Environmental Modelling & Software 85 (2016) 156–171

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Can urban pluvial flooding be predicted by open spatial data and weather data?

S. Gaitan*, N.C. van de Giesen, J.A.E. ten Veldhuis

Department of Water Management, Delft University of Technology, Postbox 5048, 2600 GA, Delft, The Netherlands

ARTICLE INFO

Article history:
Received 19 January 2016
Received in revised form
11 August 2016
Accepted 11 August 2016

Keywords:
Urban flood modelling
Open spatial data
Data mining
Multivariate analysis
Climate change adaptation

ABSTRACT

Cities are increasingly prone to urban flooding due to heavier rainfall, denser populations, augmenting imperviousness, and infrastructure aging. Urban pluvial flooding causes damage to buildings and contents, and disturbs stormwater drainage, transportation, and electricity provision. Designing and implementing efficient adaptation measures requires proper understanding of the urban response to heavy rainfall. However, implemented stormwater drainage models lack flood impact data for calibration, which results in poor flood predictions. Moreover, such models only consider rainfall and hydraulic parameters, neglecting the role of other natural, built, and social conditions in flooding mechanisms. This paper explores the potential of open spatial datasets to explain the occurrence of citizen-reported flood incidents during a heavy rain event. After a dimensionality reduction, imperviousness and proximity to watershed outflow point were found to significantly explain up to half of the flooding incidents variability, proving the usefulness of the proposed approach for urban flood modelling and management.

1. Introduction

Cities are increasingly prone to urban flooding due to heavier rainfall, denser populations, augmenting imperviousness, and infrastructure aging (ten Veldhuis et al., 2011; Ashley et al., 2005). To overcome this challenge, cities need to design and implement proper and smart adaptation measures (e.g. Melo et al., 2015; Gaitan et al., 2014; Jacobs, 2012; ten Veldhuis et al., 2011; Wong and Brown, 2009). This requires a comprehensive understanding of the urban response to heavy rainfall events (Gaitan et al., 2015; Ochoa-Rodriguez et al., 2015; Spekkers et al., 2013; ten Veldhuis et al., 2011). Such understanding is limited by uncertainties in implemented drainage models and a lack of damage data (Freni et al., 2010).

Due to the lack of impact data, drainage models are often not calibrated and their uncertainty is poorly known (Dotto et al., 2012; Deletic et al., 2012), particularly for complex urban drainage systems. Uncertainties in currently implemented drainage models result in poor predictions of local floods occurrence during heavy rain events (Maksimović et al., 2009; Fontanazza et al., 2011; Gaitan et al., 2012; Ochoa-Rodriguez et al., 2015). Additionally, explaining urban flooding risks requires better understanding of additional factors such as the influence of natural, built, and social characteristics of the urban environment on flooding impacts (Cherqui et al., 2015).

1.1. Modelling of urban flooding risks and the use of open data

Recent works have used spatially distributed data to study the occurrence of pluvial flooding incidents and damage. Spekkers et al. (2014) have used decision tree analyses to determine to what extent multiple environmental and socio-economic variables can explain variability in insurance claim data, associated with rainfall-related damage. The developed model in that study explained close to 25% of variance in claim occurrence, improving from an 18% explained variance by multiple regression models. Gaitan et al. (2015) have analyzed citizens' complaints of local flooding incidents in relation to urban topography, finding no spatial autocorrelation in the location of complaints along overland flowpaths. Both studies suggest that pluvial flooding incidents, in the investigated Dutch areas, can only partly be explained in terms of rainfall intensity or urban topography. Merz et al. (2013) identified important variables influencing building direct damage due to river flooding using decision tree models and a thousand records dataset at a national level in Germany. Explanatory power of these tree-based models outperformed that of two linear models;

* Corresponding author.
E-mail address: S.Gaitan@tudelft.nl (S. Gaitan).

http://dx.doi.org/10.1016/j.envsoft.2016.08.007
1364-8152/

differences among performance of all models, however, were not statistically significant. Fontanazza et al. (2012) used Bayesian inference to reduce the uncertainty of depth-damage models on relatively small datasets applied to the city of Palermo. Uncertainty of damage estimation was reduced remarkably during the first and second (up to 40%) Bayesian updates, stabilizing by the third update, ensuring model robustness and reliability.

Spatial datasets of urban characteristics are becoming more attainable. Formerly scarce or inaccessible data-sources are currently available even as part of Open Data policies (Vitolo et al., 2015). In the case of The Netherlands, for example, open socioeconomic data has been aggregated into grids with 1 Ha or 0.25 km^2 cells. (e.g. Dutch Ministry of Interior and Kingdom Relations, 2014). The public availability, coverage, and spatial resolution of open data, enables flexibly using them in scientific research (Gaitan and ten Veldhuis, 2015). The integration of these heterogeneous data can be done at the Urban Water System level, ensuring an inter- and multidisciplinary approach for addressing urban floods (Bach et al., 2014).

1.2. Exploratory analysis techniques in heterogeneous spatial data

The use of exploratory tools, such as multivariate exploratory analysis and data mining techniques, is a key component for articulating existing, disparate models and data, under an integrated modelling approach Hamilton et al. (2015). There are different techniques that can be used to explore association patterns in multiple variables. Multivariate analyses can classify or ordinate multivariate information, or describe the response of a variable as a function of other functions. Classification techniques can provide insights about the structure of studied data by partitioning variable values into groups given their concurrence at sampling sites. Ordination analyses can be used to quantify the comparative variance of a set of multiple variables. Multiple regression analysis tests whether the distribution of a response variable is linked to a set of descriptor variables (Tongeren, 1995; Ramette, 2007; ter Braak, 1995; Legendre and Legendre, 2012a, b).

The aim of this paper was to assess the degree in which openly available datasets explain the occurrence of flood incident reports by using exploratory data analysis. To that end, classification, ordination, and regression techniques were applied to study the occurrence of flood incidents, using datasets representing a range of environmental and socioeconomic characteristics. Data and methods used for this study are presented in Section 2. Obtained results are presented and discussed in Section 3. Finally, conclusions are drawn in Section 4.

2. Data and methods

2.1. Data gathering and preprocessing

A highly localized, heavy rain event, with total rainfall varying from 125 to 140 mm in several rain gauges, and an estimated return period of 2000–5000 years (Netherlands Royal Meteorological Institute, 2014), hit the city of Amsterdam on July 28 to 29 2014. This event is used as case study in this work. Intensities peaked up to 100 mm/h during 15 min intervals in some areas of the city, causing considerable impacts such as interrupted highway traffic and tram lines, delays at Amsterdam airport, as well as flooded train stations and streets (see Waternet, 2015; Het Parool, 2014). During and shortly after the event, hundreds of citizen reports about location of flooding incidents were registered. These reports can be used as indicators of urban flooding incidence.

Meteorological, socioeconomic and cadastral spatial-data are available from open data sources. Rainfall intensities for 15 min and 60 min time windows, number of inhabitants per km^2 and average building age per km^2 were derived from these sources. Detailed descriptions of these data sources can be found in Gaitan and ten Veldhuis (2015). Additionally, this study also used polygon representations of water bodies and green areas coverage available from land registries, and a digital elevation model (DEM) from which average measures of imperviousness, distance to watershed outflow point and catchment area per km^2 were computed. Additional details about these variables can be found in following sections. The area of study was delimited by following the canals, highways, and train lines as close as possible to administrative borders. The goal of such delimitation was to allow the modelling of overland flowpaths to work on continuous paths. An overview of data characteristics is shown in Table 1.

Initial data clipping, the filtering of the digital elevation model, and the delineation of watersheds and overland flowpaths were done using ArcGIS 10, its spatial analyst tools (ESRI, 2012), and QGIS (QGIS Development Team, 2014). Data structuring and matrix algebra for the computation of overland flow path distances was done in Python (Python Software Foundation, 2014); reading of HDF5 weather radar imagery employed h5py (Collete, 2015), all remaining geographic data was read using Fiona (Gillies, 2014), and spatiotemporal queries were done with a combination of pyproj, Shapely, RTree, and Pandas (Whitaker, 2014; Gillies, 2013; Gillies et al., 2014; McKinney, 2015). Multivariate analysis were performed using the Vegan and stats packages in R (Oksanen et al., 2015; Ihaka and Gentleman, 2015).

Structuring the data for analysis and modelling required spatially aggregating studied data sources (Vitolo et al., 2015). All

Table 1
Data sources and variables (indicated with *s* and *v* respectively) used in this study. Total number, mean, and standard deviation of data points refer only to case study area.

Data source (*s*) or variable (*v*)	Spatial, temporal resolutions	Metric or unit	Data points, mean ± std. dev.
Incident reports (*s*)	Address points, time-stamped	Phone call register with address	336, mean and std. dev. N.A.
Max. rainfall intensity (*s*)	1 km^2, every 5 min	mm/h × km^2	292, 40.1 ± 20.5
Inhabitants (*s*)	1 Ha, year 2013	Individuals/Ha	6127, 131.2 ± 82.7
Age of construction (*s*)	Building polygons, year 2012	Years since built	234,736, 105.6 ± 217.6
Buildings area (*s*)	Building polygons, year 2012	m^2	34,952, 837.36 ± 2103.20
Roads area (*s*)	Single roads, year 2012	m^2	71,732, 488.84 ± 1038.21
Interpolated digital elevation model (*s*)	0.5 × 0.5 m grid	m	50,999 × 35,056, −1.25 ± 2.26
Aggregated incident report (*v*)	1 km^2 cells	Individuals/Ha	80, 5.30 ± 5.30
Maximum rainfall intensity at 15 min (*v*)	1 km^2 cells	mm/h × km^2	80, 15.13 ± 15.13
Maximum rainfall intensity at 60 min (*v*)	1 km^2 cells	mm/h × km^2	80, 8.67 ± 8.67
Average population density (*v*)	1 km^2 cells	Individuals/Ha	80, 103.32 ± 59.42
Average building age (*v*)	1 km^2 cells	Years since built	80, 124.38 ± 198.20
Impervious ratio (*v*)	1 km^2 cells	Ratio (dimensionless)	80, 0.43 ± 0.19
Average distance to outflow point (*v*)	1 km^2 cells	m	80, 459.54 ± 338.83
Average catchment size (*v*)	1 km^2 cells	m^2	80, 808.62 ± 664.27