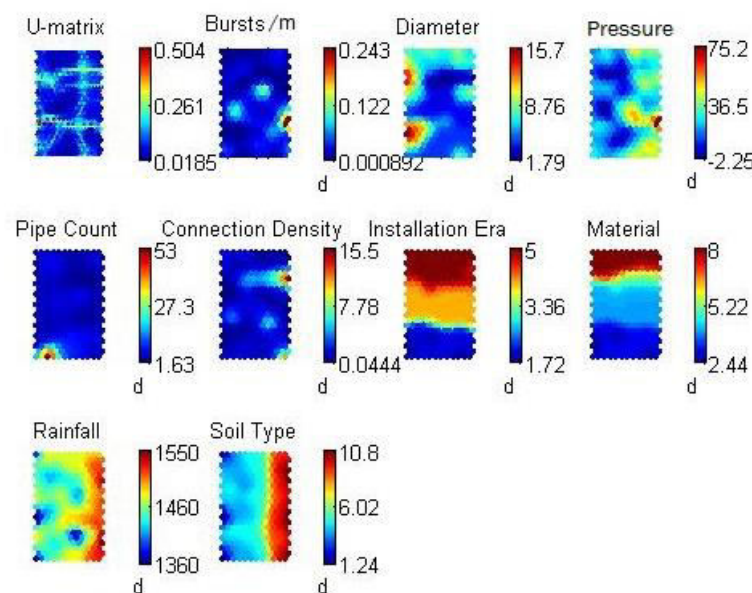


Figure 3 - SOMs results of physical, operational, and environmental factors influencing pipe bursts for all sub-districts in Barbados.

The parallel bands observed in the material and installation era component planes support that iron pipe materials (e.g. CIP, DIP, galvanised iron) were installed mostly between 1860 and 1969, with a transition to plastic pipe materials (HDPE and PVC) which were more recently installed.

The component planes of rainfall and soil type are similar in distribution with values increasing from left to right. That is, total rainfall is higher in inland ‘clay’ areas and lower in coastal areas. This follows the normal climatic distribution of rainfall in Barbados.

Interestingly pipes of smaller diameter sizes between 2 and 5 inches, which are mostly used as water service pipes, appear to be associated with areas of higher bursts, pressures, and connection density.

4.2 Implications of SOMs Analysis for all Sub-districts in Barbados

Observation of the bursts component plane in Figure 3 showed 3 distinct clusters – 1 of peak bursts values and 2 of moderate bursts values. Given that the cluster of peak bursts values also aligned with peak values of pressure, rainfall, and clay soils (particularly the Scotland District – which is also situated at the island's highest elevation), all of which are geo-spatially referenced data, it was postulated that these clusters can be isolated and referenced to the area representing these attributes. This sub-area in Figure 4 was used to identify the sub-districts (i.e. observations) that would have most likely contributed to this peak cluster of values. It followed that more insight may be obtained by splitting the model testing data into two: (i) Dataset 1: observations within Box A – likely associated with peak bursts values; and (ii) Dataset 2: observations outside of Box A – likely associated with moderate bursts values.

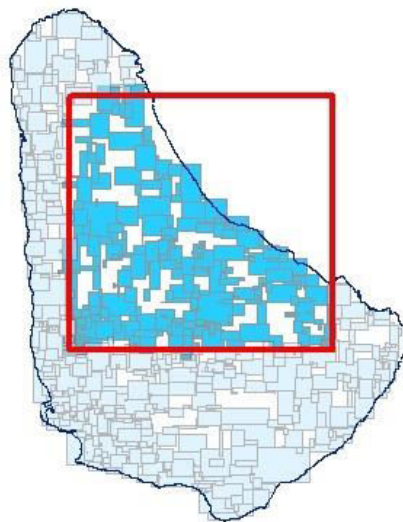


Figure 4 - Box A (in red) containing sub-districts likely associated with peak values of bursts, pressure, rainfall, and Scotland Clay District.

4.3 Results of SOMs Analysis for the Sets of Model Testing Data

The results for the SOMs analysis of Dataset 1 are shown in Figure 5. As expected, one cluster of peak values of bursts is aligned with peak values of pressure. This supports the observations by the BWA of pressure management issues in this area due to pressures being extremely high and exceeding standard operational practices for customers downhill, in order to service customers uphill. Other significant observations show that peak bursts are associated with 3-inch and 4-inch diameter pipes, and these pipes are situated in areas of higher rainfall within the Scotland District.

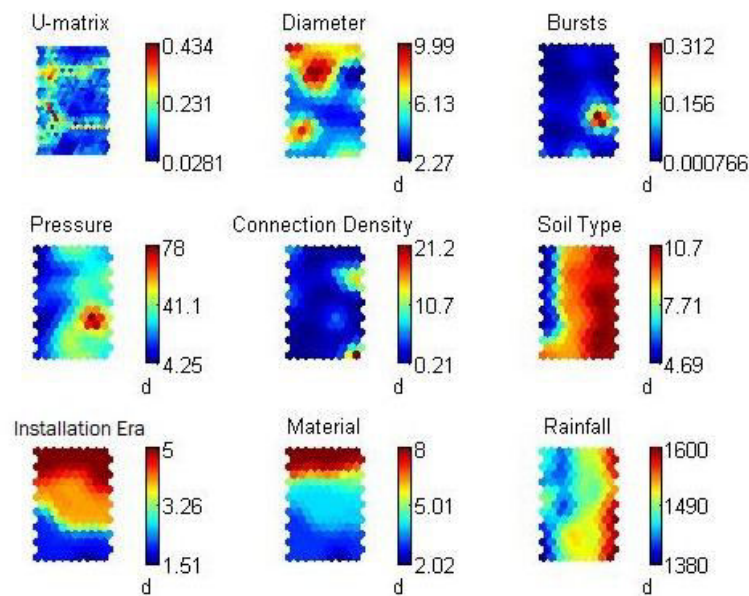


Figure 5 – Results of SOMs analysis of Dataset 1.

The results for the SOMs analysis of Dataset 2 are shown in Figure 6. The component plane for bursts showed 2 clusters that likely relate to the 2 clusters of moderate bursts values observed in Figure 1. The top-most burst cluster corresponds with pipes of mid-sized diameters between 6 and 9 inches, high pressure, low connection density, high rainfall, and installed within the past 40 years. Whereas the bottom-most cluster corresponds with pipes of smaller diameters between 2 and 4 inches, low pressure, mid-to-high connection density, moderate rainfall, and installed between 1860 and 1969.

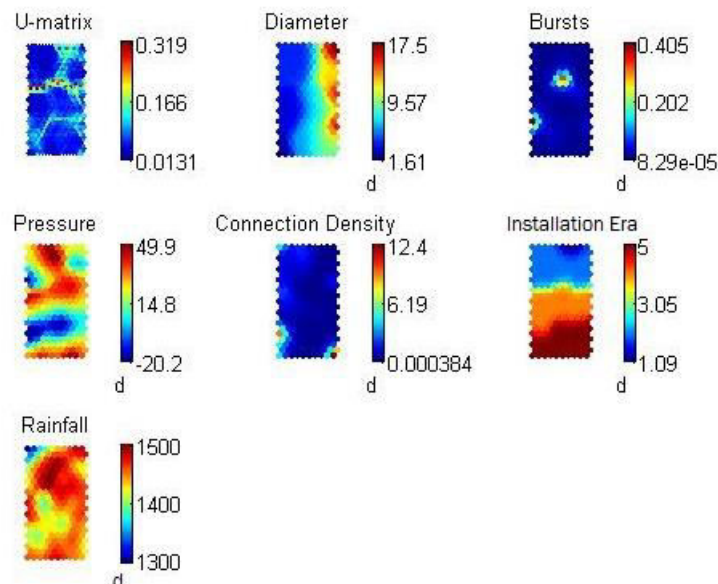


Figure 6 - Results of SOMs analysis of Dataset 2.

5 CONCLUSIONS

The use of GIS to model all input factors, especially operational and environmental factors which are typically excluded in the analysis of pipe bursts, proved an innovative and effective method to capture the contribution of factors such as connection density, soil type, and rainfall. Subsequently,

incorporating geo-spatially referenced data into a SOMs analysis also proved insightful and effective for the descriptive modelling of pipe bursts. Despite data limitations, the SOMs analysis identified the major drivers of pipe bursts at an island-wide scale. Splitting of the database into Dataset 1 and Dataset 2 allowed for the identification of possible key drivers of pipe bursts at smaller geographic scales.

Results showed the main contributing factors of reported pipe bursts in Barbados are high pressures, clay soils, high rainfall, aged mains, and smaller diameter pipes. Downscaling results identified that pipes in and around the Scotland District are aged and mostly affected by the high operational pressures. The swelling and shrinking of clay soils due to high rainfall may also play a major role in contributing to pipe bursts in this region. Bursts of pipes outside of the Scotland District seemed driven by other factors. Midsize diameter pipes that are likely distribution mains may burst with higher values of pressure and rainfall. Whereas bursting in smaller diameter service pipes may be attributed to pipe age and high connection density - factors which are characteristic of the densely populated city centres found in the low-lying, coastal areas.

Future work will consist of exploring quantitative data-driven methods that can complement SOMs so as to inform the hotspot mapping of areas of the pipe network that are most susceptible to bursts.

Acknowledgements

The authors are appreciative of the support of the Barbados Water Authority (BWA). This work was funded by the Australian Agency for International Development (AusAID).

References

- [1] W. Bank, "Institutions, Performance, and the Financing of Infrastructure Services in the Caribbean," World Bank, 2005.
- [2] GWP-C (Global Water Partnership Caribbean), *Water Financing Experiences in the Caribbean*, vol. 3, GWP-C, 2012.
- [3] A. Cashman, "Water Scarcity and Services in the Caribbean," *Water*, vol. 6, pp. 1187-1203, 2014.
- [4] T. Mackey, A. Cashman and R. Cumberbatch, "Identification of factors contributing to deterioration and losses in the water distribution system in Barbados," Centre for Resource Management and Environmental Studies, The University of the West Indies., Barbados, 2014.
- [5] J. Boxall, A. O'Hagan, S. Pooladsaz, A. Saul and D. Unwin, "Estimation of burst rates in water distribution mains," *Proceedings of the Institution of Civil Engineers*, vol. Water Management 160, no. WM2, pp. 73-82, 2007.
- [6] T. Kohonen, "The Self-Organising Map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990.
- [7] J. Kangas and T. Kohonen, "Developments and applications of the self-organising map and related algorithms," *Mathematics and Computers in Simulation*, vol. 41, no. 1-2, pp. 3-12, 1996.
- [8] K. Vernon and D. Carroll, "Soil and land-use Surveys No. 18 Barbados," The Regional Research Centre of the British Caribbean at the Imperial College of Tropical Agriculture, Trinidad and Tobago, 1966.