Exploring urban flooding incidence through spatial information

A complementary view to support climate adaptation of lowland cities

Gaitan Sabogal, Santiago

DOI
10.4233/uuid:ee9e2ff4-256b-4ed7-8049-a9ab86820cc8

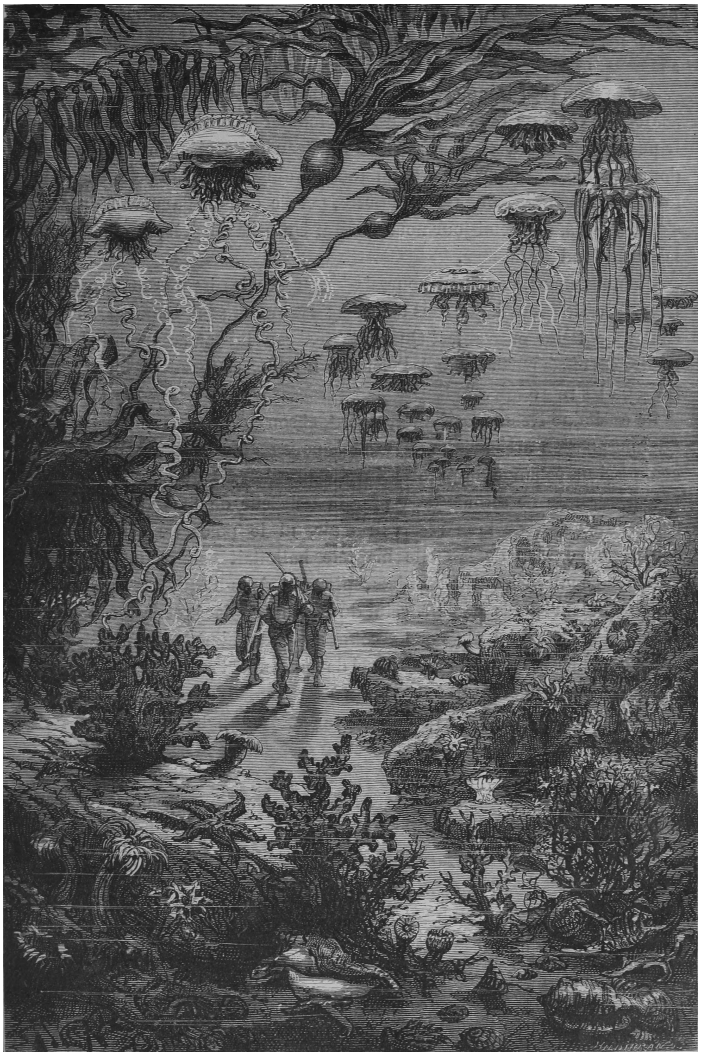
Publication date
2017

Document Version
Final published version

Citation (APA)
Gaitan Sabogal, S. (2017). *Exploring urban flooding incidence through spatial information: A complementary view to support climate adaptation of lowland cities*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ee9e2ff4-256b-4ed7-8049-a9ab86820cc8

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Exploring urban flooding incidence through spatial information



A complementary view to support climate adaptation of lowland cities

Exploring urban flooding incidence through spatial information

A complementary view to support climate adaptation of lowland cities

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 6 december 2017 om 12:30 uur

door

Santiago GAITAN SABOGAL

Bachelor of Science in Biology,
Nationale Universiteit van Colombia, Bogota, Colombia,
geboren te Bogota, Colombia.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. N.C. van de Giesen

Copromotor:

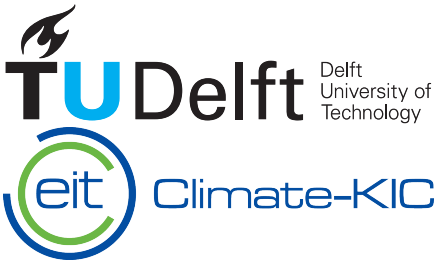
Dr. ir. J.A.E. ten Veldhuis

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. N.C. van de Giesen,	Technische Universiteit Delft
Dr. ir. J.A.E. ten Veldhuis,	Technische Universiteit Delft

Onafhankelijke leden:

Prof. dr. J. Aerts,	Vrije Universiteit Amsterdam
Dr. W. Buytaert,	Imperial College of London
Prof. dr. ir. A.P. de Vries,	Radboud Universiteit
Prof. dr. B. van de Walle,	Technische Universiteit Delft
Prof. dr. ir. M. Bakker,	Technische Universiteit Delft



Keywords: Urban flood modelling
Open spatial data
Data mining
Pattern recognition
Climate change adaptation

Printed by: Ipskamp.

Front & Back: Cover illustration by Alphonse de Neuville, extracted from a 1870 edition of Twenty Thousand Leagues Under the Sea.

ISBN 978-94-6186-877-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To my family.

Contents

Summary

Cities are vulnerable to local floods due to heavy rainfall. Urban flooding causes damage to buildings and contents, and also disturbs daily city activities as it entails drainage, transportation, and electricity interruptions. Urban flooding is expected to increase as climate change drives heavier rainfall events. Population and assets densification, as well as infrastructure aging, increasingly hamper cities from tackling pluvial flooding. Climate adaptation measures can help cities to face the challenge of heavier weather and urban flooding. Examples of those measures are: smart drainage maintenance and emergency responses, urban climate-proofing and retrofitting, and provision of real-time flooding information to citizens and government officials, among others. To plan and perform such measures it is required to know, and even predict before a heavy storm is onset, where, when, and why urban flooding occurs. This knowledge is not always available though. Required knowledge to design and implement adaptation measures against urban flooding is insufficient in cases such as Amsterdam and Rotterdam. In these cities, urban drainage models are limited to certain districts or uncalibrated; they cannot validly predict where or when the drainage system will surcharge or flood, and thus, they cannot be used for flood damage modeling. Moreover, urban flooding may not only depend on hydraulic parameters of underground drainage systems; other physical and socioeconomic characteristics of the urban fabric may also influence the flooding likelihood at a particular urban location. Urban flooding can be better understood by using non-hydraulic and unconventional sources of information. Available public data, curated by statistics, cadastral, or municipal call-center services, can provide insights about urban flooding damage. Using mainstream technology, such as web, traffic, and smart-phone cameras, can also afford for valuable data about urban flooding impacts, which contributes to the development of climate adaptation measures in lowland cities. This dissertation aimed to determine the potential of such alternative data sources in better explaining urban flooding incidents. Employed

methods combined techniques from geographic information systems, graph theory, community ecology, and computer vision. The exploration done in this research follows three main steps: testing previously proposed models, exploring currently available data sources, and evaluating the usefulness of attainable and affordable technology to gather key, nonexistent data about the timing, location, and extent of urban flooding incidents.

1

Introduction

1.1 Research context and objective

Cities are vulnerable to urban flooding due to heavy, localized rainfall. Urban flooding causes damage to buildings and contents, and disturbs daily city activities as it entails drainage, transportation, and electricity interruptions (Ashley et al., 2005; ten Veldhuis and Clemens, 2010). Urban flooding is expected to increase in many areas worldwide as climate change drives heavier rainfall events (Berggren et al., 2012). Forecasted climate change scenarios show that in The Netherlands, the intensity of extreme, short-termed, convective rainfall events leading to urban flooding is expected to increase (Attema et al., 2014; Romero et al., 2011).

Apart from the influence of heavier rainfall events, other factors may affect the occurrence and severity of urban flooding incidents. Population and assets densification, impervious covers advance, as well as infrastructure aging, increasingly hamper cities from tackling urban flooding. Lowland cities, with their characteristic geography, infrastructure layout, and growing population, are particularly vulnerable to urban flooding incidents. Delta cities, for instance, have 10-times the population density of the world average, hosting more than half a billion people and key infrastructure and services of global economic importance (Ericson et al., 2006). In The Netherlands, 18% of total surface corresponds to urban areas laying at or below 1 m below the sea level; these areas are inhabited by a third of the total country population (Center for International Earth Science Information Network - CIESIN - Columbia University, 2013). Urban drainage systems in The Netherlands have been designed designed to cope with rainfall events with return periods of 2 to 5 years (RIONED Foundation, 2004; ten Veldhuis and Clemens, 2010), which implies frequent urban flooding incidents.

Localized rainfall flooding can cause considerable damage. For instance, estimations of damages due to heavy rain in autumn 1998 in the Netherlands, accounted for €408 million (European Central Bank, 1998; Jak and Kok, 2000). Likewise, in the UK the annual average damage from intra-urban flooding is about a quarter of the total flood-related annual average damage (Blanc et al., 2012). Other studies claim that 40% of flood damage and associated economic losses are attributable to urban pluvial flooding (Douglas et al., 2010).

Climate adaptation measures can help cities to cope with the challenge of increasing urban flooding risks. Adaptation measures include climate-

proofing of urban infrastructure, such as the development of water plazas, underground water storages, green roofs and façades, and floating buildings (Jacobs, 2012). Implementing smarter drainage maintenance measures, by dynamically focusing on areas and sewer system components with higher incident occurrence (see ten Veldhuis et al., 2011), can also improve the prevention of urban flooding damage. Providing real-time information to support emergency responses, as illustrated in Melo et al. (2015); ten Veldhuis et al. (2011), is another strategy to cope with urban flooding impacts. Also, an insurance with vulnerability-graduated premiums, can secure repairing funds and catalyze mitigation investments (Botzen et al., 2009).

At present, planning and performing these measures is limited by an incomplete knowledge of the damage-generating process of urban floods. Proper measures require precise models about the location, timing, and causes of urban flooding incidents. Yet, damage models suffer from uncertainties in implemented drainage models and sparse damage data (Freni et al., 2010). Drainage models are often not calibrated (Dotto et al., 2012) and the associated uncertainty is either unknown (Dotto et al., 2011) or poorly known at best (Deletic et al., 2012). Uncertainties in currently implemented drainage models hinder realistic predictions of local flooding occurrence during heavy rain events (Fontanazza et al., 2011; Maksimović et al., 2009; Ochoa-Rodriguez et al., 2015).

Modeling urban flooding damage is particularly arduous in lowland environments where drainage networks are highly looped and rely heavily on pumps, as a result of negligible differences in terrain elevations. Besides, due to the subtle flooding depths and frequently short duration of flooding incidents, well-known stage damage functions used in river and coastal flooding are not applicable for urban flood risk assessment(ten Veldhuis and Clemens, 2011).

Under these circumstances, urban flood risk analysis cannot rely on the use of hydraulic models only. Alternative, non-hydraulic information sources must be considered. The use of crowd-sourced data is a promising source as it brings information about flooding timing and location, enriching the data produced by conventional rainfall or water level sensors. Natural and socioeconomic features of the built environment, which influence flooding incidence at particular urban locations, can be described by open and

public databases. Additionally, affordable, *ad hoc* technology, such as street video footage using mainstream cameras, can deliver valuable measurements about puddles due to urban flooding (Elmore et al., 2014; Horita et al., 2012; Lo et al., 2015; Michelsen et al., 2016; Muller, 2013; Muller et al., 2015; Shibata et al., 2014; Spekkers et al., 2012; ten Veldhuis et al., 2011).

Crowd sourced data can be used to analyze urban flooding incidents. ten Veldhuis et al. (2011) showed that reports about flood incidents made by citizens provide a valuable source of information for flood risk analysis. Reports can be used to analyze the impacts related to the typically subtle water-depths of pluvial floods and even account for intangible damages (Arthur et al., 2009; Caradot et al., 2010; ten Veldhuis, 2011; ten Veldhuis and Clemens, 2010). While these studies focused on report counts and textual content, they did not consider their geographic location. Using the position of citizen reports about urban flooding incidents, which is retrievable by geocoding reported addresses, offers a valuable information source for flooding analysis.

Public and open data include information about environmental, land-use, and social characteristics of urban environments. Statistical, cadastral, environmental, and municipal databases bear information about multiple, spatially distributed variables: income, housing market prices, building age and extents, roads, rainfall intensity and distribution, among others (e.g., Centraal Bureau voor de Statistiek, 2013; Kadaster Nederland, 2013; Netherlands Royal Meteorological Institute, 2013). Part of such information is provided in The Netherlands as part of open-data public policies (e.g., Dutch Ministry of Interior and Kingdom Relations, 2014).

Additionally, mainstream technology can deliver additional information about urban flooding incidents. Web, traffic, and smart-phone cameras, can provide valuable observations of street conditions (e.g., Horita et al., 2012; Shibata et al., 2014). Thus, street video footage, captured by conventional cameras, can be automatically processed to detect puddles, providing a valuable tools for the risk analysis and management of urban pluvial flooding.

This dissertation aims to determine the potential of crowdsourcing and open data sources to provide key information about the timing, location, and extent of urban flooding incidents. The employed methods combine techniques from geographic information systems, graph theory, community

ecology, and computer vision. Three research questions are tackled in chapters 2, 3, and 4. Chapter 2 statistically analyzes whether overland flow-paths constrain the spatial distribution of flood incidents in the case of a delta city, characterized by small ground level variations. Chapter 3 evaluates to which degree openly available spatial datasets, including environmental and socioeconomic information, explain the occurrence of flood incident reports by using exploratory data analysis. Chapter 4 assesses the potential of mainstream image and video recording, and well-known and accessible computer vision tools, to deliver key information about localized urban flooding incidents.

1.2 Thesis outline

The structure of this thesis is as follows. In Chapter 2, a highly detailed digital elevation model and and a set of citizen reports on flood incidents are used to evaluate the influence of urban topography on flood occurrence. The implemented analysis used the notion of urban watersheds and overland flow paths. The outcome of this test shows that in spite of the relevance of topography in other types of flooding, it is not a main factor for urban flood incidents in lowland cities. Evaluating other variables is required.

Chapter 3 develops a model for explaining variability of urban flooding incidents occurrence, based on multiple, publicly available, socio-environmental datasets. To achieve this, information patterns of the conditions underlying flood-prone areas were studied. Available information explained up to half of the flood incidents variability in the case study of Amsterdam. Even though this represented an improvement in the understanding of urban flooding incidents, it confirmed that currently available information is not enough to fully explain the occurrence of urban flooding incidents.

Chapter 4 presents an evaluation of mainstream and affordable technology to automatically gather information about the occurrence of urban flooding puddles. This included a proof of concept on the usefulness of well known computer vision techniques to automatically recognize flooding puddles in web images and videos recorded with conventional mainstream cameras. Findings showed that while certain easily implementable computer vision techniques achieved remarkable performances on automatically identifying web images containing puddles, the automatic detection of the

extent of the latter in video footage is not trivial, requiring a careful implementation with an elaborated machine learning framework that exceeds the scope of this thesis.

Finally, chapter 5 presents conclusions and recommendations. This chapter is of particular interest to government stakeholders, on decision making position on adaptation measures.

2

Flooding incidents along overland flowpaths

An increase of urban flood risks is expected for the following decades not only because climate is becoming more extreme, but also because population and asset densities in cities are increasing. There is a need for models that can explain the damage process of urban flooding and support damage prevention. Recent improvements in flood modeling have highlighted the importance of urban topography to properly describe the built environment. While such modeling has mainly focused on the hazard components of urban pluvial floods, the understanding of damage processes remains poor, mainly due to a lack of flood impact information. Citizen's reports about flood incidents can be used to describe urban flooding impacts. In this study a database of such type of reports and a digital elevation model are used as main inputs to analyze the relationships between urban topography and occurrence of pluvial flood impacts. After a delineation of urban subwatersheds at a district level, the amount of reports along the overland flow-paths is studied. Then, the spatial distribution of reports is statistically assessed at district and neighborhood levels, in Euclidean and network-constrained spaces. This novel implementation computes the connections of a network of subwatersheds to calculate overland flow-path gradient distances, which are used to test whether the location of reports is constrained by those gradients. Results indicate that while reports have a clear clustered spatial distribution over the study area, they are randomly distributed along overland flow-path gradients, suggesting that factors different from topography influence the occurrence of incidents.

This chapter is based on:

- *Gaitan, S., ten Veldhuis, J., Spekkers, M., and van de Giesen, N. C., 2012. Urban vulnerability to pluvial flooding: complaints location on overland flow routes. In Proceedings of the 2nd European Conference on Flood Risk Management FLOODrisk2012, Rotterdam, The Netherlands, 19-23 November 2012, pages 338–339, The Netherlands, 2012. CRC Press.*
- *Gaitan, S., ten Veldhuis, J., and van de Giesen, N., 2015. Spatial Distribution of Flood Incidents Along Urban Overland Flow-Paths. Water Resour Manage, pages 1–13, May 2015. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-015-1006-y.*

2.1 Introduction

The changes in precipitation patterns expected for the following decades (Bates et al., 2008; Hurk et al., 2006; Murphy et al., 2009; Romero et al., 2011), as well as urban growth, and higher population and assets densities, increase the risks of urban pluvial flooding (Ashley et al., 2005; ten Veldhuis and Clemens, 2009). This kind of floods can give rise to considerable damage in cities. Estimated damage due to heavy rain in autumn 1998 in the Netherlands accounted for 408 million Euros (European Central Bank, 1998; Jak and Kok, 2000). Likewise, in the UK the annual average damage from intra-urban flooding is about a quarter of the total flood-related annual average damage (Blanc et al., 2012). Other studies claim that 40% of flood damage and associated economic losses are attributable to pluvial flooding (Douglas et al., 2010). Such damage levels highlight the need for devising reliable models that can predict how heavy rains lead to pluvial flooding and damage.

There is relatively wide scientific knowledge covering hazard and damage modeling of coastal and river flooding (e.g., Apel et al., 2004, 2009; Aronica et al., 2002; Booij, 2005; Freni et al., 2010; Hoes and Schuurmans, 2006; Horritt and Bates, 2001; Jonkman et al., 2008a,b; Knebl et al., 2005; Kok et al., 2009; Maaskant et al., 2009; Merz et al., 2004; Pistrika and Jonkman, 2010).

Flooding in the urban environment, where overland flows depend on the complexity of the built infrastructure, is comparatively less studied. Recent availability of high resolution digital elevation models (DEMs) has allowed flood modeling research to explore urban topography to an increased level of detail (Bellos and Tsakiris, 2014; Diaz-Nieto et al., 2011; Dongquan et al., 2009; Jeong et al., 2010; Kunapo et al., 2009; Maksimović et al., 2009; Neal et al., 2011; Pistrika et al., 2014; Ravazzani et al., 2014; Tsakiris, 2014; Tsakiris and Bellos, 2014).

An important bottleneck in flood risk analysis is the scarcity of data about damages (Pistrika et al., 2014). Spekkers et al. (2014) analyzed damage reported in insurance claims and different environmental and socio-economic characteristics, which explained close to a quarter of the variance of claim occurrence. One of the reasons for this low explanatory power is the low spatial resolution of rainfall grids (1 km^2) and damage data (postal-district aggregations) used for the study.

Reports about flood incidents made by citizens, hereafter referred to as 'reports', provide a valuable source of information about flood occurrence and damage aspects. Reports can be used to analyze the impacts related to the typically subtle water-depths of pluvial floods and even account for the intangible caused damage (Arthur et al., 2009; Caradot et al., 2010; ten Veldhuis, 2011; ten Veldhuis and Clemens, 2010; ten Veldhuis et al., 2011).

In spite of the proven importance of topography in coastal and river flooding, and the availability of high resolution DEMs and flood reports, an analysis of the location of pluvial flooding incidents and the topographic conditions of the underlaying terrain has not been done yet. This work builds on results from previous exploratory analyses made at a municipal level, which displayed higher densities of reports counts in areas towards the outflow points of urban overland flow-paths (Gaitan et al., 2012). The present study statistically analyzes whether overland flow-paths constrain the spatial distribution of flood incidents in the case of a delta city, which is characterized by small ground level variations. This is a novel implementation that tests spatial autocorrelation on drainage distances between connected subwatersheds, including non-adjacent, along urban overland flow networks. This chapter is structured as follows: section 2 presents the area of study, data inputs and models used; section 3 presents and discusses the results, and conclusions are finally provided in section 4.

2.2 Data and Methods

The general approach used in this study is to aggregate reports into urban subwatersheds and then compute report counts and respective catchment areas. Those two variables are compared to determine if there are trends in the location and occurrence of reports over the underlying topographic conditions. The count of reports is used as a proxy of pluvial flooding damage. Locations towards the downstream end of intra-urban watersheds, which have bigger catchment areas, are likely to be exposed to higher overland flows during heavy rains, and therefore they are expected to account for higher reports occurrence.

2.2.1 Area of study

In this study, data for a set of urban catchments in Rotterdam are analyzed. Rotterdam is located along the final 40 km of the course of the New

Meuse river in the Rhine-Meuse Delta (Figure 2.1.a). It is one of the biggest cities in The Netherlands and has the largest European port. It is inhabited by close to 600 thousand people. Being a polder, its terrain elevations range from -6 to up to 10 meters above sea level. The city is a low lying environment, heavily urbanized, densely populated, vulnerable to pluvial flooding. Citizen's reports about rain-related incidents, as well as a very detailed digital elevation model (DEM) are available for research. Rotterdam's polder structure creates land areas with isolated surface waters, that enable straightforward overland-flows analysis. Ground level differences are small, with an average slope of 1.8% and standard deviation of 2.8%. In such flat terrains, flow-paths and watersheds can only be modeled from highly detailed DEMs. These characteristics make Rotterdam an interesting case for testing possible links between the location of flood reports and underlying terrain features. This study focuses on two different spatial scales: the District of Kralingen-Crooswijk, and the Neighborhood of Kralingen-West, covering aprox. 13 and 1 km^2 respectively. The first one will be referred to in this chapter as the 'district level', whereas the second as the 'neighborhood level'. Kralingen-Crooswijk is a district in Rotterdam comprising densely urbanized, industrial and park areas. Overland flows in this district are isolated from the adjacent areas. Only Rubroek, one of the district neighborhoods, shares overland flow with the Centrum District. This neighborhood was excluded from analysis. The neighborhood of Kralingen-West mostly consists of residential and commercial areas.

2.2.2 Available data sources

A database of transcripts of telephonic reports about pluvial flooding made by Rotterdam's inhabitants was made available for this study. It comprises 38,657 reports made from 2004 to 2011, and includes fields describing the neighborhood, street name, house number, short description of flooding incident, and reporting and problem solving dates. Of these, 36 registers did not have addresses, 12,663 did not have house number and could not be used for analysis, resulting in a final dataset of 25,958. A Python script was programmed to access and query the on-line Dutch public geo-information services (Publieke Dienstverlening Op de Kaart Loket, 2013) to geocode the reports having street name and house number. 21,577 reports were successfully geocoded. The remaining unrecognized 4,417 reports could not

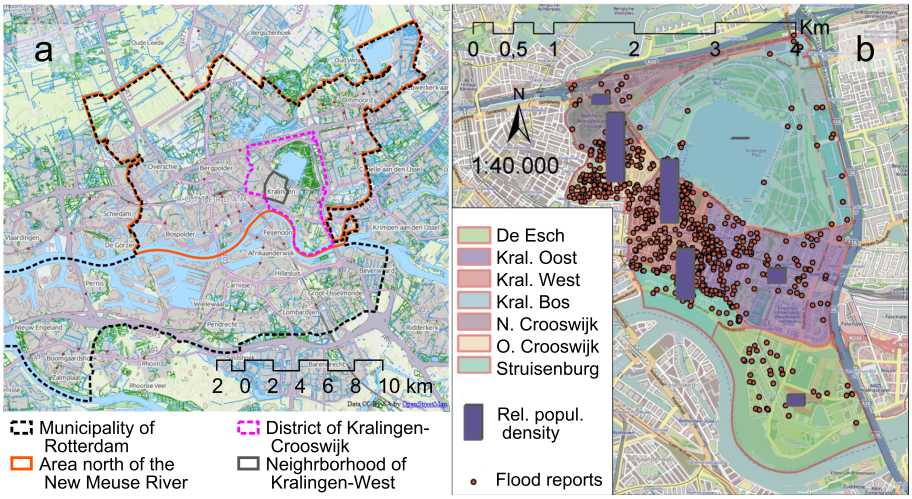


Figure 2.1: **a** View of eastern Rotterdam; municipal borders and areas of study are enclosed with different line colors **b** Visualization of flood reports and relative population density in neighborhoods of the Kralingen-Crooswijk district (Rubroek is excluded)

be used in the analysis either; they included 2,922 registers with zeros as house number and 1,495 registries with addresses that were not available in the public register. Additionally, a DEM was used. This DEM was produced by means of Light Detection and Ranging (LiDAR) of ground levels from an aerial platform. The DEM includes heights of urban objects such as streets, sideways, buildings, cars and trees. Some blank areas in the DEM, represented by no-data cells, are associated with signal noise due to response of wet surfaces, reflective materials and shadow effects at the base of tall objects with the LiDAR imaging. The DEM is characterized by a spatial resolution of 0.5 m $\times$ 0.5 m, a vertical precision of 5 cm, a systematic error of 5 cm, a random error of 5 cm, and a minimum precision under two standard deviations of 15 cm (Zon, 2011). A land-use maps was also available for Rotterdam. The map included polygons for each of the land-use classes.

2.2.3 Extraction of hydrological characteristics from the DEM

Some definitions are required for the rest of this chapter. The term "overland flow-paths" refers to the routes followed by rainfall running off over the watershed surface due to underlying slope aspects. A "subwatershed" refers to the hydrological subunits composing a watershed, that are discretized by drainage boundaries, and that drain into specific outflow points along the overland flow-paths of that watershed. In this work, those outflow points are set at a minimum drainage area threshold, which implies that sizes of enclosing areas of subwatersheds are generally similar. The area enclosed by the delineation of a subwatershed can be different from its drainage area. The former is simply the area enclosed within the subwatershed boundaries, while the latter is the total overland area draining into its outflow point including the drainage areas of upstream subwatersheds. The delineation of flow-paths and watersheds follows the approach proposed by Jenson and Domingue (1988) and Tarboton et al. (1991). Such delineation results in a tree-shaped network of subwatersheds that allows differentiating places in a city in terms of underlying overland drainage areas, which is suitable for analyzing the vulnerability of a given subwatershed to flooding as a result of depression-filling (ten Veldhuis et al., 2011). Pistrika et al. (2014) and Bellos and Tsakiris (2014) used DEMs, which include heights of building and other urban objects, for flood risk assessments in built-up areas to describe their topographic complexity. The following assumptions were made for the delineation of overland flow routes:

- Inputs and outputs from/to the underground sewer network are blocked or saturated. This assumption was also made by Diaz-Nieto et al. (2011). This implies that reports are assumed to be made during sewer surcharge or sewer blockage conditions (ten Veldhuis et al., 2011).

- Rainwater fallen on the buildings, tree canopies, and cars drains to the streets. The delineation of urban overland-flow routes is done on the basis of an elevation model, which includes urban features such as buildings, cars and trees. Changes of these features over time are not considered in this study. The used DEM represents the situation sensed by a series of LiDAR missions during 2008.

- Rainwater in surface water channels does not overflow onto the streets. Water in canals is supposed to be managed by a system independent from the sewers, which is normally the case in polder systems. Canals are considered as outputs of the overland-flow paths.
- The surface waters in the studied areas are isolated hydrological units, without interaction with adjacent hydrological units.

The DEM was prepared by first clipping the study areas and removing areas related to canals, lakes and rivers, using administrative and land used maps. Since the original DEM is a representation of the terrain under dry conditions (Zon, 2011), a direct processing of a run-off direction model would lead to a model composed of isolated urban ponds. With continuing rainfall, local ponds fill-up until the water exits by the lowest point of water divides, flowing into a nearby urban pond or into a body of water (Maksimović et al., 2009). For that reason, the DEM was treated with a filling process. This process raises the water levels of urban subwatersheds that initially do not have an outflow point, until they are connected to an urban water body or to another subwatershed. The run-off direction model was then processed from the prepared DEM to develop a flow accumulation model using the D8 algorithm (e.g., Olivera and Maidment, 1999; Tarboton et al., 1991). A threshold for the minimum flow accumulation value was established at a catchment area of 2,000 m^2. This is an area comparable to a 100 m long and 20 m wide street. This threshold allowed us to delineate subwatersheds. The ending point of each overland-flow route was considered as the exit point of the corresponding subwatershed.

Definition of non-adjacent connectivity

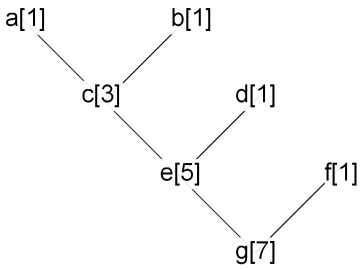
An example of a tree graph representation of the connections between subwatersheds is shown in Fig. 2.2.a. In this graph, each of the subwatersheds has one unit of enclosing area. Numbers in brackets indicate drainage areas at the exit of each subwatershed. c, for instance, has a drainage area of 3 units of area, which equals the sum of the enclosing areas of itself, a, and b. For the case of g, while its enclosed area is 1 unit, its drainage area equals 7 units, which is the sum of the areas enclosed by all the subwatersheds in this watershed. On the other hand, for f, which has no upstream subwatersheds, enclosed and drainage areas have the same size. An adjacency matrix was built for the full network of subwatersheds on the basis of the adjacent

connectivity along flow-paths. Figure 2.2.b shows the connectivity matrix of the tree presented in Figure 2.2.a. This matrix represents whether the subwatershed of a given row is connected downstream to another one of a given column; a value of 1 means there is a downstream connection; a value of 0 means the opposite. See, for example, that subwatersheds a and b are adjacently connected to c; the latter, however, only shows a connection to e. A watershed matrix can be computed from an adjacency matrix using the expression: $W = (I - A)^{-1}$, where A is the adjacency matrix, and I is the identity matrix of A. $(I - A)^{-1}$ is the inverse matrix of $(I - A)$. W accounts for the full downstream connectivity of subwatersheds; for this reason, it is different from the adjacency matrix, which only indicates adjacent connections. The watershed matrix in Fig. 2.2.c has been calculated using the adjacency matrix of Fig. 2.2.b. In this example, while a is connected to c, e, and g; g has no downstream connections. Upstream tributaries can be found by looking into the columns; column e, for example, shows that this subwatershed receives overland flows from a, b, c, and d. The watershed matrix permits identifying each of the studied trees and their internal connections. A watershed matrix was computed for the area of study to determine all possible non-adjacent, downstream connections between subwatersheds. This matrix was then used to compute the differences in catchment areas between connected subwatersheds.

2.2.4 Analysis of spatial distribution of reports in relation to overland-flow paths

Vulnerability due to depression filling is expected to be higher at locations catching higher overland inflows. Therefore, subwatersheds located further downstream the overland-flow network are expected to receive higher report counts than the ones located upstream. This hypothesis assumes that reports are not randomly distributed throughout the urban space. This can be checked by testing whether report data display spatial structure under a spatial weighting based on the overland-flow network. Spatial distances and units of analysis to be studied in such approach must take care of underlying overland-flow networks rather than Euclidean distances.

Three different tests were performed to assess whether the spatial distribution of reports displays patterns. Those tests were run at the district and neighborhood spatial scales mentioned in Section 2.2.1. First, a



(a)

$$A = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \end{array} \begin{array}{c} \begin{array}{ccccccc} a & b & c & d & e & f & g \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

(b)

$$W = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \end{array} \begin{array}{c} \begin{array}{ccccccc} a & b & c & d & e & f & g \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

(c)

Figure 2.2: **a** Example of a tree of subwatersheds. The arcs represent downstream connections between adjacent subwatersheds. Literals indicate arbitrary names given to the subwatersheds. The root of this tree is g. **b** Adjacency matrix (A) of the network presented in (a). Subwatersheds have been labelled in rows and columns. **c** Watershed matrix (W) of the tree in (a)

simple Average-Nearest-Neighbor test was applied for checking clustering of reports. In this test, the distance between the location of each report and its nearest neighbor is measured. An average for all the nearest neighbor distances of each report is then computed and compared with a random distribution. Further details of this method can be found in Illian et al. (2008, p. 126-127) and Sinclair (1985).

Then the magnitude of the spatial autocorrelation in reports aggregated into subwatersheds was tested using a Global Moran's I test. The report counts per subwatershed, and the distance between watersheds centroids on a Euclidean space, were used as input variables for this test. Further detail on Spatial Autocorrelation and the Global Moran's I test can be found in O'Sullivan and Unwin (2010, p. 195-206) and in Okabe and Sugihara (2012, p. 137-152). If the spatial distribution of reports is clustered given the

arrangement of subwatersheds, the Global Moran's I hypothesis of random distribution should be rejected.

As the overland flows between subwatersheds are determined by their connectivity, a second Globals Moran's I test was performed using "drainage distances" along overland flow-paths instead of Euclidean distances: the test was run only over pairs of subwatersheds found to be connected in the watershed matrix, and the distance used was the difference of their drainage areas. This type of distance quantifies the separation that two subwatersheds have in their relative position along the overland flow gradient. As an example, while the length of the two flow-paths connecting subwatersheds e and g, and subwatershed f and g, may be similar; the difference in catchment areas is 2 and 6 units respectively (see Fig. 2.2.a). In other words, two connected subwatersheds can be geographically close to each other, and still be wide apart in terms of the situation of their catchment sizes. Using the difference in catchment areas as a distance metric for the spatial autocorrelation test allows us to check if the occurrence of flood reports is influenced by the underlying overland drainage condition. Comparing the results of the Global Moran's I test on a Euclidean space with the ones constrained to the overland flow networks enable us to analyze the influence that depression filling may have in the occurrence and distribution of reports.

2.3 Results and Discussion

2.3.1 Computation of non-adjacent connections at the district level

After computing the watershed matrix, the number of independent trees identified was 1,717. There was an average of 3 subwatersheds per tree. The total number of actual connections between subwatersheds was 115,282. This large number can be explained by the increasing connections due to branching in a watershed. For example, in a single branched tree, made of 5 nodes, one of them being the unique leaf, the number of downstream connections equals 10: $\sum_{i=leaf}^{i=root} n_i$, where n is the amount of downstream nodes at each node. If it had two more leaves, the tree would have just two extra nodes, but the total number of connections would be 18, almost twice the original amount. In reality every branching does not occur at

Table 2.1: Description of clustering tests and results.

Test	Average nearest neighbor		Global Moran's I			
Aggregation level: single reports (S); subwatershed-aggregation (A)	S	S	A	A	A	A
Distance metric: Euclidean distance (E); Flow-path gradient distance (F)	E	E	E	E	F	F
Spatial scale of analysis: district level (D); neighborhood level (B)	D	B	D	B	D	B
z-score	-50.09	-19.75	2.93	1.29	0.05	0.23
p-value	0.00	0.00	0.00	0.19	0.95	0.81
Null hypothesis rejected at 99% confidence?: Yes (Y); No (N)	Y	Y	Y	N	N	N

the tips, but watershed networks are ideally more branched towards the tips. For the area of study, the presence of outliers with large numbers of subwatersheds can explain the large number of connections.

2.3.2 Testing of spatial patterns of reports distribution

Results obtained for the different performed clustering tests are presented in Table 2.1.

The average nearest neighborhood test, applied to non-aggregated reports, resulted in high z-scores of -50 and -20 at the district and neighborhood scales respectively. Associated p-values for both cases are extremely low. The magnitude of average distances between the nearest pairs of reports is higher at the district than the neighborhood scale. This result strongly suggests that single reports are not randomly distributed over the Euclidean space.

Results from the Global Moran's I test showed that the null hypothesis of a random pattern in the spatial distribution of subwatersheds-aggregated reports is rejected at the district level, but not at the neighborhood level under a confidence of 99%. However, there is 80% probability of spatial autocorrelation in the latter case.

Such patterns do not hold when the Global Moran's I test is constrained to the flow-paths gradient space. The hypothesis of reports being randomly distributed along overland flow-path gradients cannot be rejected. p-values,

at 0.95 and 0.81 for the district and neighborhood levels respectively, are far from being significant. These results clearly show that flood reports are clustered when observed in an open, Euclidean space, but this clustering is not related to the modeled overland flow gradient.

2.3.3 Discussion

Other factors that can explain the observed clustering are the distribution of urban mosaics composed by buildings, streets, and green areas; the spatial variation of population density; and the layout of water infrastructure, such as canals and sewers.

Differences in the urban mosaic composition can explain the clear rejection of the null hypothesis in the average nearest neighbor test at the District level. The extent of green areas is considerably different between neighborhoods; e.g., while the neighborhood of Kralingen Bos mainly consists of a park, Kralingen West hosts dense residential and commercial infrastructure (see Fig. 2.1). Highly impervious, dense residential are possibly more prone to local pluvial flooding than green areas, characterized by a higher permeability. Land uses of low imperviousness are not randomly distributed over the district; their location has been determined by urban planning and development processes, resulting in a permeability heterogeneity. This can explain the non-random pattern of reports locations at the District level.

Population density is another factor that can explain the outcome of the Nearest Average Neighbor test. Reports are made by citizens; therefore, more highly populated areas are likely to present higher report counts. In Fig. 2.1.b the comparatively low amount of reports in neighborhoods with lower population density is evident. This Figure also shows that areas with less green areas tend to account for higher populations.

Despite of being less strong than in the latter level, the z-score of the Average Nearest Neighbor at the neighborhood level is still substantial. Reports keep a strong clustering pattern within the neighborhood level. This suggests that the factors driving higher incidence of flood reports at the district level may also present a high spatial heterogeneity at the neighborhood scale. If imperviousness and population density heterogeneity are responsible for a structured spatial distribution of reports at the district levels, then results suggests that this heterogeneity is likely to be found,

and influencing the distribution of reports, in the neighborhood level.

Results of the second test are consistent with the latter. At the district scale, where the urban heterogeneity is greater, a clustering pattern is recognizable, despite the spatial aggregation into subwatersheds of approximately 2000 m^2. When the spatial level is focused into the neighborhood level, the effect of such aggregation is observed in a weaker, yet still considerable, p-value of 0.2. This suggests that an aggregation into 2000 m^2 subwatersheds regions tends to overlook the spatial structure clearly recognizable in the average neighborhood distance test. On the other hand, the weaker p-value can be also due to less marked variation of the factors influencing report occurrence at the neighborhood level. While subwatersheds are used to aggregate reports in this second test, the discussion about the influence of the overland flow gradient can be better made on the basis of this third test.

The third test demonstrates the strong lack of evidence to support the idea that incidence of reports is linked to overland flow-paths; reports occurrence has no preference for downstream locations along overland flow-paths. Such random spatial distribution is further explored in Fig. 2.3, which presents cumulative counts of subwatersheds', enclosing areas, and report counts for the district level. The increasing rate of reports closely follows the cumulative area, suggesting that reports occur evenly along the overland flow gradient. Reports density along such gradient (see bars in Fig. 2.3), does not show an increasing trend. Given the high number of reports per year, many of them are likely to be associated with relatively small rain events that do not trigger a depression-filling process. This is confirmed by results of ten Veldhuis et al. (2011), who found that local blockage of sewer inlets was the main reported cause of flood incidents, occurring even during small rainfall events.

This result can also be explained by the low sloping values of the city, which probably limits the onset of significant overland flows. Besides, the existence of canals throughout the city, which are heavily regulated by pumps, can mitigate the outbreak of puddles due to sewer blockage, malfunction, or overloading during heavy rain events. Serving as outflow receivers of overland flow-paths, canals can explain the low average of subwatersheds per tree in the studied area (see Section 2.3.1).

Discerning the effects of imperviousness, population density, and the

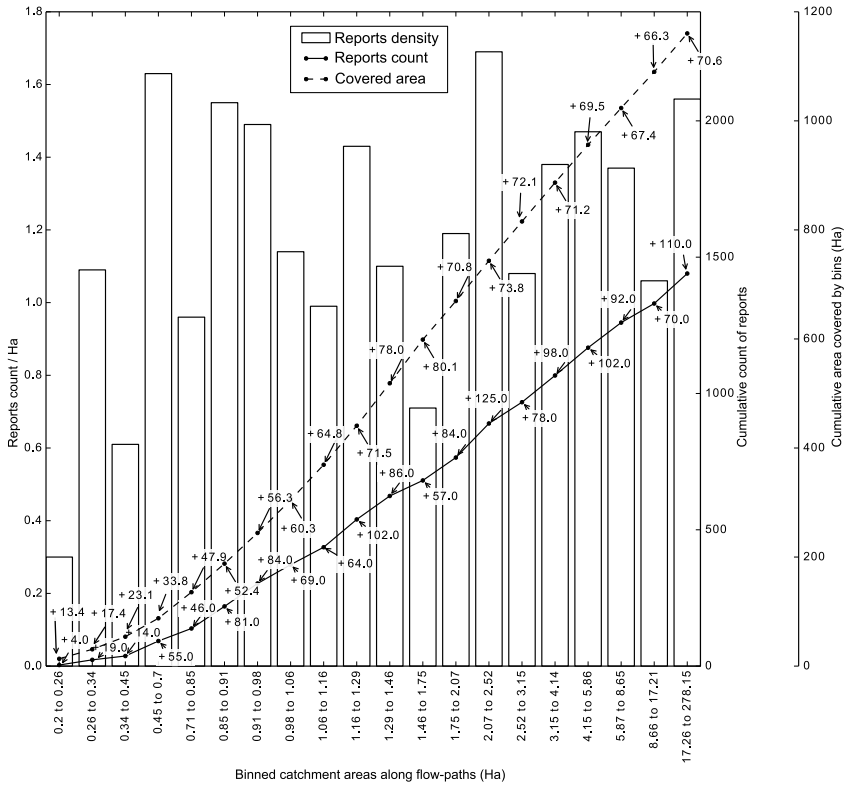


Figure 2.3: Cumulative sum of reports and area, and report density in binned drainage areas. Bins have the same number of subwatersheds.

proximity of a canal on the location of flood incidents cannot be done on the basis of the evidence obtained by this study, but it certainly is an analysis that might be revisited by future research.

2.4 Conclusions and outlook

In this chapter, the spatial distribution of reported local flooding incidents was investigated in relation to overland flow-paths and associated subwatersheds. Spatial clustering tests were performed on areas reportedly susceptible to urban flooding to determine if their location was linked to the underlying topographical conditions, in a city characterized by very low slopes. Those tests were based on datasets of flood reports and a highly de-

tailed DEM. The methodological implementation presented in this study can be used to test whether the spatial distribution of a variable is determined by the underlaying urban overland drainage conditions. In spite of the documented importance of topography in the analysis of flood occurrence and risks in environments from mild to accentuate slopes, this study showed that this factor does not determine the location of reported flood incidents in a polder environment such as Rotterdam. This conclusion follows from the results obtained from the Global Moran's I constrained to the flow-paths connection space. On the other hand, the average nearest neighbor test, and the Global Moran's I applied to the subwatershed-aggregation on a Euclidean space, probed that reports are definitely clustered. This suggests that factors different than the overland flow-path gradients, varying even at the sub-neighborhood scale, may contribute to the incidence of flood reports. Future research must focus in reports associated with heavy rain events, assessing the potential of population density, imperviousness, and water infrastructure to explain the occurrence of urban pluvial flood incidents.

3

Multiple spatial datasets to explain flooding incidents

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. Besides, such models only consider rainfall and hydraulic parameters, neglecting the role of other natural, built, and social conditions in flooding mechanisms. This chapter 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.

This chapter is based on:

- *Gaitan, S. and ten Veldhuis, J.A.E., 2015. Opportunities for multivariate analysis of open spatial datasets to characterize urban flooding risks. Proceedings of the International Association of Hydrological Sciences, 370:9–14, June 2015. ISSN 2199-899X. doi: 10.5194/piahs-370-9-2015.*
- *Gaitan, S., van de Giesen, N.C., ten Veldhuis, J.A.E., 2016 Can urban pluvial flooding be predicted by open spatial data and weather data? Environmental Modelling & Software 85, 156–171. doi:10.1016/j.envsoft .2016.08.007.*

3.1 Introduction

Cities are increasingly prone to urban flooding due to heavier rainfall, denser populations, augmenting imperviousness, and infrastructure aging (Ashley et al., 2005; ten Veldhuis et al., 2011). To overcome this challenge, cities need to design and implement proper and smart adaptation measures (e.g. Gaitan et al., 2014; Jacobs, 2012; Melo et al., 2015; 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 (Deletic et al., 2012; Dotto 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 (Fontanazza et al., 2011; Gaitan et al., 2012; Maksimović et al., 2009; 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).

3.1.1 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; 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 nowadays 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).

3.1.2 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 descriptor variables (ter Braak, 1995; Legendre and Legendre, 2012a,b; Ramette, 2007; Tongeren, 1995).

The aim of this chapter 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 3.2. Obtained results are presented and discussed in Section 3.3. Finally, conclusions are drawn in Section 3.4.

3.2 Data and Methods

3.2.1 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 to 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 Het Parool, 2014; Waternet, 2015). 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 3.1.

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 (Collette, 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 (Gillies, 2013; Gillies et al., 2014; McKinney, 2015; Whitaker, 2014). Multivariate analysis were performed using the Vegan and stats packages in R (Ihaka and Gentleman, 2015; Oksanen et al., 2015).

Structuring the data for analysis and modelling required spatially aggregating studied data sources (Vitolo et al., 2015). All spatial data were aggregated to the grid used in the weather radar imagery (see Section 3.2.1 for more details). Grid-cells were considered the units of analysis in this study, and are referred to as “sites" in this chapter (Gaitan and ten Veldhuis, 2015). Figure 3.3 shows the layout of the grid over the city of Amsterdam. An overview of value distributions and correlations of studied variables is presented in Figure 3.4.

Table 3.1: Data sources and variables (indicated with s and v respectively) used in this study. Variables were processed from data sources. 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	234736, 105.6 ± 217.6
Buildings area (s)	Building polygons, year 2012	m^2	34952, 837.36 ± 2103.20
Roads area (s)	Single roads, year 2012	m^2	71732, 488.84 ± 1038.21
Interpolated digital elevation model (s)	0.5×0.5 m grid	m	50999×35056, -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

Incident reports (Section 3.2.1) were used as response variable in this chapter; all other variables (Sections 3.2.1 to 3.2.1) were considered as descriptor variables. We used the terms response and dependent, and descriptor and independent, interchangeably throughout this chapter.

Flood incident reports

The Amsterdam water authority maintains a call-line for registering citizen's reports about urban drainage issues. This register was used in this chapter. Reports are initially classified by personnel in the call-center. Reports include fields for a unique identifier, time, address, a succinct textual description of the incident, and the response operation taken. By querying the Dutch public OpenLS geo-information services (Publieke Dienstverlening Op de Kaart Loket, 2013) with a Python script, addresses of reports were translated to projected coordinates according to Gaitan et al. (2015).

To use incident reports as response variable we needed to make two assumptions. First, we considered the number of incident reports as a proxy for the impact of urban flooding; an incident location reported several times during a heavy rainfall event is probably referring to a more severe impact than that of an incident reported just once. Second, we assume that available reports are a representative sample of the total set of flood incidents. It is possible that, during urban flooding emergencies, phone lines become saturated, and incident reports are missed. Other reports can be made to phone-lines different than that of the water authority: e.g., to the firebrigade, to the police, to the municipality.

Rainfall intensity

Rainfall intensity measurements are derived from two C-band Doppler weather radars operated by the Netherlands Royal Meteorological Institute (2013). Rainfall depths are provided with a temporal resolution of 5 min, on a grid of 1 km^2 spatial resolution using a custom geographic projection (Overeem et al., 2009b). Information is available since 08:00:00 1 January 2008 UTC, through a FTP server in HDF5 format (Roozekrans and Holleman, 2003). This data set was aggregated into 15 min and 60 min steps, as suggested by Overeem et al. (2009a). Maximum rainfall intensity during the event is computed for both aggregations.

When reprojected to the standard Dutch coordinate system (Amersfoort/RD New, EPSG:28992), the area of the radar cells became variable, and ap-

proximately 8.5% smaller than the 1 km^2 in the custom KNMI projection. This was taken into account for calculations including the other variables, which are provided with the Dutch projection by default.

Socioeconomic and cadastral data

As population density and building age were found to be the socioeconomic variables that better defined the multivariate data structure in the area of analysis Gaitan and ten Veldhuis (2015), these variables are selected from the sets available at the Central Bureau for Statistics (Centraal Bureau voor de Statistiek, 2013) and the Netherlands Land Registry (Kadaster Nederland, 2013). Spatial aggregation and sampling of these two variables was done following Gaitan and ten Veldhuis (2015). Points with secret or not available data have been excluded from the analysis. Only buildings in use have been considered. Additionally, cadastral geoinformation was used to model the imperviousness ratio and to filter the DEM as described in following sections.

Imperviousness

The ratio of highly impervious areas per site is used as a proxy for its imperviousness. In this study, paved areas were considered as highly impervious. Dutch cadastral data includes a land use model, composed of layers representing buildings, roads, green areas, and water bodies, among others. Roads include most streets, squares, and wide bike- and foot-paths. Green areas include parks, and other open, unpaved spaces (Publieke Dienstverlening Op de Kaart Loket, 2015). An estimation of the highly impervious area per cell was obtained by computing the sum of road and building areas intersecting each cell. Green areas and water bodies are not included in this sum. Obtained values were then divided by the total cell area. This is, thus, a broad representation of the overall imperviousness trend in each cell.

Distance to watershed outflow point

Water bodies, such as canals and urban ponds, receive rainfall that runs off from adjacent areas, following the urban topography. Therefore, it is expectable that flooding occurrence in the proximity of such outflow points is different than at farther locations. Defining a quantitative measure of the proximity to watershed outflow points required computing urban overland

flowpath networks (see Gaitan et al., 2015). Additionally, a buffer of 10 km around the limits of the city of Amsterdam was set to the extent of the input DEM to ensure no flowpaths were cut by the city boderline. The DEM used is the AHN2 (Publieke Dienstverlening Op de Kaart Loket, 2014), with spatial resolution of 0.5 m $\times$ 0.5 m and a minimum precision of 15 cm within two standard deviations (Zon, 2011).

Null pixels in the AHN2 model, associated with signal shadows and noise, were filtered combining the raw and interpolated AHN2's versions, and the Land Registry's vector representations of buildings and water areas such as ponds and canals. This filtering was devised and implemented especially for this chapter. Filtering was applied as follows. Null values in the interpolated version were replaced by raw values in building areas. Null values in water areas were preserved. Remaining null values were replaced by an iterative filtering. They were assigned with the average of non-null values within a window of 11 $\times$ 11 pixels, centered on the treated null value pixel. The latter was repeatedly applied until no null values were found, except for those corresponding to water bodies.

The digital elevation model used was the most accurate and up to date for Amsterdam at the moment of study. The 1 km^2 averaging we use for computing distance to outflow point (see details below) and average catchment area, was set to capture the general trend within each cell. We assumed that typical changes in urban topography, since the LIDAR data gathering to compose the DEM, until the date of the rainfall event, were not big enough to affect the topographic trend of a whole km^2.

The filtered DEM was used to delineate subwatersheds and overland flowpaths, computing a flowpaths adjacency matrix. Full downstream overland flowpaths connectivity was modelled in a watershed matrix (W). It was computed using the following expression:

$$W = (\mathbb{I} - A)^{-1} \tag{3.1}$$

where A is the adjacency matrix, and $\mathbb{I}$ is the identity matrix of A (see Section 2.2.3, Jenson and Domingue (1988), and Tarboton et al. (1991) for more details).

The watershed matrix was used to compute average distances to watershed outflow point per site. The proximity measurement for a single subwatershed was set to be the mean network distance from itself to the

outflow point. This was calculated as the distance from the midpoint of a subwatershed flowpath arc to the outflow point. The average distance to watershed outflow point was calculated for every site. These mean values were weighted by the proportional contribution of each flowpath arc to the total flowpath length within the site. Flow path arcs were considered to belong to a site when their centroid was in that site. Flowpath arc centroids are located in the middle point of the mainstream in each subwatershed. For this reason they were chosen as indicators of stream locations, instead of the watershed centroids. The following form expresses the computation of $\bar{d}$, the average distance to watershed outflow point in each cell:

$$\bar{d} = \sum_{i}^{n} \frac{(a_i + b_i)}{2} w_i \tag{3.2}$$

where $\sum_i^n$ is a sum over the n subwatersheds in a given cell, a_i and b_i are the distances from the beginning and end of the flowpath arc of the $i–th$ subwatershed within the cell to its watershed outflow point, and w_i is the weighting factor for the mean length of a_i and b_i.

a_i, b_i, and w_i can be calculated as shown in equations 3.3, 3.4, and 3.5, respectively:

$$a_i = l_i + b_i \tag{3.3}$$

where l_i is the length of the flowpath arc of the $i–th$ subwatershed. The distance from the end of a subwatershed to its outflow point, b_i, equals the sum of lengths of the flowpath arcs of all downstream subwatersheds:

$$b_i = \sum_{j}^{m} l_j \tag{3.4}$$

where $\sum_j^m l_j$ is the sum of lengths of the m flowpath arcs downstream the $i–th$ subwatershed until its watershed outflow point. The weighting factor, w_i, is the contribution of l_i to the total length of flowpath arcs in a cell:

$$w_i = \frac{l_i}{L} \tag{3.5}$$

with L being the sum of lengths of all flowpath arcs whose centroid is within the site.

Here we can use example of Figure 3.1 to better visualize these compu-

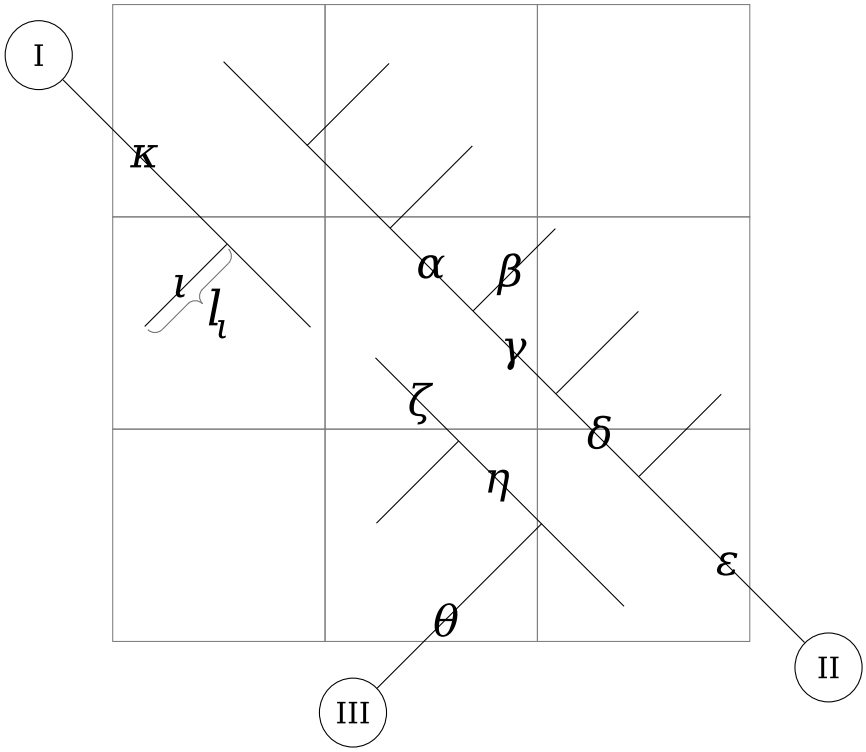


Figure 3.1: Simplified example of watershed flowpaths under an imposed sampling grid. These flowpaths can be extracted from a DEM as described in Jenson and Domingue (1988) and Tarboton et al. (1991), and follow the definitions of Section 2.2.3. Three watersheds are visualized. Their outflow points are represented by circles marked with Roman numbers. A flowpath arc in watershed I is labeled with Greek letter ι. Following the notation used in the main text, the length of ι is denoted by l_ι. In this example, α, β, γ, δ, ζ, η, and ι are 1 length unit long; ϵ, θ, and κ are 2 units long.

tations. The mean distance to outlet for ι, the flowpath arc in Figure 3.1, can be obtained by averaging the distances from the beginning of ι, a_ι (see Equation 3.3), and from its end, a_ι (see Equation 3.4); this is $\frac{3+2}{2}$.

Let us now compute the weighted average distance to outflow, $\bar{d}$, for the central grid-cell in the example of Figure 3.1. Flowpath arcs with centroid falling within this central cell are α, β, γ, and ζ. Their mean distances are $\frac{5+4}{2}$, $\frac{5+4}{2}$, $\frac{4+3}{2}$, and $\frac{4+3}{2}$, respectively. The sum of flowpath lengths, L, in this cell is 4. As all flowpath arcs in this cell are 1 unit long, the weighting factor, in this case, is the same for all of them: $\frac{1}{4}$ (see Equation 3.5). $\bar{d}$ equals $\frac{9}{2}\frac{1}{4} + \frac{9}{2}\frac{1}{4} + \frac{7}{2}\frac{1}{4} + \frac{7}{2}\frac{1}{4}$ (see Equation 3.2), which is 4. This means that in average, objects in this cell are 4 length units apart from outflow point. Also note that even though ζ belongs to watershed III, it is computed together with α, β, and γ, which outflow to II.

Computing the distance to outflow point for every flowpath arc in more complex cases, such as in urban overland flowpath networks, can be done using basic matrix algebra and a watershed matrix (see Equation 3.1). The distance downstream to the watershed outflow point ($\sum_j^m l_j$) was computed in this chapter as follows:

$$\sum_j^m l_j = (F_{1,k} \circ (W_i \circ (1_{k,k} - \mathbb{I}_W)_i)) \bullet 1_{k,1} \tag{3.6}$$

where $F_{1,k}$ is the vector of lengths of the k flowpath arcs of all subwatersheds, W_i is the watershed matrix row of the element i (see Equation 3.1), $1_{k,k}$ is an all-ones square matrix of k, k dimensions, $\mathbb{I}_W$ is the identity matrix of W, and $1_{k,1}$ is an all-ones column vector of length k. $(1_{k,k} - \mathbb{I}_W)_i$ is a row vector of length k with a zero value in the position of the element i and ones elsewhere. It is used to exclude i's arc length from the sum, as we were only interested on downstream arc lengths. $\circ$ and $\bullet$ are element-wise product (or Hadamard product) and scalar product, respectively. It is useful to note here that the order of elements in F and rows and columns in W must be consistent.

Thus, $\bar{d}$ from Equation 3.2 can be expressed as follows:

$$\bar{d} = L^{-1} \sum_i^n l_i \left(\frac{l_i}{2} + ((F_{1,k} \circ (W_i \circ (1_{k,k} - \mathbb{I}_W)_i)) \bullet 1_{k,1})\right) \tag{3.7}$$

This computation avoids $\bar{d}$ being reduced due to extremely frequent

short-length arcs. Weighting in terms of the total length of flowpath arc within each site normalizes the contribution of each segment to the average distance to outlet in each particular cell. Given that different cells have different total flowpath lengths, this approach does not account for the differences in total drained areas, which is better captured by the measurement of the percentage of impervious areas within each site.

Catchment area

In case of an urban drainage failure, rainwater flows along overland flowpaths until it reaches a water body or a permeable area (Maksimović et al., 2009). In such cases, larger catchment areas have a higher potential overland flow and are expected to be more susceptible to flooding.

The catchment area at given point can be also seen as a measure of how low that point is along the watershed in terms of the area it drains from upstream locations. A weighted average of catchment areas, $\bar{A}$ was computed for each site as follows:

$$\bar{A} = \sum_{i}^{n} C_i g_i \tag{3.8}$$

where C_i is the catchment area at the centroid of the i–th subwatershed flowpath in a given site. This selection aimed to capture the average catchment area of each subwatershed, given that it is different in every subcatchment pixel. Flow paths were considered to belong to a site if their arc centroid was within that site. g_i is the weighting factor of a subwatershed i in a given site, which is the ratio of its individual enclosing area (a_i) to the sum of subcatchment areas in that site (A). Thus, $\bar{A}$ in equation 3.8 can be expressed as:

$$\bar{A} = A^{-1} \sum_{i}^{n} C_i a_i \tag{3.9}$$

The applied weighting enforces subwatersheds with smaller enclosing areas to contribute less to the average catchment values in a site. Terminology and procedure used to compute catchment areas follow Section 2.2.

Standardization of variables

The variables handled in this study have different measuring scales; they refer to different physical and socioeconomic dimensions. They were standardized to make them comparable in the multivariate analyses. Standardized values of those variables were obtained by using the following expression:

$$y_{std} = \frac{y - \bar{Y}}{S_Y}; \forall y \in Y \tag{3.10}$$

where y_{std} is the standardized value of y, which is one value from the set of observations (Y). $\bar{Y}$ and S_Y are the mean and standard deviation of Y.

Data transformation

In order to check the normality of variables, and the suitability of fitting a linear model on them, a Shapiro-Wilk test was performed. Non-normal variables were transformed and used in a multiple linear regression, which is detailed in Section 3.2.2. Box-Cox transformations were applied as follows:

$$T(y_i) = \begin{cases} (y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \ln(y_i) & \text{if } \lambda = 0 \end{cases} \tag{3.11}$$

where y_i are each of the individual values of the response variable, and λ is an arbitrary parameter whose value is adjusted to provide maximum correlation between the distribution of transformed response values and standard normal distribution (Box and Cox, 1964). Obtained lambda was rounded to one decimal and used for the transformation.

3.2.2 Multivariate analysis techniques

Multivariate analysis techniques are described in following subsections. As rainfall intensity and spatial distribution might change from event to event, we run tests including and excluding rainfall variables. This allowed us to evaluate if incidents significantly responded only to urban fabric conditions without considering rainfall.

The multivariate techniques described below were used to assess variable relationships from different angles. The overview obtained with the correlogram of Figure 3.4 provided a description of links between variables that was limited to pair-wise relationships, and that ignored possible local arrangements of descriptor values and reports. Such arrangements were

visualized using cluster analysis. Obtained clusters were used to describe the spatial distribution of descriptors over the urban landscape, and the local occurrence of incident reports. Principal component analysis (PCA), and multiple linear regression (MLR) gave insights into relationships between variables from another perspective. PCA was used to evaluate the contribution of each variable to the overall variability across all variables, to identify collinearity between variables, and to select the variables to be fed to the MLR. Finally, the latter provided a quantification of the significance and performance of a combination of descriptors in explaining the occurrence of incident reports.

Cluster analysis: grouping sites according to available independent variables data

Following Gaitan and ten Veldhuis (2015), a square similarity matrix was obtained for the sites. Given the quantitative nature of the variables, Euclidean distances between variable realizations in studied sites was used as a measure of similarity. The aim was to identify areas in the city with similar characteristics in terms of variables under study, and to explore whether occurrence of flooding differs among areas with particular environmental configurations (Braak and Looman, 1995; Legendre and Legendre, 2012a).

All descriptor variables were considered in cluster analysis to classify urban regions by recognizing patterns in all available information about independent variables. This was done regardless the strength of descriptor-response relationships.

Euclidean distances between sites, in which cluster analysis were based, were calculated by means of the equation:

$$D_{(x_1,x_2)} = \sqrt{\sum_{j=1}^{p}(y_{1j} - y_{2j})^2} \tag{3.12}$$

where j indicates the j–th descriptor, and y_{1j} and y_{2j} are the two sites for which the pairwise distance is calculated.

Given the lack of previous research about the importance of different variables in explaining the occurrence of flooding incidents, an unweighted pair group method with arithmetic average (UPGMA) was carried out using all the variables except for the response (incident reports). In this technique, pairs of objects, or groups, are successively classified together

when their distance, or the distance between group averages, is the smallest. Two cluster analyses were performed: one including rainfall variables, and another excluding them.

Profiles of obtained clusters were further analyzed and compared with the occurrence of incidents per cluster. For building these profiles, independent variables were standardized to mean 0 and unit variance. In this way, their distributions could be compared using a standardized scale. The distributions of standardized values of different variables, in every cluster, were then visualized using box-plots. Incident reports were not included in the computation of clusters to depict classes of the urban landscape only in terms of independent variables. As incidents were assumed to be dependent on descriptor variables, excluding them from clustering avoided obtaining groups on the basis of their similarity between sites. Once the classification of sites was completed, the incidence of reports was studied in each of them.

Principal components analysis: inspecting dimensionality reductions

Two PCAs were performed. The first one was applied to a correlation matrix of all variables, including the response variable, to discover relationship patterns among them. In the second one, rainfall variables were excluded. This aimed at evaluating the reconfiguration of PCs when rainfall is not taken into account, which implies a focus on the urban fabric.

PCA extracts a theoretical axis, or Principal Component (PC), from the data, which explains, in a linear fashion, most of the variability in observed variables. Additional, uncorrelated PCs can be computed, which account for remaining variance, not explained by the previous PC. The goal of PCA is to display most of the multidimensional data variance in a few PCs. Note that explanatory variables in PCA are theoretical and mutually orthogonal. This makes PCA different from linear regression (ter Braak, 1995; Legendre and Legendre, 2012b).

PCA was used to visualize correlations among descriptors, and contributions of each of them to the PCs. This allowed us to recognize data redundancy and select variables highly correlated with the response variable.

Multiple regression: modelling linear relationships on multiple descriptors

MLR estimates the parameters of a linear function, which model follows the form $E_y = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p$, where the expected value of y depends on an intercept and a number of p terms, consisting on an estimated parameter b_p and descriptor variable x_p. A least squares estimate of the vector of parameters ($\hat{\beta}$) is $(X'y)/(X'X)$, where X is a matrix of values of sites vs descriptors (i.e. the data matrix) and y is a column vector of response values in the n sites (Montgomery and Runger, 2003).

MLR was applied to a selection of variables based on the PCA results. Variables were standardized to properly compare regression coefficients and determine the importance of analyzed terms. Additionally, a MLR was applied to the transformed data.

3.3 Results and discussion

Figure 3.2 presents parsed rainfall radar intensity series. Values were aggregated at 15 min and averaged over the cells on Amsterdam administrative borders. The gray bars indicate a 15 min aggregated count of incident reports. Five rainfall intensity peaks can be differentiated in this figure. The first two peaks had an average between 3 and 6 mm/h, with deviations in the scale of 10 mm/h. Third peak, at 11:00 h of Monday July 28 2014, was the highest, with maximum average intensity of 20 mm/h and a deviation around 20 mm/h. At this peak, 5% of the Amsterdam area was impacted by rainfall intensities over 75 mm/h (see quantile 95% in Figure 3.2). While the initial two peaks did not triggered a single incident report, the period associated with the third peak entailed a wave of them; within two hours around this peak, incident reports crested to more than 30 every 15 min. This rainfall intensity and reports peaks decreased around 15:00 h. A small fourth rainfall intensity peak, with an intensity of less than 1 mm/h, took place around 02:00 h of Tuesday July 29; no incidents were reported then. Between 07:00 h and 08:00 h of the same morning, the fifth peak registered an average intensity of 1 mm/h, followed by six hours in which 1 to 2 incidents were reported every 15 min.

Figure 3.2 provides an ambiguous picture of the possible relationship between rainfall intensity and reports incidence. The coincidence of the highest peaks of average rainfall intensity and reports may suggest that they

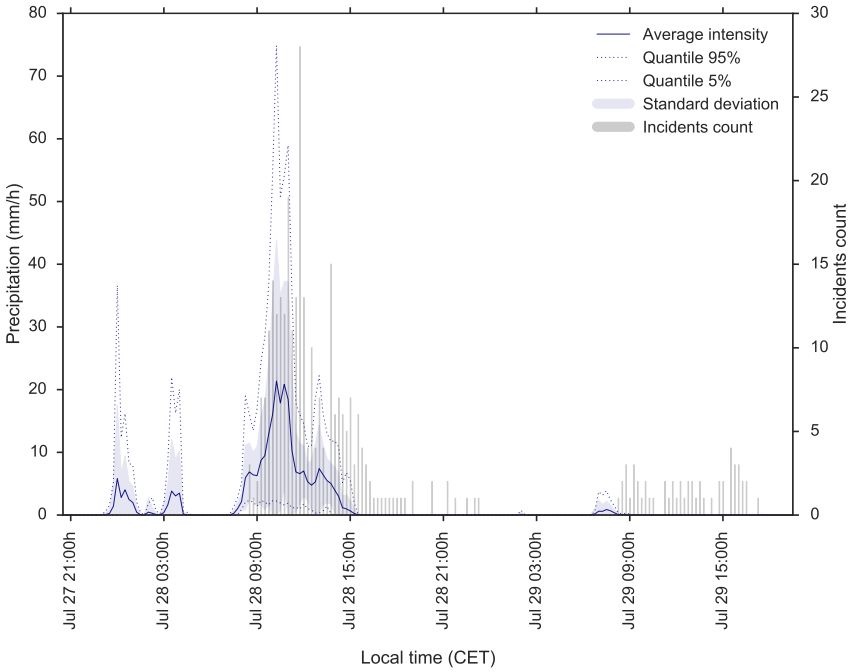


Figure 3.2: Average rainfall intensity during the studied event. Values of the rainfall radar grid cells where aggregated at 15 min time-steps and averaged to obtain the visualized plot. Bars indicate incidents count, also aggregated at 15 min time-steps.

are associated. One could even speculate that the absence of reports during the two first peaks could be explained by a lower rainfall intensity than in the third peak; nevertheless, this is opposed by the numerous reports following the fifth rainfall peak, which was five times less intense than the first two peaks. This suggests that variables different than rainfall intensity might influence the incidence of reports.

Cells of the radar-grid where incidents were registered were considered as sampling sites. The latter accounted for a total of 80 cells, comprising a total of 336 incident reports. Such sites are numbered in Figure 3.3, which also presents a view of the distribution of 15-min stepped maximum rainfall intensities and the location of geocoded incident reports. A direct relationship between rainfall intensity and reports incidence cannot be perceived in this visualization: the number of reports in sites located north of the river Ij, where maximum rainfall intensities were higher, seems to be less than in central Amsterdam.

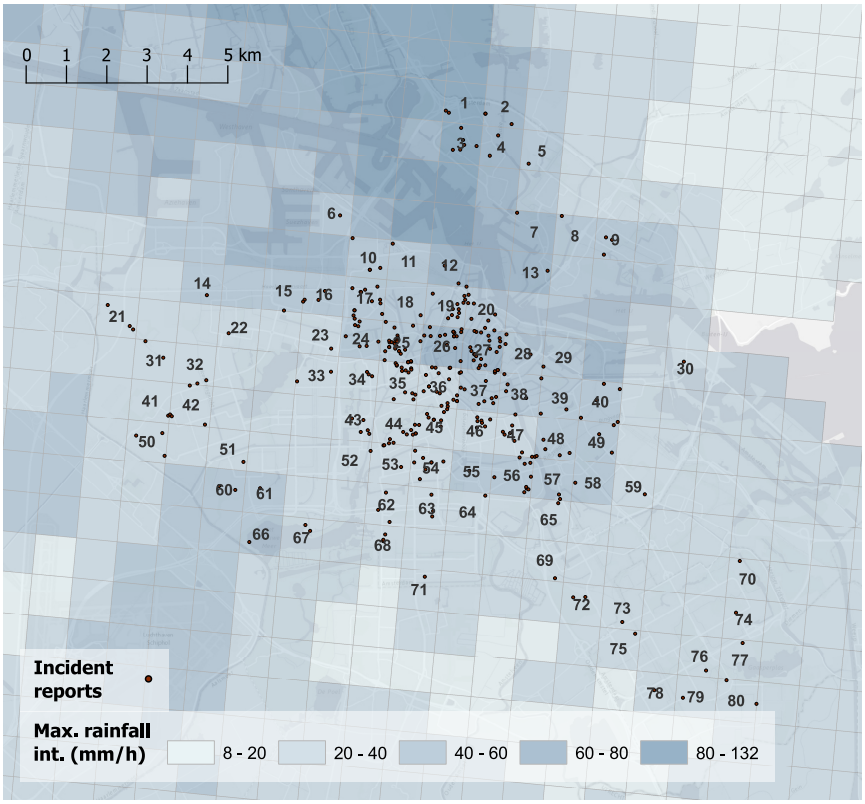


Figure 3.3: View of maximum rainfall intensities during the event (15 min time steps) and reports of incidents. The 80 sampling sites are numbered with an arbitrary index.

An initial overview of distributions and associations of studied data sources can be found in in Figure 3.4. It includes histograms and pairwise correlograms of all preprocessed data; Pearson correlation values and the p-value of a test for association between paired samples using Pearson's coefficient are shown under the diagonal.

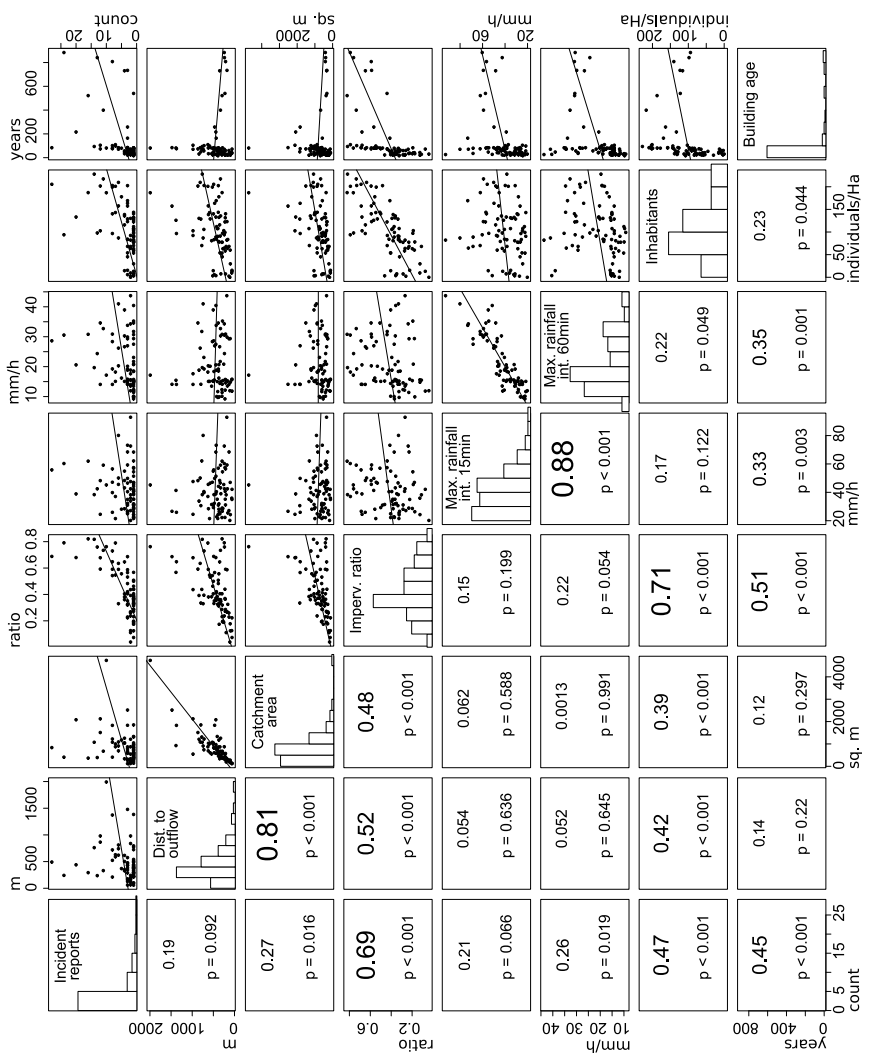


Figure 3.4: Scatter plots of pairwise combinations of studied variables. Pearson's correlation and the significance of pairwise associations are shown under the diagonal. Scatterplot units are shown along the figure borders. Distribution of variables is plotted in the diagonal histograms. Units in the bottom and top borders can be also used to determine value intervals of histogram bins. Bin heights are relative to value frequencies in each dataset, and do not relate to units in left and right borders.

The associations between reports and rainfall intensity found in the correlogram of Figure 3.4 were weak but significant. Correlations were 0.21 in the case of 15 min rainfall intensity and 0.26 in for 60 min correlations. Pair-wise correlation significance were 1% and 2%, respectively. These results suggested that rainfall intensity could have played a small but meaningful role in the distribution of rainfall incidents. This possible relationship was further assessed in the cluster, principal component, and multiple regression analyses.

3.3.1 Normality tests and data transformations

Results of normality tests showed that from all variables only inhabitants and imperviousness can be considered approximately normal. The Shapiro-Wilk's null hypothesis, which assumes the distribution of sampled data to be normal, was rejected at 99% confidence for all variables except for imperviousness and inhabitants. Obtained *p–values* were as follows: incident reports, 1.78×10^{-12}; distance to outflow, 1.94×10^{-8}; catchment area, 1.12×10^{-10}; impervious ratio, 5.10×10^{-2}; 15 min stepped max. rainfall intensity, 6.86×10^{-4}; 60 min stepped max. rainfall intensity, 1.66×10^{-5}; inhabitants, 3.69×10^{-2}; and building age, 5.15×10^{-15} .

As linear regression assumes that studied variables are approximately normal, non-normal variables selected for MLR in Section 3.3.3 were transformed: incident reports, maximum rainfall intensity within 15 min windows, building age, and distance to outflow. Respective obtained transformation lambdas were: -0.5, -0.2, -0.5, and 0.1. For transformation of 15 min rainfall intensity and distance to outflow, the lambda value was approximated to 0; this value fell within a 95% confidence interval centered in the lambda with maximum likelihood of yielding a distribution closer to a normal one. This approximation eased their interpretation as this lambda value represents a simple logarithmic transformation in the Box-Cox procedure (see Equation 3.11).

3.3.2 Cluster analysis

Cluster analysis including rainfall variables

Results of cluster analysis for all descriptors are shown in Figure 3.5. Trimming of the dendrogram was set to obtain 6 clusters, avoiding groups to be made of a single or too few sites. Other trimming levels were tested,

resulting in several very small groups. Obtaining one group with a single site was unavoidable because of the large dissimilarity of site 56 (see Figure 3.5). A similar cluster analysis was performed excluding rainfall variables. Box-plots in the top half of Figure 3.7 present profiles of standardized variables within the six groups obtained from the cluster analysis. Note that, even though variables were standardized to be used in the profile visualization, the y-axis in the box-plots is not centered in 0. This occurs because each cluster comprises a different subset of variable realizations, whose distributions differ from the standardized, overall dataset distribution. Profiles of groups that accounted for higher numbers of reports are analyzed below. As a reference, the mean density of reports per site is 4.2 (which is the total number of studied reports over the number of sites: 336/80).

Group 3 contained the greatest number of reports (42%). Its sites are located mostly surrounding the city center towards southeast and southwest (see Figure 3.6). This group had a report density close to 5, which does not differ much from the expected density of 4 reports/km^2. This means that the high amount of reports can be explained by the size of this group. When compared to groups 4 and 2, the second and third groups with the greatest report count, group 3 can be differentiated for having relatively recent building ages. Distance to outflow and catchment areas tended to be lower than in group 2, and clearly higher than in group 4. Average imperviousness were comparable in groups 2, 3, and 4, though this variables was less dispersed in group 4.

Group 4 accounted for the second greatest amount of reports (23%). Comprising 8 km^2 (10% of studied sites), reports density of this group was very high: close to 10 reports per km^2. As it can be seen in Figure 3.6, group 4 is located in the city center. This group was characterized by small catchment areas and short distances to watershed outflow point. This group also comprised the oldest buildings in the study area. This is clearly the outcome of urban layout of Amsterdam's old city center, where the dense network of canals limits the length and size of distances to outflow and catchment areas.

Group 2 comprised 33 sites (40% of total sites), making it the largest in covered area. It accounted for 15% of incident reports, and a report density close to 2; which is half of the mean reports density. This group extents over the periphery of the studied area, mainly towards Amsterdam

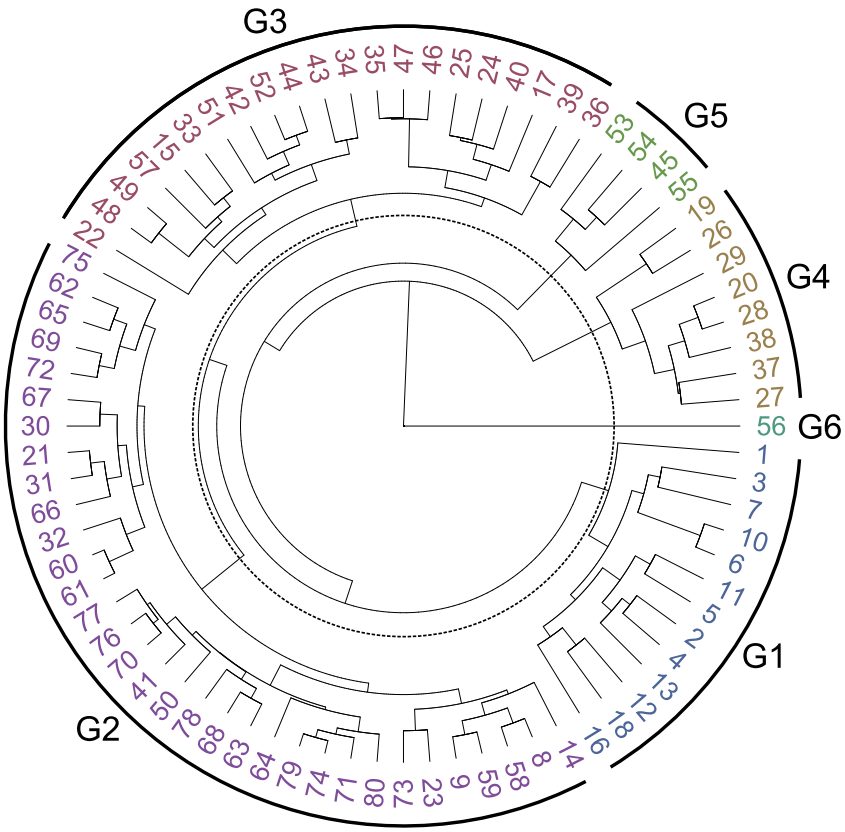


Figure 3.5: Dendrogram of hierarchical cluster analysis with UPGMA. Colors of site numbers correspond to the different clusters. The dashed line indicates a similarity threshold of about 62%, at which different clusters are differentiated.

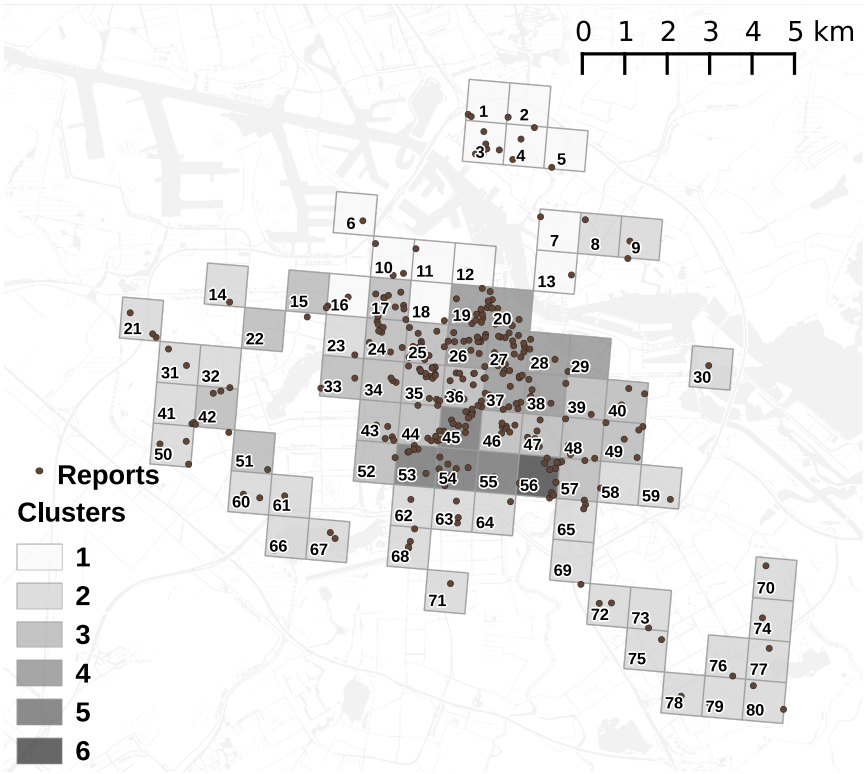


Figure 3.6: Sites classified according to cluster analysis (see Figure 3.5).

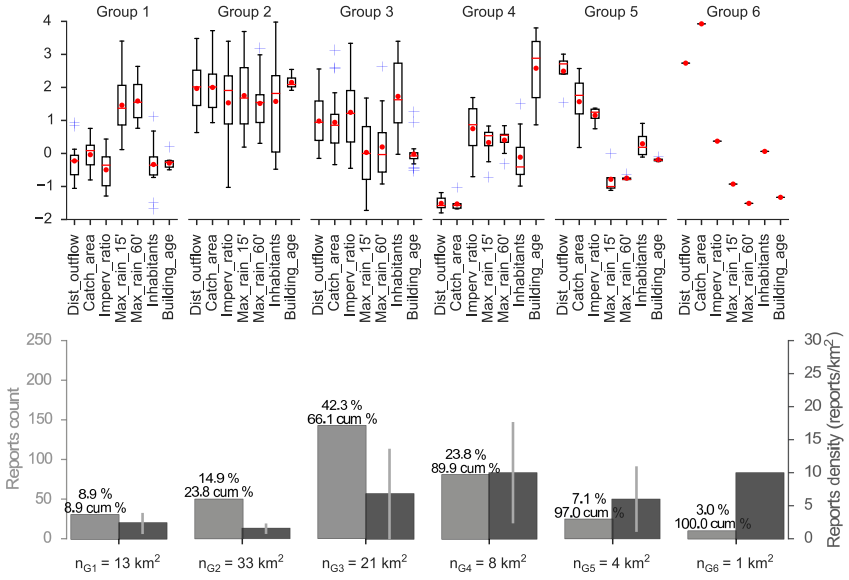


Figure 3.7: Profiles of classified groups. Top half of the figure includes box-plots of standardized variables ($\bar{x} = 0$ and $s^2 = 1$) among classified groups. Bottom half shows the count of incident reports in each group (lighter bars), as well as the average report count per site (darker bars), which can also be considered as 'reports density'. Individual percentage and cumulative percentage sums (cum) of reports per group, are indicated on top of reports count bars. Standard deviation of report density in each group is indicated by error bars (thinest, lighter bars over density bars). The area (number of cells) covered by each group is shown in the bottom.

South-West, South, and South-East, as well as over two square kilometers in Amsterdam North (see Figure 3.6). It is characterized by old buildings, with a narrow age distribution. This may suggest that this group corresponds to an urban expansion development occurred relatively simultaneously in its sites. Despite the low report density in this group, its distance to outflow, catchment area, and population tended to be higher than in groups 3 and 4. In fact, these two variables in group 2 tended to be the largest among all groups. Imperviousness presented a wide dispersion, with an slightly longer lower tail (see box-plot in Figure 3.7).

Group 1 displayed a similar reports density, within half the area covered by group 2. Rainfall intensity in group 1 also showed similar trends as in group 2. The rest of the variables, including distance to outlet, catchment area, and distance to outlet, presented relatively low values. Imperviousness, presented some of the smallest values Sites of group 1 were situated North of the Ij river, and Northwest of the city center. Groups 1 and 2 covered 60% of studied area, accounting only for 25% of reports.

Reports in groups 5 and 6 had the lowest count, with 7 and 3% respectively. Comprising 4 km^2, group 5 had a reports density around 5, which is close to the mean value of reports density. This group was characterized by relatively large distances to outflow and catchment areas, and slightly high impervious ratio, low rainfall intensities, close to average population densities, and average building ages. Group 6 represented a single site with relatively large catchment areas and distances to outflow, extremely low rainfall intensities, and very recent buildings. Reports density in this site was as high as that observed in group 4.

Cluster analysis excluding rainfall variables

Cluster analysis excluding rainfall variables delivered the following changes to previous clusters: 8 of the 13 sites of original group 1 were allocated in group 2, the remaining 5 sites of group 1 were allocated in group 3, and a new group (referred to as $1_{\text{no rain}}$), consisting of 2 sites originally in group 4, emerged. Groups 5 and 6 stayed the same as in the above described classification; the distribution of variables within them did not change. This classification produced groups that resembled location of Amsterdamer boroughs.

Changes in variable distributions of group 2 and 3 were similar: averages of all variables moved closer to 0 as a result of the contribution of small

values of variables of sites originally in group 1. Overall differences between both groups remained the same, with group 2 having higher distances to outflow, catchment areas, and imperviousness, and older buildings. In the case of group 3, there was an increment of 6.0% in the area covered and a 4.4% in reports incidents. As for group 2, the increment was 10.0% and just 4.4%, respectively. This made new group 2 an interesting class, as it covered more than half of the study area while accounting for only a fifth of incident reports. The rearrangement of groups 2 and 3, due to the exclusion of rainfall from cluster analysis, approximately set group 2 around the city center, and group 3 around group 2.

Group $1_{\text{no rain}}$ consisted of sites 19 and 26. It was located over the neighbor of Jordaan, in the city center. Covering only 2.5% of studied area, this group comprised 8% of reports. Particularly, the values of variables defining this group had a very narrow dispersion: it was a cluster clearly defined by low distance to outflow and catchment areas, high imperviousness and population density, and relatively old buildings. This seems to be highly vulnerable area to pluvial flooding.

In comparison with group Group $1_{\text{no rain}}$, average values of variables in Group 4 where clearly lower for distances to outlet and catchment areas, much lower for imperviousness and population density, and higher building age. These results, from group $1_{\text{no rain}}$ and 4, agreed with those found by PCA and MLR, discussed below. The exclusion of possible information noise related to rainfall, the narrow distribution of variables, and the relatively high report count in these two small cases provided a picture of how independent variables can be linked to the incidence of reports.

In general, profiles of existing groups in previous classification did not exhibit remarkable difference with the just discussed profiles. By excluding rainfall from the elaboration of clusters, obtained classes apparently reflected the underlying city landscape better. The distribution of rainfall intensity is highly dynamic and random, and interfered with the classification of the urban landscape. The rearrangement of groups 2 and 3, and the emergence group $1_{\text{no rain}}$, changed the geographic layout of the initial classification. This new layout seemed to follow the historical concentric development of Amsterdam. There is a roughly coincidence of groups 4 and $1_{\text{no rain}}$ with Amsterdam-Centrum, 5 and 6 with Amsterdam-Zuid, group 3 with Amsterdam-West and Amsterdam-Oost, and group 2 with more peripheric

areas.

Average distance to outflow, catchment area, and building age are clearly differentiable among biggest groups. In this sense, the use of cluster analysis depicted different classes composing the urban landscape in terms of the studied descriptor variables. Overall, in spite of these differences, the variability in the count of incident reports, particularly in the largest groups (see gray bars in the bottom of Figure 3.7), limits its use to draw conclusions about a possible increased reports incidence given clusters.

3.3.3 Principal component analysis

Results of PCA are shown in Figure 3.8. The six groups obtained via cluster analysis (Figure 3.5) are also plotted enclosed in polygons.

Results of PCA showed that variance in data can be mostly explained by the first two or three eigen vectors. Average catchment size, and average distance to outlet are highly correlated, as well as maximum rainfall intensities within 15 and 60 minutes. This is also the case for population density and imperviousness ratio, reflecting that areas more densely built tend to host more people. The sum of incidents tends to be orthogonal to the rainfall intensity as well as distance to outflow and catchment area. Building age vector is slightly closer to the response variable. Imperviousness ratio and population density vectors are most closely related to the occurrence of incidents.

Figure 3.9.a shows PC1 and PC2 being relatively higher than average explained variance. PC3 does not differentiate from this average. PC1 and PC2 explain 40% and 27% of variance each. PC3 accounts for 15% of variability, which is slightly above the expected average variability accounted for by one in eight variables. These three principal components explain 82% of variability. None remaining PCs is close to the average eigenvalue; those components do not represent important gradients in the data dispersion.

In Figure 3.9.b, imperviousness and inhabitants are the descriptors with contributions greater than the average score. Reports also score more than average in PC1, suggesting that their distribution tends to follow a theoretical environmental gradient aligned with imperviousness and population density.

Figure 3.9.c indicates that the two catchment-related variables, the two rainfall variables, and building age scored higher than average for PC2.

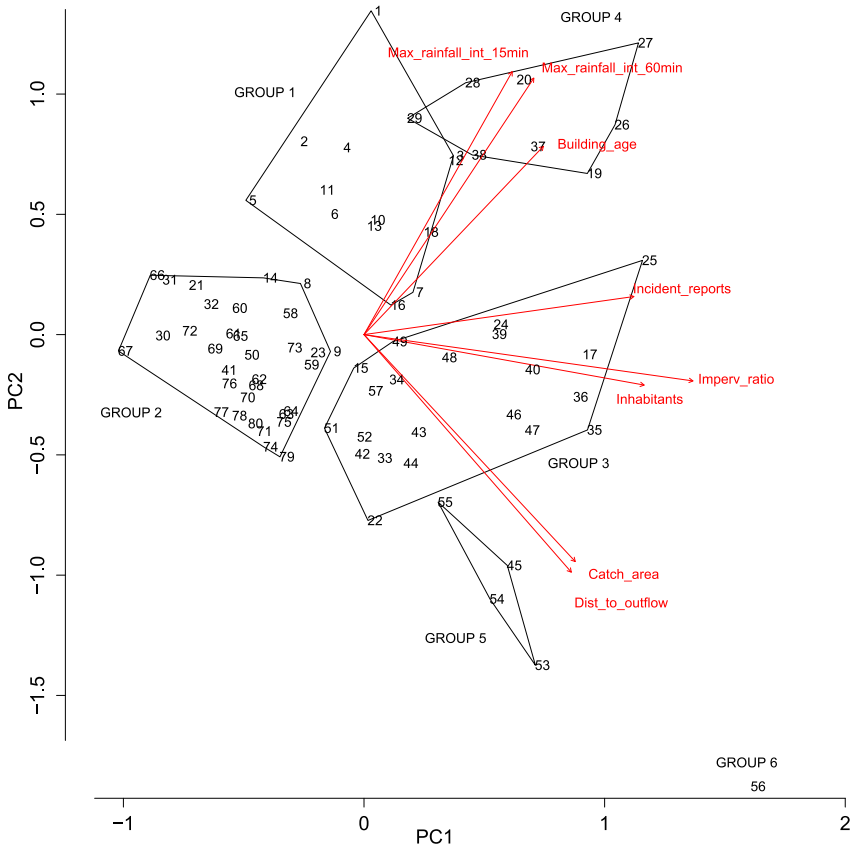


Figure 3.8: Biplot of scaled principal component analysis of sites including rainfall variables. Polygons enclosing sites represent groups obtained from the cluster analysis (see Figure 3.5).

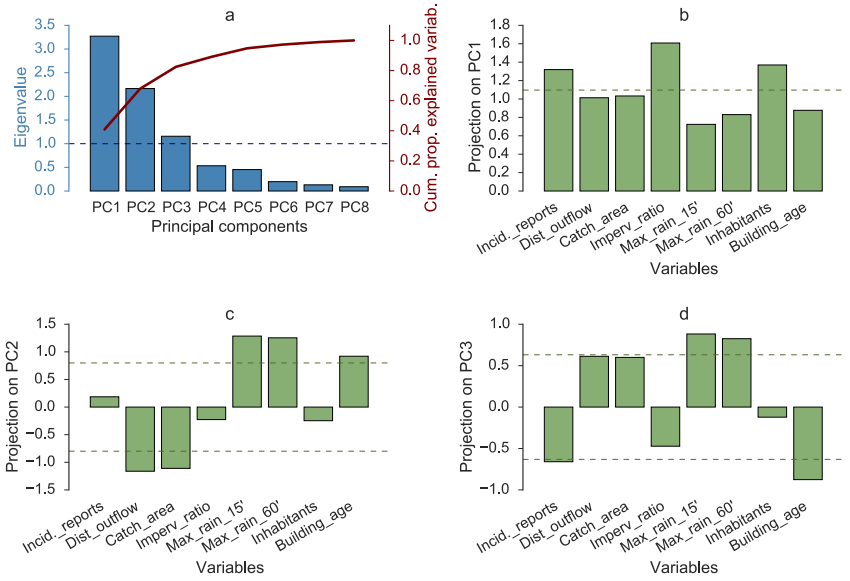


Figure 3.9: a. Eigenvalues (blue) and cumulative proportion of explained variability (maroon) of PCA applied to all variables. The dashed blue line indicates the average eigen value. The projection, or contribution, of variables to each of the first three principal components is shown in b., c., and d., respectively. Green dashed lines indicate the average contribution of the variables to the eigenvectors.

Distance to outflow scored slightly higher than catchment area (-1.16 and -1,11, respectively), as it did 15 min windowed rainfall over the 60 min windowed one (1.29 and 1.25, respectively). The sign of the scores indicate the direction of the variable vectors in the respective PC (e.g., see vector of catchment area in Figure 3.8).

Figure 3.9.d shows that rainfall variables and building age scored high in PC3. In contrast to PC2, in PC3 incident reports have a contribution at least as big as the expected average and, as well as building age, tend to be opposite to the rainfall intensity.

Results from PCA applied to the dataset excluding rainfall variables resemble the outcome of the PCA applied to all variables.

The layout of clusters in Figure 3.8 resembled the patterns found in Figure 3.7. Group 2, whose profile showed to have a low report density, also presented this pattern here. The high variability of reports values for group 3 is visualized here again; group 3 was dispersed along PC1, the component on which the response variable contributes the most. It was also evident that group 3 displayed some of the highest impervious ratios and population densities. Response variable does not varies in groups 4 and 5 as in group 3: the range of the two earlier over PC1 is shorter. As well as in Figure 3.7, group 5 was represented in Figure 3.8 by the highest average distances to outflow and catchment areas, and recent building ages. Group 6 presented the highest average catchment area value, and a very high count of reports, which explained its location in the lower-right corner in Figure 3.8. The latter clearly shows that the six different groups are easily differentiable in terms of two theoretical variables, which relatively coincide with some of the studied descriptors. Groups 1 and 4, were however overlapping, which can probably be resolved by checking both groups distribution over PC3.

Dimensionality reduction

Results discussed above aided the selection of descriptors used for MLR by avoiding variable collinearity. Correlation between imperviousness and population was significant and strong (see Figure 3.4); these variables also showed close collinearity (see Figure 3.9). Imperviousness scored higher than inhabitants in all PCs (see Figure 3.9). Nevertheless, both variables were selected for MLR to clearly check whether areas with higher population density report flooding incidents more frequently.

From Figures 3.4 and 3.8 it was clear that catchment area and distance

Table 3.2: Explaining power, estimated parameters and respective significance obtained for the regression applied to all descriptors selected in Section 3.3.3. The second half of the table shows results for transformed variables, as described in Section 3.3.1. Significances with $P < 0.5$ are indicated in bold typeface.

Results of MLR applied on non-transformed variables: Adjusted R^2 =0.50		
Parameter	Coefficient	Significance (*p–value*)
Intercept	7.515×10^{-17}	1.00
Impervious ratio	0.86	$\mathbf{1.6 \times 10^{-6}}$
Distance to outflow	-0.24	**0.03**
15 min rainfall int	0.09	0.29
Inhabitants	-0.04	0.71
Building age	-0.04	0.71

Results of MLR applied on transformed variables: Adjusted R^2 =0.48		
Parameter	Coefficient	Significance (*p–value*)
Intercept	-0.29	1.00
Impervious ratio	-0.73	$\mathbf{1.07 \times 10^{-5}}$
Distance to outflow	0.24	**0.03**
15 min rainfall int	-0.05	0.62
Inhabitants	-0.08	0.53
Building age	-0.54	0.67

to outflow are closely correlated. The contribution of distance to outflow to PC2 was greater than that of catchment area. For this reason, distance to outflow was selected for MLR.

15 min time-windowed rainfall intensity was selected because it scored slightly higher than rainfall intensity windowed at 60 min both PC2 (1.29 and 1.25, respectively) and PC3 (0.88 and 0.83, respectively), where these variables scored higher than average. Given the contribution of building age to PC2 and PC3, it is also selected for MLR.

3.3.4 Multiple regression analysis

The adjusted coefficient of determination of the model, R^2, was 4.99×10^{-1} . The significance of a F-test, whose null hypothesis is a fewer-parameters-model with better fitting, was 5.14×10^{-11} . This proved that the model was able to explain close to half of the variability of incidents occurrence, and that a model based on fewer variables does not perform better. This

suggested that excluding rainfall intensity was not enhancing the MLR explanatory power.

Values of coefficients and associated *p–values* (see Table 3.2) show that imperviousness was the strongest and most significant descriptor, followed by distance to watershed outflow point. This outcome is in agreement with results of PCA. While in the latter both variables were clearly collinear (Figure 3.8), inhabitants seconded imperviousness in the contribution to PC1, indicating that imperviousness carried most of information in that component.

This was confirmed by results of an additional MLR run, in which imperviousness was excluded. It resulted not only in inhabitants being significant with a *p–value* of 2.23×10^{-3}, but also population density with a *p–value* of 4.36×10^{-4}, with coefficients 3.43×10^{-1} and 3.77×10^{-1}, respectively. This result actually accords with findings of Chapter 2, which observed indications of an association between incident reports with building age and population density in a dataset that did not include descriptors about imperviousness nor topography. Nevertheless, adjusted R^2 accomplished by this MLR was 3.23×10^{-1}; this is close to 17% less explaining power than the achieved by the MLR including imperviousness.

Links between incidents and remaining descriptors were insignificant, indicating that additional parameters should be explored to better explain flooding incidents. Distance to outflow's negative coefficient is in agreement with PCA results, but contradicted the initial working hypothesis where areas closer to water bodies were expected to be less susceptible to flooding incidents (see Section 3.2.1). Rainfall intensity shows limited significance in explaining flooding incidents occurrence. The strong link found for imperviousness agrees with previous research, where this variable has been identified as a factor heavily influencing the sensitivity of storm water models (e.g. Dotto et al., 2011).

MLR was also applied to the dataset where non-normal variables were transformed using Box-Cox transformation (see Section 3.3.1). Results were similar to previous MLR applied to non-transformed variables (see Table 3.2). Obtained adjusted R^2 was 4.80×10^{-1}, slightly less than for MLR based on non-transformed data.

Imperviousness and distance to outflow point were stronger descriptors of flooding reports occurrence than rainfall intensity and population density.

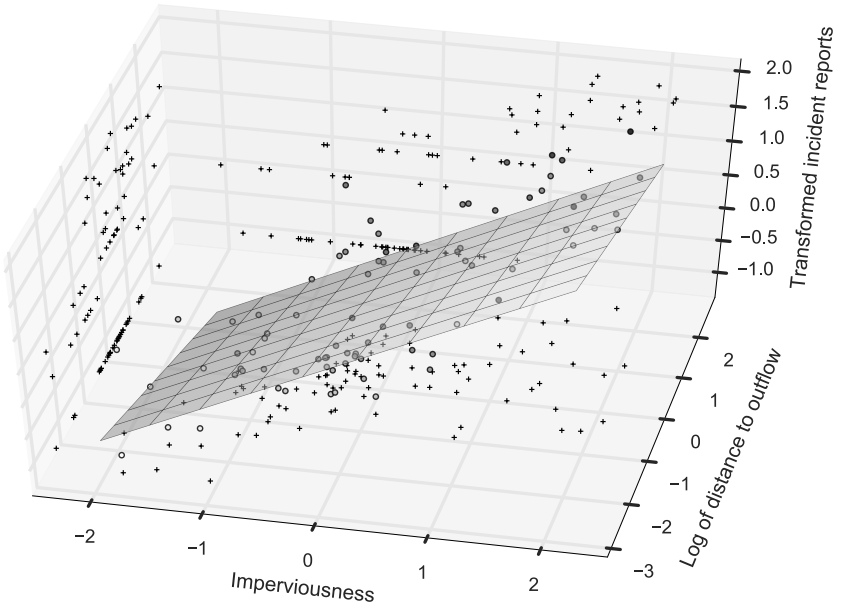


Figure 3.10: 3D representation of response data and the two most significant descriptors. Modeled response surface is also shown. Crosses represent projections of 3D floating points on 2D scatter-plots. The tone of floating points indicate their alignment to the distance to outflow axis: the darker the point, the higher the log of distance to outflow value. Lighter tones of the response surface indicate a higher the response value.

This results opposes the *a priori* expectation that areas exposed to higher rainfall intensity, and with higher population, would deliver higher report counts. Regarding rainfall intensity, Spekkers et al. (2013) found rainfall intensity to explain no more than 34% of building and content damage, including roof leakage cases. It is then reasonable that rainfall intensity does not play a major role in the case of street flooding incidents.

Figure 3.10 shows the response surface for imperviousness and distance to outflow variables, and scattered 3D points for those two and the response variable. The surface was made using the coefficients obtained in the MLR applied to transformed variables. This figure shows that the slope between imperviousness and incident reports is positive and relatively steep, while the one between distance to outflow and reports is negative and moderate. This figure also includes 2D scatter-plot projections for all variables in the graph. Lighter color of response surface indicates higher incident reports

values. Darker color of 3D points indicates higher transformed values of distance to outflow.

In both 2D scatter-plots of descriptors against response, there is a series of low response values that is not sensitive to changes in descriptor values. Figure 3.10 shows that the response surface is drawn downward due to those low response values. If those low response values were excluded from analysis by setting a minimum threshold (e.g. $y_{selected} \geq 2$), correlation of distance to outflow and imperviousness against reports would both be positive.

Low incident report values, representing about half of the response variable dataset, could be influenced by variables different from those linked to the higher response values. Analyzing these two types of response values separately could reveal different underlying flooding processes. Decision tree analysis is an adequate tool to assess response variables in cases like this one.

3.4 Conclusion and outlook

In this chapter multiple open environmental and socioeconomic spatial datasets were explored using multivariate analysis techniques. The aim of this chapter was to assess the degree in which openly available datasets explain the occurrence of flood incident reports by using exploratory data analysis.

The analyses identified that impervious ratio, and distance to watershed outflow point, significantly explained up to half of the variability in the spatial distribution of flooding incidents. This doubles the explaining power achieved by Spekkers et al. (2014), who used decision trees on confidential insurance claims and household-level statistics to explain urban flooding damage to buildings and contents. Decision trees have been also used by Merz et al. (2013) for explaining damage due to river flooding, on the basis of telephone interviews, hydrological, socioeconomic, and building variables, among other data sources. These datasets comprised approximately a thousand samples. The latter study achieved remarkable explaining powers; from 50% to 75% in average. This might suggest that incident reports bear a stronger signal about street flooding incidents than insurance claims do about pluvial flooding damage to buildings and contents. Under this scenario, decision trees should be used to further explore street flooding

incidents.

Rainfall intensity was found to be a weak predictor of the spatial distribution of flooding incidents. Even though incident reports are made by people, population density was not a significant descriptor of reported flooding incidents.

This study devised a novel tool to approximate watershed measurements in grid-cell aggregations; this tool is useful for future research involving data mining, spatial aggregations, and overland flowpath networks. The required mathematical implementation was developed specifically for this chapter. Mining of open spatial data to model complex environments frequently requires data aggregation and indexation. Integrating the concept of watersheds into the data mining process poses the challenge of subsetting them without breaking their network connectivity. Aggregated average distance to outlet and reports incidence presented a significant relationship; this indicates that the earlier bears actual information about processes taking place in the urban environment, and signals its utility for urban environmental modelling.

Results of cluster analysis were hampered by wide descriptor variability and meager response data. Groups identified by cluster analysis agreed with locations of distinct areas of the city. Particularly, it differentiated areas in the historic city center from recently developed suburban areas. A wide dispersion in descriptor values and in the average count of incident reports within groups hampered the use of cluster analysis to characterize susceptibility of identified groups to flooding incidents. Clusters 1 and 2 delivered a relatively clear result; they covered the majority of the area (60%), accounting merely for a quarter of complaints. Response data was a bottleneck for the present study. If abundant response data becomes available for future research, partitioning analysis methods, such as cluster analysis or decision tree analysis, can be used more consistently. MLR could be run within urban subclasses, differentiated by cluster analysis. Future use of cluster analysis can also focus in imperviousness and distance to outflow, plus additional variables not explored in this study. Information about underground drainage networks, for instance, is a key aspect that must be considered in the future. Also, smaller grid-cells for data aggregation and sampling could also be used if enough information about both independent and response variables becomes available.

Most of data variability was captured by two theoretical gradients in performed PCAs, proving the relevance of its results for reducing the number of descriptor variables from 7 to 5. PCAs also showed two variables scoring high in the two first PCs: imperviousness and distance to watershed outflow point. This agreed with results obtained later in MLRs. These results showed that PCA is useful in urban flooding research to explore multiple variables with unknown relationships with the response variable, and identify descriptors with strong explanatory power.

MLRs provided significant insights into the relationships between studied descriptors and the occurrence of incident reports. Impervious ratio and average distance to outflow were the only significant descriptors, with coefficients close to 0.7 and -0.2, respectively, and 50% explaining power. The sign of the latter coefficient indicates that areas close to watershed outflows are more prone to flooding; it suggests that flooding from surface waters, or overloading or backwater effects in the drainage system in these areas, are main flooding mechanisms. The two significant variables were found not to be collinear in the PCAs, which indicates that they describe two different phenomena, probably affecting the occurrence of flooding incidents in two different ways. Results of MLRs indicated that additional variables must be explored to better explain occurrence of flooding incidents. Variables describing hydraulics, age, and maintenance conditions of drainage systems are key descriptors that should be further investigated.

An important source of uncertainty in this study lies in the nature of response data; reports are made by citizens. Their motivation is subject to personal conditions that cannot be fully explained only by adding explanatory environmental and socioeconomic information to a MLR. Still, significance and explaining power found in this study revealed the value of flooding incident reports for addressing the problem of urban flood modelling, which is commonly hindered by meager response data. The limitations posed by the uncertainty of incident reports could be overcome by combining incident reports with other response datasets; further exploration of multiple response data, such as insurance claims and data collected through social sensing (smart-phone applications, information retrieved from social media), can improve significance and explaining power.

Grid cells size influence modeled outputs. In this sense, quantity of openly information, specially response data, limits the assessment of the

cell size on results. As response data are scarce, aggregating them at higher spatial resolutions increases the number of cells accounting for single reports or no reports at all. On the other hand, increasing the cell size reduces the number of sampling sites, which impacts statistical robustness of analysis. For future research, this limitation can be overcome by focusing in multiple heavy rainfall events along, e.g., a decade. Other alternative is to enrich response data with possible addition information sources such as social and news media, fire brigade and 112 phone call records, or with images and video from traffic, security, and other cameras installed in the city.

Further improvement can possibly be achieved by relying on better urban imperviousness estimations. Even though it was based on an accurate land use model, the imperviousness ratio computed in this study is a rough approximation to urban imperviousness, and it was not validated. Given the significance of imperviousness in explaining the variability of incident reports, future research must be based on a better imperviousness model.

The observed significance of variables derived from open spatial data sources, in explaining half of the variability of flooding incidents occurrence, is a remarkable outcome. Datasets had different types, formats, and resolutions, and were produced independently by different organizations for disparate purposes. In spite of the spatial aggregations made, and the heuristic processing to compute weighted averages of impervious ratio per km^2 and distance to watershed outflow point per km^2, these metrics yielded significant insights for explaining incidents occurrence. Given the lack of precise urban hydrodynamic models and urban-focused flooding impact models, this data-driven outcome is highly valuable. This also highlights the importance of open and social-sensed data for future urban flooding modelling.

An important contribution of the present exploratory study is the indication of valuable information and methods in the modelling of urban pluvial flooding. This chapter tackled a problem with scarce experimental data available, uncontrolled conditions, and unknown relevant variables. It represents a first step research approach in which a set of previously available information is methodically evaluated. On the basis of achieved results, future research could engage in predictive modelling, in which obtaining higher accuracy could be a success indicator. Contrastingly, in a case with restricted prior knowledge about the urban pluvial flooding

phenomenon, achieving 50% of explanatory power is a clear, significant sign of the value of a selection of open available information to explain the occurrence of flooding incidents.

Insights provided by this study can support future design and implementation of efficient measures against urban flooding incidents. As the work presented in this chapter is of an exploratory nature, a direct application of the obtained insights obviously requires additional research steps. However, urban planners can consider found links between imperviousness and proximity to outflow, and the reports occurrence, to focus in effective measures against urban flooding. Results clearly point at the reduction of imperviousness in the urban environment as a possible effective flooding management alternative in Amsterdam. The proximity to outflow overland flow points can be used as a metric to prioritize preventive drainage maintenance. Emergency response management can use imperviousness maps to pinpoint critical areas during pluvial flooding events. A more sustainable and transparent design of premiums for insurance against pluvial flooding could use incident reports and imperviousness spatial distribution. These are some examples of how evidence provided by this chapter can be useful in practical applications.

Open data and social sensing data can provide valuable response data to validate urban flood models, foster realism in descriptor variables, and enhance the validity of flooding predictions. Further investigating these sources is thus key for planning of climate adaptation measures and efficient operation of existing drainage infrastructure.

4

Automatic detection of urban flooding from street images and video

Increasing urban flooding impacts challenges cities worldwide to devise smarter adaptation measures. Urban floods disrupt drainage, transportation, electricity, and medical services, and may trigger waterborne infectious disease outbreaks. Design of effective adaptation measures require precise and reliable urban flooding models. These are currently unavailable due to a lack of sufficient flood incident data. Crowd sourced images and surveillance videos are a potential source of such data. They can provide details about the timing, location, and extent of urban flooding. Computer vision techniques have been used to automatically classify images and video content to detect road conditions and water levels in embankments. Their usefulness for urban flooding has not been tested yet. In this chapter, the potential of mainstream image and video recording, and well-known and accessible computer vision tools for delivering key information about localized urban flooding incidents, is explored. Scene classification was applied to a dataset of images queried in a common web search engine, and foreground detection was performed on the video of a controlled street flooding experiment. Obtained results suggest that image and video information, such as crowd-sensed pictures and surveillance footage, can be used to retrieve important information on flooding incidents. Data sources and required computer vision tools are currently readily available for application in cities worldwide, and represent an opportunity to leverage existing sensing infrastructure to better understand and prevent urban flooding impacts.

4.1 Introduction

Cities need to adapt to changes in climate that lead to increasing likelihood of intense rainfall, associated with higher urban flooding risks. Urban floods affect critical daily activities as they interrupt drainage, transportation, electricity, and medical services (Ashley et al., 2005; ten Veldhuis et al., 2011). Pluvial flooding may also trigger waterborne infectious disease outbreaks in cities (de Man et al., 2014; Sales-Ortells et al., 2015; ten Veldhuis et al., 2010; Wade et al., 2014). Urban flooding risks are increasing due to heavier weather, lower permeability, higher population and assets densities, and aging infrastructure (Bates et al., 2008; Murphy et al., 2009). Smart infrastructure, preventive maintenance, and real time emergency responses can help citizens and governments to reduce urban flooding impacts (Gaitan et al., 2014; Jacobs, 2012; Melo et al., 2015; ten Veldhuis et al., 2011; Wong and Brown, 2009).

However, such adaptation measures require precise urban flooding information to develop reliable models for flood prediction that are currently not available. Design and implementation of such models is hindered by a lack of sufficient data on flood incidents. Apart from precise and timely meteorological information, and an adequate knowledge of the urban drainage infrastructure, the calibration and validation of urban drainage and flooding models requires data about the timing, location, and extent of local flooding taking place when the drainage system fails (Deletic et al., 2012; Dotto et al., 2012; Fontanazza et al., 2011; Gaitan et al., 2015; Maksimović et al., 2009; Ochoa-Rodriguez et al., 2015).

Monitoring local urban flooding implies technical challenges. Localized flooding is typically rapid in onset and of short duration. Local floods can also occur simultaneously at different city locations, demanding multiple installed detection units. Typical monitoring in current urban drainage systems include pumping logs, discharge measurements at a limited number of locations towards the end of urban catchments, and water levels in certain components. Actual extents or depths of flooding on the streets are seldom measured. Besides, acquisition and maintenance of monitoring devices is costly.

Cameras have been used to monitor water bodies. Different approaches

have been used in unmanned ground navigation systems, usually with demanding hardware requirements. Matthies et al. (2003), for example, employed four techniques to automatically identify the location of water bodies from image data:

- image classification to detect sky reflections in water bodies during the day;
- light detection and ranging (LiDaR) to measure reduced return signal due to specular reflection and absorption through the water column, and diffuse reflection on the bottom of the water body;
- short-wave infrared imagery to detect water bodies due to specific absorption coefficients of relatively deep and clear water bodies at these wave-lengths;
- and mid-wave infrared to differentiate thermal emissivities of water bodies from surrounding terrain during the night.

Sarwal et al. (2004) sensed the partial polarization effect of light reflected on the surface of water bodies by using three polarization filters at three different angles.

Such approaches have proved to be capable of labeling water body pixels in recorded images with satisfactory performance under different lighting conditions. However, they require either the acquisition and deployment of cameras of very high specifications, or a dedicated setup of conventional cameras using polarization filters, and registering and warping image pre-processing. This makes these approaches costly and technically demanding for a city-wide scale implementation for water management purposes.

More recently, photo- and video-cameras have been used to measure rainfall intensity and river levels in urban environments during flash floods. Allamano et al. (2015) detected raindrops in image frames recorded by a dedicated conventional camera pointing at a static background, and computed their size to derive rain rates with an associated uncertainty. Lo et al. (2015) used video frames recorded by closed-circuit television cameras, on which they set virtual markers at known locations of a river embankment, and checked when they were covered by water-classified image segments, to obtain dynamic flooding levels.

Street video imaging offers a potential source of unexploited information to detect street flooding. Video recordings from traffic, web, and smartphone cameras contain information about timing and extent of local flooding. The cases of The Netherlands and Nairobi, in Kenya, are only two examples of the spread of urban imaging infrastructure. In The Netherlands, several thousands of public traffic monitoring cameras stood on built areas across the country in 2013 (Dutch Ministry of Security and Justice, 2014). Another 325 street cameras in this country, installed on amateur weather stations (Weather Underground), represent a valuable source of street video footage. In Nairobi, Kenya, an infrastructure including 1,500 urban surveillance cameras is being installed (Kenya National Government, 2015). Other privately owned cameras are also installed and streaming live time lapses. Additionally, smartphones and other mobile devices gather images into data stores that can be made available for research (e.g., Michelsen et al., 2016). Videos posted on social or collaborative media (e.g., Instagram, YouTube, Twitter, Mapillary) are another potentially useful source, if information about geographical position is included.

Surveillance camera images have been used recently to distinguish road conditions under different lighting situations for intelligent transport systems. For example, (Horita et al., 2012) used the signal of headlight reflections to distinguish road conditions, and (Shibata et al., 2014) used passive lighting reflections. Both works are based on a machine learning approach using training sets to classify texture features associated with dry, wet, and snow covered road conditions. These works proved that weather effects on roads can be detected using computer vision techniques.

Using advanced computer vision techniques implies knowledge and software prerequisites that may be hard to fulfill, in the case of environmental research. Nowadays, well known computer vision routines are becoming available as easy-to-use programming tools. This facilitates agile implementations in tests in interdisciplinary problems, such as in the monitoring of urban flooding impacts. OpenCV (Intel Corporation et al., 2014), for example, is a C++, open source software library providing a wide range of common computer vision routines. It offers wrappers for Python, C, Java, and MATLAB. Since its initial release in 1999, OpenCV's users community has been joined by major research centers and IT companies (Bradski and Kaehler, 2008).

The availability of street image and video material, and accessible computer vision tools, offer the opportunity for testing their feasibility for retrieving information on flooding location and timing in cities. Given the enormous amount of video frames that need to be stored and processed when searching for local floods in street videos in the case of a city-wide application, flood detection should be assisted by computer vision techniques. The aim of this study was to explore the potential of mainstream image and video recording, and well-established and accessible computer vision tools, to deliver key information about localized urban flooding incidents. This chapter focuses on two computer vision techniques: scene recognition and foreground detection. Two questions are addressed. First, can computer vision techniques recognize local flooding scenes in the type of imagery gathered in social media? Second, can these techniques detect the extent of a puddle in a series of videoframes, of similar characteristics of those captured by traffic and surveillance cameras.

This chapter is organized as follows: Section 4.2 presents details about image gathering and preprocessing and local flooding scene and puddles recognition. Section 4.3 presents and analyzes results. And finally, Section 4.4 discusses these results and provides an outlook of the application of these technologies in the field of smart urban water management.

4.2 Methods

After acquiring images and video material (see Section 4.2.1), detection of local flooding from street video footage was approached from two angles. First, we determined the presence of a puddle in a given set of scenes. Then we attempted to approximate the visible puddle region in a series of video frames. The earlier approach employed the scene recognition technique discussed in Section 4.2.2. The latter extracted the background from each frame of a video, leaving pixels corresponding to objects with changing reflectivity, labeled as flooded (Szeliski, 2010). Section 4.2.3, documents such implementation.

4.2.1 Data acquisition

Annotated Internet images

Annotated images were obtained from the Internet with a common web search engine (Google Images, June 2015. Available at images.google.com). The search terms “street puddle" and “street" were used to query respective image sets. Each set was labeled accordingly, ensuring proper annotation. A search filter was set to retrieve only photos, excluding clip-arts, line drawings, and animations. No image-size filters were set. After downloading, images were checked. The ones including people and editing such as typographies or evident graphic effects, were excluded.

Recording puddle videos

Local flooding was simulated by blocking the inlet to a gully pot, on the street of an experimental lot close to the building of the Faculty of Civil Engineering and Geosciences, at Delft University of Technology. A flooding simulation was produced by spraying and pouring water with a hose connected to a nearby water tap.

Recording was done with a mainstream videocamera, mounted on a tripod on a fixed position, pointing to the blocked gully pot and the created puddle. Employed camera was a GoPro HERO3+ Silver Edition, featuring 8 bits color depth, up to 1920×1080 pixels video resolution and 3680×2760 pixels photo resolution (applicable to time lapse recording), up to 170^{o} field of view, 2.8 to 6 focal ratio, and built-in Wi-Fi.

Image preprocessing

Image preprocessing was done using Python (Python Software Foundation, 2014) and the Python Image Library (Lundh, 1995). Annotated images were translated to white and black using the ITU-R 601-2 luma transform (Lundh, 1995; Solem, 2012), in which the final pixel value (L) equals $0.299 \times R + 0.587 \times G + 0.114 \times B$, where R, G, and B are the values of red, green, and blue channels of a given pixel. Images were then cropped by one pixel in the left or top sides, when width and/or height, in terms of the number of pixels, were odd. This was required before cropping images to a square aspect, avoiding differences of original dimension ratios. Finally, following Solem (2012), resulting images were resized to 100×100 pixels; this was done with a nearest neighbor interpolation (Lundh, 1995). This resizing ensured that all images fit the same sampling grid during the computation of local features, as described in Section 4.2.2.

4.2.2 Scene recognition in web images

Computer vision can be used to automatically recognize images. Scene recognition is based on binary classification where images are labeled as belonging to a previously defined class or not. This procedure is done in two steps. First, local image descriptors, or features, are computed. Then, these are used to locate each image in a multidimensional space, with as many axes as computed features, allowing for a straightforward K-nearest neighbor classification (Solem, 2012).

Description of local features

Local image descriptors define local features, which are used to compare similarities between images. Scale Invariant Feature Transform (SIFT) describes features consistently across different scales, rotations, intensities, and view points. The core of SIFT consists of finding key points with local minima and maxima differences, between Gaussians of the images across adjacent scales and rotations (Lowe, 2004; Solem, 2012). Local image features are stored as histograms indicating the magnitude and orientation of features gradients. These histograms are used in the classification described below. A dense SIFT (DSIFT) was applied to images after preprocessing. DSIFT consists of the computation of local features within the cells of a grid imposed on the images. The number of described features in all images is the same given that they had the same size and resolution after preprocessing. Descriptor arrays bear a representation of the Gaussian difference gradients in an approximation of eight directions. This representation is stored in a vector that concatenates the values of each region, of each descriptor. Features in each image were thus represented by a list of feature vectors (Lowe, 2004; Solem, 2012).

K-nearest neighbor classification

For the K-nearest neighbor classification, the data set was divided in two groups, training-set and test-set, with approximately the same number of samples. This classification works by comparing the positions of a given test image with those of all training images in a Euclidean space, whose axes correspond to the computed visual features. The test image was tagged with the label of the closest training image. This process was repeated on all images in the test set (Solem, 2012).

4.2.3 Background subtraction and puddle detection in video footage

Background regions were subtracted from videos obtained as described in Section 4.2.1, leaving detected foreground regions behind. Foreground extraction aimed at delimiting wet and flooded areas. As treated videos were recorded under controlled conditions, most of these foregrounds were expected to represent puddles. In real street conditions, different moving objects, such as cars, bicycles, and animals, are expected to pass by through scenes. In the experiment, the signals produced by a pedestrian were compared to those of pixels becoming humid or flooded.

Two techniques were tested to differentiate background and foreground areas in analyzed videoframes: Gaussian mixture models (GMM), and interactive foreground extraction using iterated graph cuts (GrabCut). The aim was to classify foreground pixels as wet or flooded areas, and background pixels as remaining areas.

K Gaussian mixture for background subtraction in video frame series

This technique compares individual pixel values at a video frame with the historic values of that pixel along the video stream. Historic values are represented with a mixed Gaussian distribution, built from the time series of the same pixel values along the video footage: the longer the combination RGB pixel values remains unchanged in the video, the higher the likelihood it is a background pixel. In contrast, pixel values of moving objects appearing in the scene deviate from normality as a result of their extremely changing value. Changes in the extent of casted shadows along the video stream are addressed by using a color model in which chromatic and brightness components are distinguished. If brightness and color distortions lay within a given statistical threshold, the sampled video-frame pixel is considered a moving shadow. The technique is described in KaewTraKulPong and Bowden (2002).

Background subtraction in single video frames

Interactive GrabCut was used to extract background areas from single videoframes. GrabCut iteratively uses a segmentation algorithm that models background and foreground areas using GMM. GrabCut requires an user-provided bounding-box (or rectangular area) enclosing the foreground

object. Initially, GMM for background are defined from values of out-of-box pixels, while GMM for foreground are provisionally based on values from within-bounding-box pixels. With each iteration, unlabeled and provisionally-labeled pixels are classified as being part of the background or not; once new pixels enter or exit each class, their GMM are updated. In a second step, a refinement of GrabCut results can be done by passing annotated values of sample background and foreground pixels (seeds). This is used to update GMM for background and definitive foreground pixels, and initialize a new series of classification iterations. Details of this method can be found in Rother et al. (2004)

Performance indicators

Ground truth and classification values obtained in Sections 4.2.2 and 4.2.3 were used to compute contingency tables. True positive (TP), false positive (FP), true negative (TN), and false negative (FN) values in analyzed case studies were derived from the contingency tables. These values were then used to obtain Sensitivity ($\frac{TP}{TP+FN}$), Specificity ($\frac{TN}{TN+FP}$), Positive Predictive Value ($\frac{TP}{TP+FP}$), and Accuracy ($\frac{TP+TN}{TP+FP+TN+FN}$) measurements. Sensitivity indicates how well the methods correctly detect puddles relative to all puddle samples. Specificity is a measure of the methods' capacity to detect non-puddle conditions, relative to all non-puddle samples passed. Positive predictive value indicates the fraction of properly detected puddle samples relative to all samples classified as puddle. Accuracy is a general measurement that indicates the proportion of proper puddle and non-puddle classifications out of all samples.

For the case of scene recognition, indicators were computed using the labels of downloaded images. Thus, for each image in the test scene, there was one true value and one classified value. For the foreground classifications, humid and flooded masks were manually drawn for a single frame. These masks were used to identify truth pixels, that were then compared with classified pixels in the same frame.

4.3 Results and analysis

4.3.1 Puddle scene recognition

A total of 58 non-puddle and 45 puddle images were obtained from the web query, as described in Section 4.2.1. Obtained images included disparate

Figure 4.1: Example of acquired and pre-processed non-pudle (a, b, c) and puddle street images (d, e, f).

image sizes and resolutions, fields of views, pavement types, geographic locations, and environmental conditions. Querying the web search engine for street puddles successfully delivered images containing puddles in most of the cases. The images were processed as discussed on Section 4.2.1. A sample of resulting images is presented in Figure 4.1).

Annotations made by people when uploading images by naming their file names accordingly, or by referring to puddles in texts of web pages where they are placed in, are used by common web search engines to index such images. These annotations imply an existing classification that can conveniently be used for research purposes. They were observed to be precise enough to be used in machine learning and training algorithms to recognize local flooding scenes.

4.3.2 Puddle scene recognition performance

The size of training and testing sets are shown in Table 4.1. The confusion matrix of scene recognition classifications can be found in Table 4.2. Table 4.3 lists performance measurements for the performed scene recognition. Performance measurements refer to the 51 images classified for the test set. These results are following discussed.

Sensitivity: Puddle scene recognition was highly sensitive; 91% of puddle scenes were properly classified as puddles, only 9% of actual puddle scenes was not properly classified as such. An alarm system based on this classifier would be able to properly recognize 91% of all pictures referring to flooding scenes.

Positive predictive value: From the 31 images classified as puddle scenes, just 20 were actual puddle scenes; resulting in a positive predictive value of 65%. If an alarm was set on the basis of this puddle detection, 35% of alarms would be false. This contrasts with sensitivity results; while 9 out of 10 puddle scenes would be detected, only 7 out of ten puddle-tagged scenes would correspond to actual puddle scenes. If, for instance, this recognition were used to identify flooding locations for real-time reaction, 3 out of 10 deployed operations would find no flooding.

Specificity: The classification properly detected non-puddle scenes 62% of times, this is a specificity of 0.62. By comparing this measurement with the sensitivity, it becomes clear that the system has a tendency to classify non-puddle scenes as puddle-scenes.

Accuracy: The overall accuracy of the technique was 75%; which implies that the scene recognition failed, jointly in positive and negative classifications, 25% of the times. The higher sensitivity value with respect to the specificity value indicated that the recognition performs better on detecting puddle scenes than non-puddle scenes.

Table 4.1: Training and testing set sizes used in preliminary classification.

	Puddle	Non-puddle	Total
Training set	23	29	52
Testing set	22	29	51
Total	45	58	103

4.3.3 Detected foreground

A series of 443 video-frames was collected, with captures at 1 frame/s. Start and ending time were 14:18:01 and 14:25:24 CEDT, September 10, 2015. Objective bearing was aprox. 30° S−E. Figure 4.2 shows an example video frame of the controlled flooding. The footage recorded an area becoming flooded, during a dry summer afternoon. Non saturated areas are observable around the puddle area. Different object shadows, including

Table 4.2: Confusion matrix of classification results.

		Truth		
		Puddle	Non-puddle	Total
Classification	Puddle	True positives (TP) 20	False positives (FP) 11	31
	Non-puddle	False negatives (FN) 2	True negatives (TN) 18	20
	Total	22	29	51

Table 4.3: Performance measurements in classification of puddle scene.

Measurement	Computation	Value
Sensitivity	$\frac{TP}{TP+FN}$	0.91
Positive predictive value	$\frac{TP}{TP+FP}$	0.65
Specificity	$\frac{TN}{TN+FP}$	0.62
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	0.75

those of moving tree leaves and branches, are also present. Seven areas with different scene characteristics areas are indicated with red squares in Figure 4.2. Characteristics of these areas are described in Table 4.4. As a result of the close-range, wide view-angle imaging employed in this experiment, objective-to-target distance and incidence angle are noticeably different for each of the sampling windows. This results in objects with apparent distorted sizes: e.g., tiles of the same size appear to be bigger or smaller depending on how close they are to the camera. Changes in digital numbers (DN) of pixels within such areas during the footage are shown in Figure 4.3 and following described.

Description of sampling windows

Frame 2 Environment at initial state. Dry conditions. No specular reflections due to wet surfaces. No pedestrians passing by. No visible droplets.

Frame 30 No water was sprinkled up until this point; "dry", "humid", "flooded", and "reflecting" sampling windows only present non-shadowed,

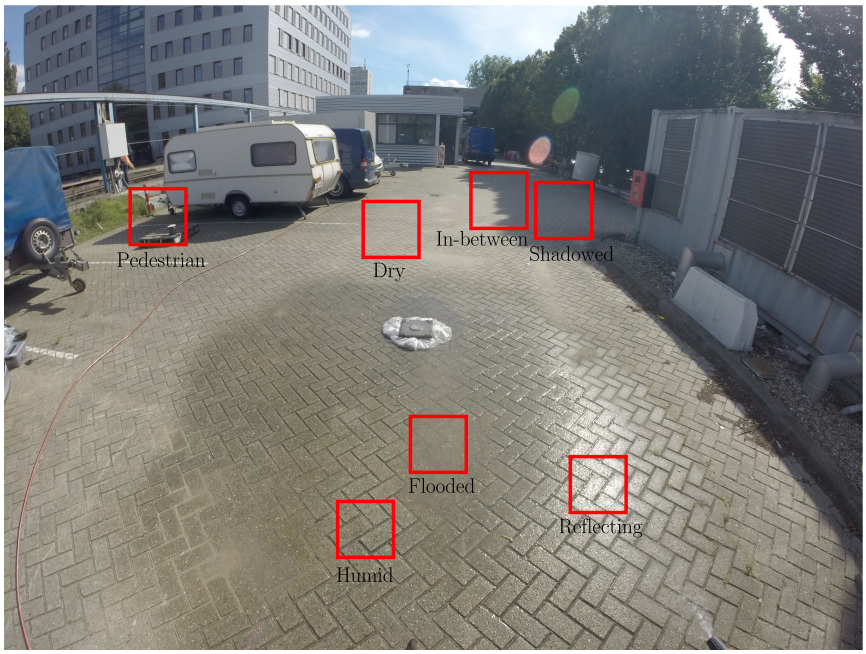


Figure 4.2: Location of sampling windows.

free of standing objects, paved areas. Signals from these four sampling windows are similar. DN deviations in the three color channels are small, with blue mean being the lower, and red mean being the higher, with DN values ranging from 120 to 160.

The "flooded" sampling window shows narrower deviations than the other three afore mentioned windows. Apart from the effects of imaging range and view angle, bricks in this area were covered by a higher amount of sand. This causes pixels to differ less. Tile edges appear to be comparatively blurred. The presence of sand in this area is a result of its location along the local overland flow-path. Sand located elsewhere is dragged towards this area before rain water enters the sewer.

"Pedestrian" sampling window has wider deviations and lower DN means. There a is small peak visible in this window, which can be related to droplets visible in the respective still-frame. These droplets were actually part of the water being spread over the area during the experiment.

DN mean values of the three color channels in the "in-between" sampling window are very close to each other, and have a wide distribution. In this sampling window, DN responses show stronger oscillations as a result of trees leaves and branches being shifted by the windblown.

The "shadowed" sampling window has the lowest DN means and narrow deviations. In contrast to other sample windows, mean of red channel tends to be lower than green and blue channels, with blue channel mean tending to be higher than the other two channels. This trend is maintained throughout the footage.

From this frame on, some tiles became wet; means of three channels in "Humid" and "Flooded" sampling windows start to drop.

Frame 85 "Dry window" signals show a slight increase, possibly related to natural increments of solar radiation. "Humid" sampling window shows a signal drop from DN means between 120 and 150 at frame 30, to 100 and 130 at frame 50, plateauing from there on. This is around 12% drop over 20 seconds, or around 6% drop every 10 seconds before plateau. The mean values drop is about 20 DN.

In the "flooded" sampling window, signals also dropped but did not reached a plateau. They drop from DN means between 140 and 160 at frame 35, to 140 and 160 at frame 85, to 110 and 130 at frame 130. This is around a 20% signal drop over 95 seconds, or around 2% drop every 10

seconds before plateau. The means drop is about 30 DN. Compared to "humid" sampling window, the signal drop is less steep but larger.

Contrary to the other two windows with areas sprinkled with water, mean DN in the "reflecting" sampling window increased. Apparently, this area became wet until around frame 85. Before that, signal increment was similar to that of "dry" sampling window.

The increment rate also observed in "pedestrian" and "shadowed" windows is similar to the one in the "dry" sampling window. Peaks in "shadowed" sampling window around frame 60 are the result of droplets with relatively bright responses.

Frame 100 The most important change at this frame is the mean of DN signals reaching a plateau in the "reflecting" sampling window. Right after the area within that window is sprinkled, DN means showed a steep increase of about 25% in just 7 frames, which means an average signal change of about 4% every second. The series of still frames for this window clearly shows the strong brightness changes in its pixels.

Frame 160 DN means and deviations in all windows are sustained since frame 100. The effect of droplets results in higher DN deviations in the "flooded" sampling window; such droplets are also visible in the still frame "e" of this window. Droplets disturbing the water mirror in the "reflecting" sampling window may be also the reason for the oscillations observed in the DN means and deviations in that area.

Frame 349 and beyond Signal responses are sustained in general. The effect of a pedestrian walking around, visible in still frame "f" of "pedestrian" sampling window, caused the signal spikes in its DN means and deviation. This characteristic signal, easily differentiable from the patterns discussed above, can easily be isolated.

Table 4.4: Description and location of sampled areas in video.

Sample	Location (min X, min Y, in pixels)
1 – Flooded area; a puddle of a certain depth takes place.	2176, 950
2 – Humid area; wet pavement below oversaturation.	1786, 495
3 – Sunlight reflecting area; wet pavement reflecting sunshine	3032, 735
4 – Dry area; dry pavement with relatively constant conditions.	1923, 2100
5 – Shadowed area; pavement mostly covered by shadows.	2847, 2201
6 – In-between area; pavement partly covered by shadows.	2500, 2255
7 – Pedestrian area; pavement on which a pedestrian passes by.	511, 1846

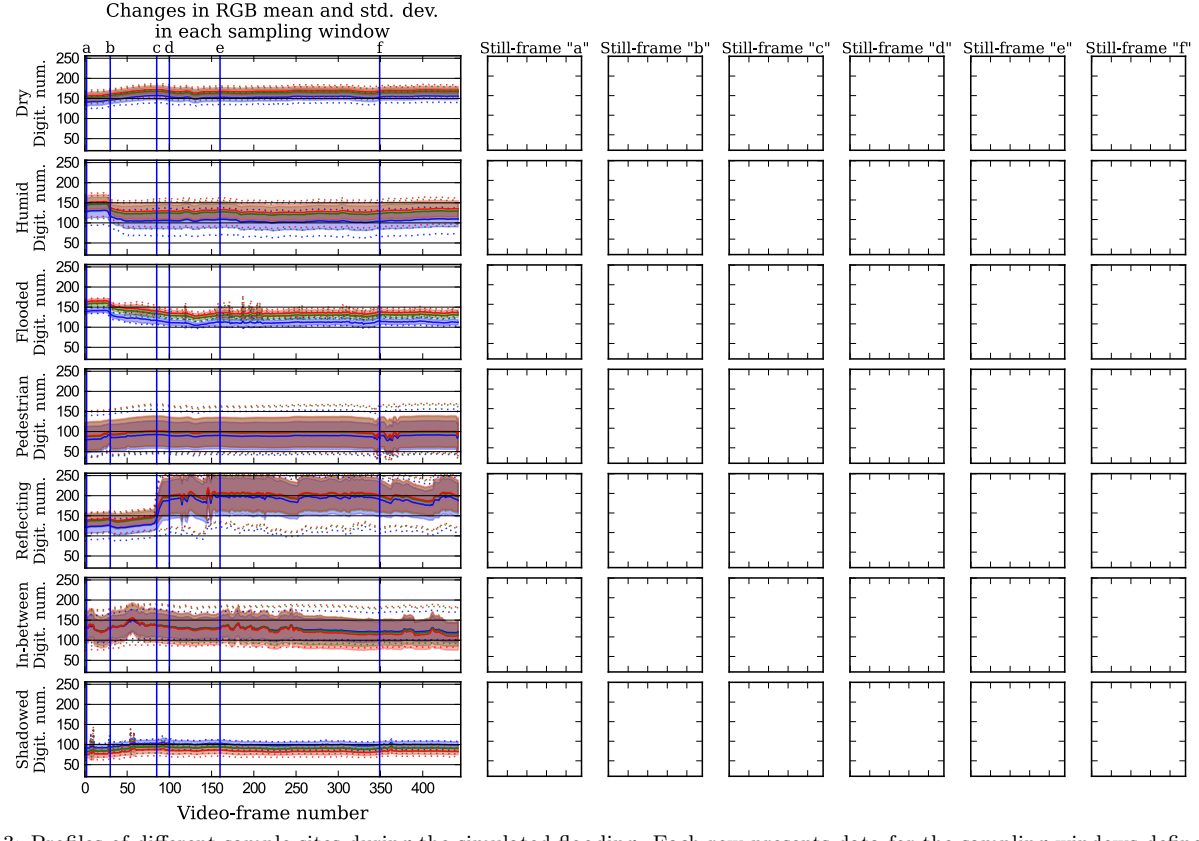


Figure 4.3: Profiles of different sample sites during the simulated flooding. Each row presents data for the sampling windows defined in Figure 4.2. The first column of plots from left to right shows the profile of DN means of red, green, and blue channels (with lines in respective colors) and standard deviation ranges (with shadowed areas in respective colors). 5% and 95% percentiles are shown with respectively colored, dashed lines. Vertical blue lines indicate timing of still-frames presented in the rest of plot columns. Literals 'a' to 'f' are used to indicate frames 2, 30, 85, 100, 160, and 349 (see top of plots in first row). A video of this sequence can be found at this URL: youtu.be/hRf-SpdYy9U (Gaitan, 2016).

GMM and GrabCut classified foreground

GMM detection Foreground detection was done for all frames of the simulation video. The model parameters were set as follows. History, which is the number of frames used to build initial background models was set to 30 so that only dry conditions are considered (see Figure 4.3). The number of GMM models was set to 5. The considered background ratio was approximated to 0.7 (see truth masks in Figure 4.6 in Section 4.3.4). The learning rate was set to 0.001 to enforce the model to maintain a dry background model. Using different parameter sets did not deliver significantly different puddle detection results.

Some of the changing areas in the scene were detected by the GMM classification. Detected foreground corresponded to specular reflection areas, moving tree branches and leaves, and passing pedestrian. The position change of the hose employed for the flooding experiment was also detected. Most of the humid and flooded areas were not detected. Figure 4.4 shows detected foreground for frame 349. GMM classification differentiated reflecting and pedestrian pixels from the background. It did not satisfactorily detect humid and flooded areas. Changes in the DN responses of specular reflections associated with wet ares are clearly greater than those becoming humid or flooded. Such differences were less accentuated when responses from pedestrian pixels were compared to flooded and humid pixels. DN mean changes due to wetness and flooding are bigger than those due to the pedestrian passing by, but the latter delivers more extreme values (see Figure 4.3). The detection of the pedestrian and the missed detection of humid and flooded pixels may be related to the extreme values produced by the pedestrian passing by.

An analysis of the performance of this detection is presented in Section 4.3.4.

GrabCut detection This interactive classification method builds background and foreground models for a single image frame on each iteration. The method was used for frame 349, which presents dry, humid, flooded, and reflecting areas, and pedestrian passing by. One detection was done setting foreground seeds including humid and flooding areas. A second detection only included flooded areas in the foreground seeds, and humid areas in the background seeds. This was done to test whether GrabCut was able to differentiate humid from flooded areas in that particular frame.

Figure 4.4: Detected foreground for frame 349 using GMM. Visible pixels are the ones classified as part of the foreground. Black pixels were classified as background.

Results of this classification are shown in Figure 4.5.

In simultaneous humid and flooded detection, GrabCut classification roughly excluded the block on the sewer from the background. A part of it was still visible in the detection, in spite of having been manually labeled with seeds as background. A small paved area to the left of the frame was also classified as foreground. It is possible that this area was unintentionally sprinkled with water before starting the simulation.

Regarding the second GrabCut classification, performed specifically to distinguish flooded and humid flooded areas, results showed that this technique was not able to perform such differentiation. Only the areas in and around the background seeds, passed within the initial polygon mask, where enforced as background in the final classification. Most of the remaining area within that polygon was classified as foreground.

4.3.4 Puddle detection performance

A comparison of GMM and Grabcut performance was made for frame 349 (see Figure 4.3). Three truth masks were used for computing performance results of foreground detection in frame 349. Figure 4.6 shows truth masks drawn for that frame. Four different sets of performance metrics were computed for the GMM classification: detection of humid plus flooded areas, only–humid areas, only–flooded areas, and reflecting areas. Two sets of performance measurements were obtained for the GrabCut classification:

The selection of those sets was made given that GMM classification tended to detect reflecting areas, omitting humid and flooding ones, and GrabCut returned well-detected humid and flooded areas, without being able to differentiate one from another. Table 4.5 shows computed performance measurements. Those measurements are analyzed further.

Sensitivity GMM detection showed very low sensitivity overall. The highest sensitivity was delivered for the classification of reflecting areas, where 10% of them were properly detected. As this is a pixel-level classification, this low value can be related with the existence of non-reflecting pixels within the area used as truth-reflecting mask. If this approach was used at its current status in, e.g., a urban flooding alarm application, it would miss puddles 90 to 100% of the times. If all puddles were reflecting, only 10% of their area would be properly detected. It is worth noticing that the GMM performance values in the two first columns of Table 4.5 are highly

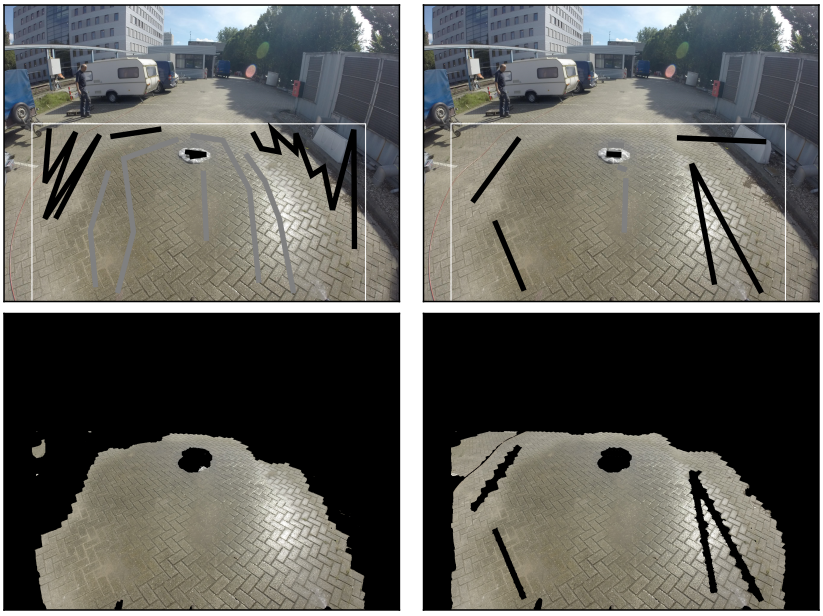


Figure 4.5: Puddle detection using GrabCut for humid–plus–flooded detection, and only–flooded detection. The area of initial mask is enclosed by a white polygon in top frames. Background and foreground seeds are shown in black and gray, respectively. Top-left frame shows how the foreground seeds include humid and flooded areas in the first GrabCut detection. Top-right frame shows the only–flooded areas used as foreground seeds in the second GrabCut detection. Bottom-frames show respective detected foregrounds.

similar. This is the result of the few pixels within the flooded truth mask classified as foreground. GrabCut classification was highly sensitive in both cases: for detecting humid plus flooded areas, and for only flooded areas. It would return 99% of the scene areas covered by humid plus flooded areas if it were used for a flooding alarm system.

Positive predictive value Part of the foreground classified using GMM referred to reflecting areas. As reflecting areas are part of humid areas, their GMM detection has a much better positive predictive value than sensitivity. From all reflecting-classified pixels, half corresponded to non-reflecting true pixels. GMM's positive predictive value on detecting flooded areas is as low as its sensitivity (0); flooded areas showed few reflections. Additionally, GMM-detected foreground included tree– and pedestrian–related areas. Positive predictive value of GrabCut classification was high for the humid plus flooded detection, but extremely low for the only-flooded detection. From all pixels classified as flooded, only 6% were actually in flooded areas. At its current state, the efficiency of this approach to detect humid pixels is not pertinent for urban flooding. As rainfall moistens the ground, all video frames would return humid detected areas.

Specificity GMM classification delivered high specificity values, especially on flooded and reflecting areas. Areas classified as foreground by this technique were small when compared to the actual truth masks. As flooded and reflecting truth masks were less extensive than the humid mask, the chances for the background classified areas to actually coincide with the true background were bigger. Such specificity values would be useful for an actual application if they would follow higher sensitivity and positive predictive values. That is the case of the GrabCut detection of humid plus flooded areas. On top of high sensitivity and positive predictive value, GrabCut delivered 99% specificity. This means that almost all pixels classified as background were actual background pixels. The high specificity obtained by the GrabCut classification on detecting flooded pixels contrasts with its extremely low positive predictive value. In spite of the high specificity and sensitivity values, just detecting humid areas does not provide an advantage for detecting puddles for urban flood management.

Accuracy The moderate accuracy achieved by the GMM classification on detecting humid plus flooded and only flooded pixels responds to the also moderate positive predictive value in those both cases. The very high

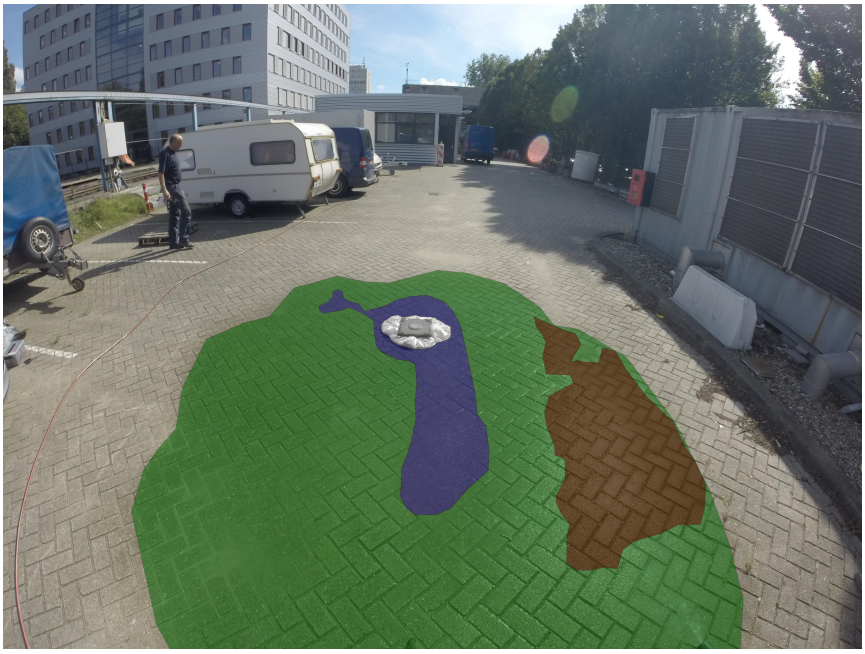


Figure 4.6: Manually–set ground truth masks for frame 349. In green: humid area. In blue: flooded area. In red: reflecting area.

accuracy of this type of classification on detecting flooded and reflecting areas can be explained by the difference of scales of TN with FP and FN, which is almost three orders of magnitude. Accuracy of GrabCut detection of humid plus flooded pixels reflects the rest of high performance measurements. Finally, GrabCut accuracy on detecting flooded pixels mirrors its very high sensitivity and very low positive predictive value.

4.4 Conclusion and outlook

This chapter tested the potential of using crowed-sourced-type images and mainstream videos to obtain information about localized urban flooding incidents. Two main methods were employed: scene recognition and foreground detection. Scene classification was applied to a dataset of images queried in a common web search engine. Foreground detection was performed on the video footage of a simulated street flooding.

Scene recognition results are promising. In spite of the limited training

Table 4.5: Performance measurements of puddle classifications.

	GMM classification				GrabCut classification	
	Humid + Flooded	Humid	Flooded	Reflecting	Humid + Flooded	Flooded
TP	5.90×10^4	5.88×10^4	1.34×10^2	5.00×10^4	3.31×10^6	2.56×10^5
TN	6.76×10^6	7.01×10^6	9.79×10^6	9.66×10^6	6.31×10^6	5.88×10^6
FP	4.44×10^4	4.45×10^4	1.03×10^5	5.34×10^4	5.00×10^5	4.02×10^6
FN	3.29×10^6	3.04×10^6	2.59×10^5	3.96×10^5	3.82×10^4	3.05×10^3
Sensitivity	0.02	0.02	0.00	0.11	0.99	0.99
Pos. pred. val.	0.57	0.57	0.00	0.48	0.87	0.06
Specificity	0.67	0.70	0.97	0.96	0.99	1.00
Accuracy	0.67	0.67	0.96	0.96	0.94	0.60

and test sets and the straightforward SIFT-KNN classification, scene recognition achieved a sensitivity of 90%. Building training sets was facilitated by annotating images as they were returned by each of both queries: "street puddle" and "street". Scene recognition indicated whether an image depicts a puddle scene or not. This technique does not provide information about puddles extent within the image. Its application can provide automatic detection of location and timing of flooding incidents. It can also be used to select images to be subjected to a puddle extent recognition.

These results prompt further exploration on the use of this technique in real applications for flooding impact prevention. This technique proved to be highly versatile. It worked with a set of disparate images, produced by different hardware, and shot at different locations, environmental conditions, and imaging setups. Considering this is a first implementation of the method, achieved high sensitivity highlights the usefulness of scene recognition for automatic puddle detection.

The usefulness of this technique should be further investigated by increasing sample-size and diversity of the image dataset, randomly changing the composition and size of training and test sets, and studying the impact this has on performance measurements. Nevertheless, achieved sensitivity results are surprisingly high for this set of images considering that it was grabbed from a straightforward and unrestrained manual query in a common search engine.

Future research should increase the data set and test other classification techniques such as Bayesian networks, Support Vector Machine, and deep learning. Image sources such as Weather Underground, Twitter, Twitpic, Instagram, and YouTube may be used.

Foreground classification delivered mixed results. GMM required the

definition of a small set of numeric parameters, delivering classified foregrounds for the series of frames in the video. GrabCut required manual input of a foreground-enclosing polygon and foreground and background seeds, delivering the classified foreground of a single video frame. In this sense, GMM classification is a more appropriate technique for an eventual city-wide flooding detection applications as it requires fewer manual inputs.

However, the best GMM performance, achieved on the detection of reflecting areas, was just moderate: 48% of pixels classified as reflecting were actual reflecting pixels, 10% of actual reflecting pixels were classified as such, and the accuracy on doing so was 96%. The detection of flooded areas had a sensitivity close to 0%.

GrabCut exhibited a remarkably high performance on detecting humid areas, while only 6% of all flooded-classified pixels were actually in the flooded area. At its current state, this approach has a very limited usefulness in an actual city-wide application.

Further research must be conducted in order to achieve satisfactory performances on automatic detection of flooded extents in videos. DN responses in characteristic areas seem to bear enough differences to distinguish flooded from humid areas. Signal filtering can be employed to that end. Additionally, the effect of GMM parameters on the classification performances requires further investigation.

Results obtained by this chapter suggest that image and video information, like crowd-sensed pictures and surveillance footage, can be used to retrieve important information about urban flooding incidents. These data, and required computer vision tools, are currently within reach of diverse cities worldwide, and represent an opportunity to leverage existing sensing infrastructure to better understand and prevent urban flooding impacts.

5

Conclusion

A precise understanding of the processes leading to urban flooding is key to formulate efficient measures to adapt cities to extremer weather, growing population densities, and aging infrastructure. Existing drainage models are often poorly calibrated due to the lack of data on water level and flow. This hampers their ability to predict when and where flooding can be expected on the streets. In this scenario, risk based approaches have begun to assist urban water management in recent years; there is a growing awareness of the usefulness of data driven approaches, given inherent uncertainties associated with complexity of urban hydrological response. Such approaches emerge as the sensitivity and effectiveness of urban flooding models and climate proofing measures require proper evaluations. These approaches entail data driven research, employing stochastic evaluations of available information. In this dissertation the potential of crowdsourcing and open data sources to provide key information about the timing, location, and extent of urban flooding incidents is evaluated.

Unconventional data sources comprise a wealth of information that can help to explain the occurrence of urban flooding incidents, particularly when availability of direct flood extent measurements and validity of hydraulic models are limited. Insights derived from alternative information sources can be integrated into urban flooding models, enhancing their accuracy and usefulness in the assessment and management of flood risk and climate change impacts. Flood incident data could indicate critical urban areas or drainage components to be retrofitted or preventively maintained, in order to reduce urban flooding damage.

This thesis investigates to which degree overland flow models, open spatial datasets, and publicly available pictures and video footage, can explain timing, location, and extent of localized flooding in cities. Part of the evaluation is based on data from Rotterdam and Amsterdam, two delta cities in The Netherlands, characterized by a flat topography, and increasing pluvial flooding occurrence. This thesis tackles three research questions: 1. What is the relation between topographic gradients and urban flood occurrence in a delta city? 2. To what extent can publicly available datasets explain the occurrence of urban flooding incidents? 3. To what extent are mainstream computer vision techniques capable of automatically detecting flood occurrence in imagery collected from web queries and typical street video footage?

The three respective main findings delivered in this thesis, their importance, and their implications, are discussed in the following three sections.

5.1 Topography does not explain flooding incidents distribution

The first research question this dissertation tackles is whether overland flow-paths constrain the spatial distribution of flood incidents in the case of a delta city, characterized by small ground level variations. The spatial distribution of flooding incident reports was related to the size of catchment areas along an urban overland flowpath network. A dataset of 21,577 flood incident reports over a period of 8 years for the city of Rotterdam was used for this analysis.

Flooding reports were used as a proxy to identify the occurrence of urban pluvial flooding. Underlying conditions at locations with higher flooding incidence should indicate key factors controlling flooding. As topography has been highlighted as a key urban flooding factor by previous studies, the first question tackled in this thesis is whether overland flowpath network structure is linked to the incidence of flooding reports.

Results show that the spatial distribution of flooding reports was clustered, but this pattern did not respond to urban overland flowpath gradients. Under the assumption of a blocked underground drainage system, a higher number of flooding incidents was not associated with low terrain elevations, in the flat topography of a delta city. Although urban topography may be assumed *a priori* to play a relevant role in the processes leading to urban flooding, it does not explain the incidence of flooding reports for the case study area in this analysis. This result contrasts with findings obtained by studies done in areas with larger elevation differences in previous studies.

Obtained results can be explained by the typical small elevation differences between streets of delta cities. Local flooding is expected to occur downstream locations of overland flow-paths, but its depth and duration is probably smaller than in cities with bigger elevation differences. Besides, this study was based on a register including flooding reports from eight complete years, not only during heavy rain events. Failures in the sewer infrastructure, such as inlet blockages, pump malfunctions, and pipe bursts, could have influenced the overall landscape of flooding locations.

This thesis shows that, in absence of validated drainage models, high resolution digital elevation models can be used to provide detailed information to model overland flow paths. More importantly, this thesis introduces an implementation of a spatial autocorrelation test, constrained to the overland flow-paths network space retrieved from such paths. This test can identify the strength and uncertainty of relationships between urban areas connected by inferred urban ephemeral streams.

Given the non-hydraulic nature of the model implemented in this thesis, and the assumption made, it is crucial to confront these results with validated hydraulic simulations in other delta cities. Additionally, as soon as rich flood incidents data becomes available, similar research should be also performed in environments with more accentuated slopes where topography might play a stronger role than in delta cities.

This implementation can also be used in other problems related to the ephemeral network of overland flow paths in cities. For instance in the case of rainwater pollution effects and epidemiology of water-born diseases in urban environments. The devised approach provides insights into the connectivity of sporadic overland flow streams that may carry pollutants or pathogens to vulnerable, unexpected locations. Models of the connectivity of overland flow paths and puddles can thus support long-term retrofitting measures, preventing maintenance, and enhanced real-time response.

The use of citizen reports in this implementation shows the value of crowd sensed data. Municipalities, and other organizations in charge of urban water management, can benefit from the use of crowd sensed data. Telephonic reports made by citizens to the emergency call centers can be used as a proxy to assess the vulnerability of individual streets to urban flooding. This in turn can support the design of adaptation measures for better protection against heavy weather. Other data sources, such as Twitter, Instagram, and news media, should be better explored and used in such models.

5.2 Open spatial data partially, significantly explain flooding incidents

The second question addressed in this dissertation is to determine to which degree openly available spatial datasets, including environmental and

socioeconomic information, explain the occurrence of flood incident reports by using exploratory data analysis.

Multivariate analysis was applied to mine diverse open spatial databases. Conditions underlying the occurrence of flooding incidents are modeled in terms of available open information, including radar rainfall imagery (1 km^2 spatial resolution grid, every 5 min), cadastral (individual building, waterbodies, and green areas geometric shapes, updated once a year), and socioeconomic data (from 1 Ha to 0.25 km^2 spatial resolution grids, updated once a year). Results show that the incidence of flooding reports in urban areas is most strongly and significantly associated with imperviousness degree and proximity to outflow points into main surface waters. These associations explain up to half of the spatial variability in flooding incidents, aggregated during the event and in a sampling grid of 1 km^2 spatial resolution.

Open spatial data deliver useful, significant information for explaining the spatial distribution of flooding reports. While rainfall intensity and socioeconomic factors, such as population density and building age, were expected to influence the occurrence of flooding incidents, they did not deliver significant explaining performance in the analyses done in this thesis. This mining of spatial data also indicates that depression filling is not an important factor for urban flooding in the case of Amsterdam. This confirms results delivered by the research of chapter 2.

Results were based on data for a single rainfall event. Response variable samples provided enough robustness to support significant statements about the relationship of flooding occurrence and imperviousness, but applying partitioning machine learning techniques, such as classification trees, require larger response variable samples. Likewise, the limited amount of available response and rainfall data also restricted increasing granularity of the analysis. A larger response dataset can also allow for the use of prefiltered flooding reports, via a manual or machine-driven classification of the text information they include. This can retrieve more detail on the type of incidents being assessed, making their links to explanatory variables more clear.

This limitation on the amount of response data can be overcome by analyzing flooding incidents during longer periods, including data from several storms. Enlarging the amount of response data can also be achieved

by including additional data sources: Twitter, Facebook, and news media. However, their use implies laborious preprocessing and geolocating efforts, as these data are usually unstructured and have poor or null spatial data. On the other hand, open spatial datasets for the environmental and social explanatory variables would afford for such techniques as these data are updated every year and are normally available at every single pixels of a 1 km^2 sampling grid.

An important contribution presented in this thesis is the routine for spatially aggregating overland flowpaths and respective contributing areas for multivariate analysis. Its usefulness is proven by the significance it brings in explaining flooding reports. The routine can also be useful for future pattern recognition studies using big data in the urban environment, in which overland flow hydrology is of interest.

Modelling of urban flooding can integrate variables from open spatial data sets to enrich the description of flooding phenomena. Open spatial data, including high resolution DEMs, provide insights into patterns of urban flooding incidents, complementing urban hydraulic models that represent only some of all possible failure mechanisms leading to urban flooding. The use of open spatial data can support the devise of more realistic representations of the urban environment. They include information about the layout of components such as building age, population density, and imperviousness, which are not taken into account in flood risk analysis only based on urban drainage models.

5.3 Street imagery provides valuable information on flooding incidents

The third research question addressed in this thesis is whether mainstream image and video recording, and well-known and accessible computer vision tools, can deliver key information about localized urban flooding incidents. Street imaging data sources and computer vision analyses have the potential to be used for detecting water on the streets. Such analyses were implemented using Python and OpenCV.

After properly set, a straightforward computer vision technique automatically detected the occurrence of urban flooding in a set of street images manually downloaded from an Internet image search engine. The use of

local descriptors and an undemanding K-nearest-neighbor classification algorithm provide promising results with a sensitivity of 91%. This shows that puddle scenes can be spotted in pictures obtained in an unstructured manner, with training set made of just 23 and 29 image samples with and without a puddle, respectively.

Another mainstream computer vision technique was not able to satisfactory detect puddle extents automatically in a video-recording of a street flooding simulation, performed under controlled conditions. In this technique, the Gaussian Mixture Models had a rather limited performance. Either if attempting to detect flooded areas in a series of video frames or in a single still frame, with little or high manual preliminary input, foreground extraction based in Gaussian Mixture Models does not bring useful results. The assumptions in the modelling of backgrounds in these techniques probably suit better the classification of pixel sets corresponding to objects with bigger digital number differences, such as grass and people, than those between wet and dry street conditions.

Results obtained provide clear evidence that computer vision techniques are able to detect flooding incidents in images typically available on the Internet. The employed technique for recognizing flooding scenes in web images does not require expensive recording devices, predetermined photography parameters, nor time-consuming preliminary human intervention. This elicits the opportunity of applying this techniques for quick detection of flood locations from web information sources, for flood management and insurance purposes. The potential value of both this information resource and the explored techniques will continue to increase due to growing social and surveillance sensing.

The puddle detection success in this study is limited to a binary classification. It discriminates whether a picture describes a flooding scene or not. Further research must be performed in order to detect the actual extent of a puddle in a video or in a picture. Additionally, the size of both training and test sets in the scene classification must be enlarged preferably to the magnitude of tens of thousands of images. This can be done by crawling the web for additional material, or by exploiting installed sensing infrastructure, such as cameras in traffic surveillance systems and in networks of amateur weather stations.

Further research can compare the performance of scene recognition

based on KNN classifications used in this study with the outcome of other classifiers, such as Support Vector Machines, Bayesian, and deep neural networks.

Retrieving the extent of puddles requires additional research. The OpenCV Python libraries available for applying GMM to foreground detection are not flexible enough to allow a non-computer-scientist researcher to straightforward use them in puddle recognition problems. Further exploration of open computer vision frameworks is required to successfully extract the extent of a scene covered by a puddle.

An interesting research outlook is obtaining 3D representation of puddles directly from images. By calibrating the camera with which recordings are made, it is possible to retrieve a 3D representation of the recorded scene. Combining the area covered by a puddle with a detailed digital elevation model, an approximation of the puddle volume, and its changes across time, can be extracted.

Water managers, urban planners, and other city stake holders can embrace the use of installed infrastructure and available image stock for mining the location and timing of pluvial flooding. Retrieved information can be used for prioritizing areas with recurrent flooding in need of urgent maintenance and retrofitting measures. To enrich the availability of key flooding information for real-time response, or for posterior research, city authorities could plan citizen sensing campaigns to be triggered during heavy weather events. Particular websites, applications, or Facebook and Twitter hashtags could be used to that end, which can help on aggregating geotagged images depicting urban flooding incidents, uploaded by citizens from their smartphones.

To conclude, this thesis has explored and presented alternative concepts and technologies, which enhance the modelling of urban flooding incidents in lowland urban environments. This thesis shows that crowd sensing, open spatial data, and street imaging represent highly valuable information sources to model, explain, and predict urban pluvial flooding. Developed and implemented methods are repeatable, and can be implemented in an information infrastructure of a smart city in need of adaptation measures against extreme rainfall events.

6

References

References

Allamano, P., Croci, A., and Laio, F. Toward the camera rain gauge. *Water Resources Research*, 51(3):1744–1757, 2015. URL http://onlinelibrary.wiley.com/doi/10.1002/2014WR016298/pdf.

Apel, H., Thieken, A.H., Merz, B., and Blöschl, G. Flood risk assessment and associated uncertainty. *Natural Hazards and Earth System Science*, 4(2):295–308, 2004. ISSN 1561-8633.

Apel, H., Aronica, G. T., Kreibich, H., and Thieken, A. H. Flood risk analyses—how detailed do we need to be? *Natural Hazards*, 49(1):79–98, April 2009. ISSN 0921-030X, 1573-0840. doi: 10.1007/s11069-008-9277-8. URL http://link.springer.com/article/10.1007/s11069-008-9277-8.

Aronica, G., Bates, P.D., and Horritt, M.S. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. *Hydrological Processes*, 16(10):2001–2016, 2002. ISSN 0885-6087. doi: 10.1002/hyp.398.

Arthur, S., Crow, H., Pedezert, L., and Karikas, N. The holistic prioritisation of proactive sewer maintenance. *Water Science & Technology*, 59 (7):1385, April 2009. ISSN 0273-1223. doi: 10.2166/wst.2009.134. URL http://www.iwaponline.com/wst/05907/wst059071385.htm.

Ashley, R., Balmforth, D., Saul, A., and Blanskby, J. Flooding in the future predicting climate change, risks and responses in urban areas. *Water Science & Technology*, 52(5):265–273, 2005. URL http://www.iwaponline.com/wst/05205/wst052050265.htm.

Attema, Jisk, Bakker, Alexander, Beersma, Jules, Bessembinder, Janette, Boers, Reinout, Brandsma, Theo, Brink, Henk van den, Drijfhout, Sybren, Eskes, Henk, Haarsma, Rein, and others. KNMI'14: Climate Change scenarios for the 21st Century–A Netherlands perspective. Scientific Report WR 2014-01, KNMI, De Bilt, The Netherlands, 2014. URL http://www.klimaatscenarios.nl/brochures/images/KNMI_WR_2014-01_version26May2014.pdf.

Bach, Peter M., Rauch, Wolfgang, Mikkelsen, Peter S., McCarthy, David T., and Deletic, Ana. A critical review of integrated urban water modelling – Urban drainage and beyond. *Environmental Modelling & Software*, 54:88–107, April 2014. ISSN 13648152. doi: 10.1016/j.envsoft.2013.12.018. URL http://linkinghub.elsevier.com/retrieve/pii/S1364815213003216.

Bates, Bryson C, Kundzewicz, Zbigniew, Palutikof, Jean, Shaohong, Wu, and World Meteorological Organisation (WMO), United Nations Environment Programme (UNEP), Intergovernmental Panel on Climate Change. *Climate change and water [Electronic resource]: IPCC Technical paper VI.* IPCC Secretariat, Geneva, 2008. ISBN 978-92-9169-123-4 92-9169-123-2.

Bellos, Vasilis and Tsakiris, George. Comparing Various Methods of Building Representation for 2d Flood Modelling In Built-Up Areas. *Water Resources Management*, pages 1–19, June 2014. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-014-0702-3. URL http://link.springer.com/article/10.1007/s11269-014-0702-3.

Berggren, Karolina, Olofsson, Mats, Viklander, Maria, Svensson, Gilbert, and Gustafsson, Anna-Maria. Hydraulic Impacts on Urban Drainage Systems due to Changes in Rainfall Caused by Climatic Change. *Journal of Hydrologic Engineering*, 17(1):92–98, January 2012. ISSN 1084-0699, 1943-5584. doi: 10.1061/(ASCE)HE.1943-5584.0000406. URL http://ascelibrary.org/doi/abs/10.1061/%28ASCE%29HE.1943-5584.0000406.

Blanc, J., Hall, J.w., Roche, N., Dawson, R.j., Cesses, Y., Burton, A., and Kilsby, C.g. Enhanced efficiency of pluvial flood risk estimation in urban areas using spatial–temporal rainfall simulations. *Journal of Flood Risk Management*, 5(2):143–152, 2012. ISSN 1753-318X. doi: 10.1111/j.1753-318X.2012.01135.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1753-318X.2012.01135.x/abstract.

Booij, M.J. Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, 303(1-4):176–198, 2005. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2004.07.013.

Botzen, W. J. W., Aerts, J., and Bergh, J.C.J.M. van den. Willingness of homeowners to mitigate climate risk through insurance. *Ecological*

Economics, 68(8):2265–2277, 2009. URL http://www.sciencedirect.com/science/article/pii/S092180090900072X.

Box, George EP and Cox, David R. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964. URL http://www.jstor.org/stable/2984418.

Braak, C. J. F. Ter and Looman, C.W.N. Chapter 3 - Regression. In Jongman, R. H. G., Braak, C. J. F. Ter, and Tongeren, O. F. R. van, editors, *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, March 1995. ISBN 978-0-521-47574-7.

Braak, C. J. F. ter. Chapter 5 - Ordination. In Jongman, R. H. G., Braak, C. J. F. ter, and Tongeren, O. F. R. van, editors, *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, March 1995. ISBN 978-0-521-47574-7.

Bradski, Gary and Kaehler, Adrian. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008. URL https://books.google.nl/books?hl=en&lr=&id=seAgiOfu2EIC&oi=fnd&pg=PR3&dq=Learning+OpenCV&ots=hTI99ibFOi&sig=dkSVMdTGdFHAUjhPHx2z4oRobjU.

Caradot, N., Granger, D., Rostaing, C., Cherqui, F., Chocat, B., et al. L'évaluation du risque de débordement des systèmes de gestion des eaux urbaines: contributions méthodologiques de deux cas d'études (Lyon et Mulhouse). In *7ème Conférence internationale sur les techniques et stratégies durables pour la gestion des eaux urbaines par temps de pluie*, 2010.

Center for International Earth Science Information Network - CIESIN - Columbia University. Low Elevation Coastal Zone (LECZ) Urban-Rural Population and Land Area Estimates, Version 2, 2013. URL https://doi.org/10.7927/H4MW2F2J.

Centraal Bureau voor de Statistiek. Statistische gegevens per vierkant - Statistischegegevenspervierkantupdate-juli2013.pdf, 2013. URL http://www.cbs.nl/NR/rdonlyres/661D884F-CF5B-4192-8138-EA959D540EFE/0/Statistischegegevenspervierkantupdatejuli2013.pdf.

Cherqui, Frédéric, Belmeziti, Ali, Granger, Damien, Sourdril, Antoine, and Le Gauffre, Pascal. Assessing urban potential flooding risk and identifying effective risk-reduction measures. *Science of The Total Environment*, 514:418–425, May 2015. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2015.02.027. URL http://www.sciencedirect.com/science/article/pii/S0048969715001631.

Collete, Andrew. H5py, 2015. URL https://pypi.python.org/pypi/h5py.

Deletic, A., Dotto, C.B.S., McCarthy, D.T., Kleidorfer, M., Freni, G., Mannina, G., Uhl, M., Henrichs, M., Fletcher, T.D., Rauch, W., Bertrand-Krajewski, J.L., and Tait, S. Assessing uncertainties in urban drainage models. *Physics and Chemistry of the Earth*, 42-44:3–10, 2012. ISSN 1474-7065. doi: 10.1016/j.pce.2011.04.007.

Diaz-Nieto, J., Lerner, D.N., Saul, A.J., and Blanksby, J. GIS Water-Balance Approach to Support Surface Water Flood-Risk Management. *Journal of Hydrologic Engineering*, 17(1):55–67, 2011.

Dongquan, Zhao, Jining, Chen, Haozheng, Wang, Qingyuan, Tong, Shangbing, Cao, and Zheng, Sheng. GIS-based urban rainfall-runoff modeling using an automatic catchment-discretization approach: a case study in Macau. *Environmental Earth Sciences*, 59(2):465–472, January 2009. ISSN 1866-6280, 1866-6299. doi: 10.1007/s12665-009-0045-1. URL http://www.springerlink.com/index/10.1007/s12665-009-0045-1.

Dotto, C. B. S., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D. T., and Fletcher, T. D. Performance and sensitivity analysis of stormwater models using a Bayesian approach and long-term high resolution data. *Environmental Modelling & Software*, 26(10):1225–1239, October 2011. ISSN 1364-8152. doi: 10.1016/j.envsoft.2011.03.013. URL http://www.sciencedirect.com/science/article/pii/S1364815211000880.

Dotto, C.B.S., Mannina, G., Kleidorfer, M., Vezzaro, L., Henrichs, M., McCarthy, D.T., Freni, G., Rauch, W., and Deletic, A. Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling. *Water Research*, 46(8):2545–2558, 2012. ISSN 0043-1354. doi: 10.1016/j.watres.2012.02.009.

Douglas, I., Garvin, S., Lawson, N., Richards, J., Tippett, J., and White, I. Urban pluvial flooding: a qualitative case study of cause, effect and nonstructural mitigation. *Journal of Flood Risk Management*, 3(2):112–125, 2010. ISSN 1753-318X. doi: 10.1111/j.1753-318X.2010.01061.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1753-318X.2010.01061.x/abstract.

Dutch Ministry of Interior and Kingdom Relations. Open Data NEXT in English | Data.overheid.nl: het opendataportaal van de Nederlandse overheid, 2014. URL https://data.overheid.nl/english.

Dutch Ministry of Security and Justice. Camera surveillance in the Netherlands | Research and Documentation Centre (WODC), April 2014. URL https://english.wodc.nl/onderzoeksdatabase/2372-caneratoezicht-slimmer-bekeken.aspx?cp=45&cs=6802.

Elmore, Kimberly L., Flamig, Z. L., Lakshmanan, V., Kaney, B. T., Farmer, V., Reeves, Heather D., and Rothfusz, Lans P. MPING: Crowd-Sourcing Weather Reports for Research. *Bulletin of the American Meteorological Society*, 95(9):1335–1342, January 2014. ISSN 0003-0007. doi: 10.1175/BAMS-D-13-00014.1. URL http://journals.ametsoc.org/doi/full/10.1175/BAMS-D-13-00014.1.

Ericson, Jason P., Vörösmarty, Charles J., Dingman, S. Lawrence, Ward, Larry G., and Meybeck, Michel. Effective sea-level rise and deltas: Causes of change and human dimension implications. *Global and Planetary Change*, 50(1–2):63–82, February 2006. ISSN 0921-8181. doi: 10.1016/j.gloplacha.2005.07.004. URL http://www.sciencedirect.com/science/article/pii/S0921818105001827.

ESRI. ArcGIS, 2012. URL http://www.esri.com/software/arcgis/arcgis10/.

European Central Bank. Determination of the euro conversion rates, 1998. URL http://www.ecb.europa.eu/press/pr/date/1998/html/pr981231_2.en.html.

Fontanazza, C.M., Freni, G., La Loggia, G., and Notaro, V. Uncertainty evaluation of design rainfall for urban flood risk analysis. *Water Science and Technology*, 63(11):2641–2650, 2011.

Fontanazza, C.M., Freni, G., and Notaro, V. Bayesian inference analysis of the uncertainty linked to the evaluation of potential flood damage in urban areas. *Water Science and Technology*, 66(8):1669–1677, 2012.

Freni, G., La Loggia, G., and Notaro, V. Uncertainty in urban flood damage assessment due to urban drainage modelling and depth-damage curve estimation. *Water Science & Technology*, 61(12):2979, June 2010. ISSN 0273-1223. doi: 10.2166/wst.2010.177. URL http://www.iwaponline.com/wst/06112/wst061122979.htm.

Gaitan, S. Color profiles of wet street objects - YouTube, 2016. URL https://www.youtube.com/watch?v=hRf-SpdYy9U.

Gaitan, S. and Veldhuis, J.A.E. ten. Opportunities for multivariate analysis of open spatial datasets to characterize urban flooding risks. *Proceedings of the International Association of Hydrological Sciences*, 370:9–14, June 2015. ISSN 2199-899X. doi: 10.5194/piahs-370-9-2015. URL http://www.proc-iahs.net/370/9/2015/.

Gaitan, S., Veldhuis, J.A.E. ten, Spekkers, M.H., and Giesen, N. C. van de. Urban vulnerability to pluvial flooding: complaints location on overland flow routes. In *Proceedings of the 2nd European Conference on Flood Risk Management FLOODrisk2012, Rotterdam, The Netherlands, 19-23 November 2012*, pages 338–339, The Netherlands, 2012. CRC Press.

Gaitan, S., Calderoni, L., Palmieri, P., Veldhuis, J. A. E. ten, Maio, D., and Riemsdijk, M.B. van. From Sensing to Action: Quick and Reliable Access to Information in Cities Vulnerable to Heavy Rain. *IEEE Sensors Journal*, 14(12):4175–4184, December 2014. ISSN 1530-437X. doi: 10.1109/JSEN.2014.2354980.

Gaitan, S., Veldhuis, J.A.E. ten, and Giesen, N. van de. Spatial Distribution of Flood Incidents Along Urban Overland Flow-Paths. *Water Resources Management*, pages 1–13, May 2015. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-015-1006-y. URL http://link.springer.com/article/10.1007/s11269-015-1006-y.

Gillies, Sean. Shapely, 2013. URL https://pypi.python.org/pypi/Shapely.

Gillies, Sean. Fiona, 2014. URL https://pypi.python.org/pypi/Fiona/.

Gillies, Sean, Butler, Howard, and Pedersen, Brend. RTree, 2014. URL https://pypi.python.org/pypi/Rtree/.

Hamilton, Serena H., ElSawah, Sondoss, Guillaume, Joseph H. A., Jakeman, Anthony J., and Pierce, Suzanne A. Integrated assessment and modelling: Overview and synthesis of salient dimensions. *Environmental Modelling & Software*, 64:215–229, February 2015. ISSN 1364-8152. doi: 10.1016/j.envsoft.2014.12.005. URL http://www.sciencedirect.com/science/article/pii/S1364815214003600.

Het Parool. Veel wateroverlast in Amsterdam. *Het Parool*, July 2014. URL http://www.parool.nl/parool/nl/4/AMSTERDAM/article/detail/3702982/2014/07/28/Veel-wateroverlast-in-Amsterdam.dhtml.

Hoes, O. and Schuurmans, W. Flood standards or risk analyses for polder management in the Netherlands. *Irrigation and Drainage*, 55(SUPPL. 1):S113–S119, 2006. ISSN 1531-0353. doi: 10.1002/ird.249.

Horita, Y., Kawai, S., Furukane, T., and Shibata, K. Efficient distinction of road surface conditions using surveillance camera images in night time. In *2012 19th IEEE International Conference on Image Processing (ICIP)*, pages 485–488, September 2012. doi: 10.1109/ICIP.2012.6466902.

Horritt, M.S. and Bates, P.D. Predicting floodplain inundation: Raster-based modelling versus the finite-element approach. *Hydrological Processes*, 15(5):825–842, 2001. ISSN 0885-6087. doi: 10.1002/hyp.188.

Hurk, B. van den, Klein Tank, A., Lenderink, G., Ulden, A. van, Oldenborgh, G. J. van, Katsman, C., Brink, H. van den, Keller, F., Bessembinder, J., Burgers, G., et al. *KNMI climate change scenarios 2006 for the Netherlands*. KNMI De Bilt, 2006.

Ihaka, Ross and Gentleman, Robert. R, 2015. URL https://www.r-project.org/.

Illian, Janine, Penttinen, Antti, Stoyan, Helga, and Stoyan, Dietrich. *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley, Hoboken, 1 edition, 2008. ISBN 978-0-470-72515-3.

Intel Corporation, Willow Garage, and Itseez. OpenCV: Open Source Computer Vision Library, 2014.

Jacobs, J.C.J. The Rotterdam approach: connecting water with opportunities. In Howe, Carol and Mitchell, Cynthia, editors, *Water Sensitive Cities*. IWA Publishing, 2012. ISBN 978-1-84339-364-1.

Jak, Martine and Kok, Matthijs. A Database of Historical Flood Events in the Netherlands. In Marsalek, Jiri, Watt, W. Ed, Zeman, Evzen, and Sieker, Friedhelm, editors, *Flood Issues in Contemporary Water Management*, number 71 in NATO Science Series, pages 139–146. Springer Netherlands, January 2000. ISBN 978-0-7923-6452-8 978-94-011-4140-6. URL http://link.springer.com/chapter/10.1007/978-94-011-4140-6_15. 00009.

Jenson, SK and Domingue, JO. Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric engineering and remote sensing*, 54(11):1593–1600, 1988.

Jeong, Jaehak, Kannan, Narayanan, Arnold, Jeff, Glick, Roger, Gosselink, Leila, and Srinivasan, Raghavan. Development and Integration of Subhourly Rainfall–Runoff Modeling Capability Within a Watershed Model. *Water Resources Management*, 24(15):4505–4527, December 2010. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-010-9670-4. URL http://link.springer.com/10.1007/s11269-010-9670-4.

Jonkman, S.N., Bočkarjova, M., Kok, M., and Bernardini, P. Integrated hydrodynamic and economic modelling of flood damage in the Netherlands. *Ecological Economics*, 66(1):77–90, 2008a. ISSN 0921-8009. doi: 10.1016/j.ecolecon.2007.12.022.

Jonkman, S.N., Kok, M., and Vrijling, J.K. Flood risk assessment in the Netherlands: A case study for dike ring South Holland. *Risk Analysis*, 28(5):1357–1373, 2008b. ISSN 0272-4332. doi: 10.1111/j.1539-6924.2008.01103.x.

Kadaster Nederland. BAG, 2013. URL https://www.kadaster.nl/bag.

KaewTraKulPong, Pakorn and Bowden, Richard. An improved adaptive background mixture model for real-time tracking with shadow

detection. In *Video-based surveillance systems*, pages 135–144. Springer, 2002. URL http://link.springer.com/chapter/10.1007/978-1-4615-0913-4_11.

Kenya National Government. Surveillance cameras to help police fight crime, May 2015. URL http://www.mygov.go.ke/?p=2360.

Knebl, M.R., Yang, Z.-L., Hutchison, K., and Maidment, D.R. Regional scale flood modeling using NEXRAD rainfall, GIS, and HEC-HMS/ RAS: A case study for the San Antonio River Basin Summer 2002 storm event. *Journal of Environmental Management*, 75(4 SPEC. ISS.):325–336, 2005. ISSN 0301-4797. doi: 10.1016/j.jenvman.2004.11.024.

Kok, J.-L., Kofalk, S., Berlekamp, J., Hahn, B., and Wind, H. From design to application of a decision-support system for integrated river-basin management. *Water Resources Management*, 23(9):1781–1811, 2009. ISSN 0920-4741. doi: 10.1007/s11269-008-9352-7.

Kunapo, Joshphar, Chandra, Shobhit, and Peterson, Jim. Drainage Network Modelling for Water-Sensitive Urban Design. *Transactions in GIS*, 13(2):167–178, April 2009. ISSN 13611682, 14679671. doi: 10.1111/j.1467-9671.2009.01146.x. URL http://doi.wiley.com/10.1111/j.1467-9671.2009.01146.x.

Legendre, Pierre and Legendre, Louis. Chapter 8 - Cluster analysis. In Legendre, Pierre Legendre and Louis, editor, *Developments in Environmental Modelling*, volume 24 of *Numerical Ecology*, pages 337–424. Elsevier, 2012a. URL http://www.sciencedirect.com/science/article/pii/B9780444538680500083.

Legendre, Pierre and Legendre, Louis. Chapter 9 - Ordination in reduced space. In Legendre, Pierre and Legendre, Louis, editors, *Developments in Environmental Modelling*, volume 24 of *Numerical Ecology*, pages 425–520. Elsevier, 2012b. URL http://www.sciencedirect.com/science/article/pii/B9780444538680500095.

Lo, Shi-Wei, Wu, Jyh-Horng, Lin, Fang-Pang, and Hsu, Ching-Han. Visual Sensing for Urban Flood Monitoring. *Sensors*, 15(8):20006–20029, August 2015. doi: 10.3390/s150820006. URL http://www.mdpi.com/1424-8220/15/8/20006.

Lowe, David G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. ISSN 0920-5691, 1573-1405. doi: 10.1023/B:VISI.0000029664.99615.94.

Lundh, Fredrik. Python Imaging Library (PIL), 1995. URL http://www.pythonware.com/products/pil/.

Maaskant, B., Jonkman, S.N., and Bouwer, L.M. Future risk of flooding: an analysis of changes in potential loss of life in South Holland (The Netherlands). *Environmental Science and Policy*, 12(2):157–169, 2009. ISSN 1462-9011. doi: 10.1016/j.envsci.2008.11.004.

Maksimović, Čedo, Prodanović, Dušan, Boonya-Aroonnet, Surajate, Leitão, João P., Djordjević, Slobodan, and Allitt, Richard. Overland flow and pathway analysis for modelling of urban pluvial flooding. *Journal of Hydraulic Research*, 47(4):512–523, July 2009. ISSN 0022-1686, 1814-2079. doi: 10.1080/00221686.2009.9522027. URL http://www.tandfonline.com/doi/abs/10.1080/00221686.2009.9522027.

Man, H. de, Berg, H.H.J.L. van den, Leenen, E.J.T.M., Schijven, J.F., Schets, F.M., Vliet, J.C. van der, Knapen, F. van, and Roda Husman, A.M. de. Quantitative assessment of infection risk from exposure to waterborne pathogens in urban floodwater. *Water Research*, 48:90–99, January 2014. ISSN 00431354. doi: 10.1016/j.watres.2013.09.022. URL http://linkinghub.elsevier.com/retrieve/pii/S0043135413007148.

Matthies, Larry H., Bellutta, Paolo, and McHenry, Mike. Detecting water hazards for autonomous off-road navigation. In *AeroSense 2003*, pages 231–242. International Society for Optics and Photonics, 2003. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=763483.

McKinney, Wes. Pandas, 2015. URL https://pypi.python.org/pypi/pandas.

Melo, Nuno, Santos, Bruno Filipe, and Leandro, Jorge. A prototype tool for dynamic pluvial-flood emergency planning. *Urban Water Journal*, 12(1):79–88, January 2015. ISSN 1573-062X, 1744-9006. doi: 10.1080/1573062X.2014.975725. URL http://www.tandfonline.com/doi/abs/10.1080/1573062X.2014.975725.

Merz, B., Kreibich, H., Thieken, A., and Schmidtke, R. Estimation uncertainty of direct monetary flood damage to buildings. *Natural Hazards and Earth System Science*, 4(1):153–163, 2004. ISSN 1561-8633.

Merz, B., Kreibich, H., and Lall, U. Multi-variate flood damage assessment: A tree-based data-mining approach. *Natural Hazards and Earth System Science*, 13(1):53–64, 2013. ISSN 15618633. doi: 10.5194/nhess-13-53-2013.

Michelsen, N., Dirks, H., Schulz, S., Kempe, S., Al-Saud, M., and Schüth, C. YouTube as a crowd-generated water level archive. *Science of The Total Environment*, 568:189–195, October 2016. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2016.05.211. URL https://www.sciencedirect.com/science/article/pii/S0048969716311482.

Montgomery, Douglas C and Runger, George C. *Applied statistics and probability for engineers*. Wiley, New York, 2003. ISBN 0-471-20454-4 978-0-471-20454-1 0-471-38181-0 978-0-471-38181-5.

Muller, Catherine L. Mapping snow depth across the West Midlands using social media-generated data. *Weather*, 68(3):82–82, March 2013. ISSN 1477-8696. doi: 10.1002/wea.2103. URL http://onlinelibrary.wiley.com/doi/10.1002/wea.2103/abstract.

Muller, C.l., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., Overeem, A., and Leigh, R.r. Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, 35(11):3185–3203, September 2015. ISSN 1097-0088. doi: 10.1002/joc.4210. URL http://onlinelibrary.wiley.com.tudelft.idm.oclc.org/doi/10.1002/joc.4210/abstract.

Murphy, James, Hadley Centre for Climate Prediction and Research, and UK Climate Impacts Programme. *Climate change projections*. Met Office Hadley Centre, Exeter, 2009. ISBN 978-1-906360-02-3 1-906360-02-2. URL http://ukclimateprojections.defra.gov.uk/content/view/824/517/index.html.

Neal, J., G. Schumann, T. Fewtrell, M. Budimir, P. Bates, and D. Mason. Evaluating a new LISFLOOD-FP formulation with data fromthe summer

2007 floods in Tewkesbury, UK. *Journal of Flood Risk Management*, 4: 88–95, 2011. doi: 10.1111/j.1753-318X.2011.01093.x.

Netherlands Royal Meteorological Institute. KNMI Product Catalogus, 2013. URL http://www.knmi.nl/datacentrum/catalogus/catalogus/catalogus-gegevens-overzicht.html.

Netherlands Royal Meteorological Institute. Hoe vaak komt extreme neerslag zoals op 28 juli tegenwoordig voor, en is dat anders dan vroeger?, August 2014. URL http://www.knmi.nl/cms/content/120817/hoe_vaak_komt_extreme_neerslag_zoals_op_28_juli_tegenwoordig_voor_en_is_dat_anders_dan_vroeger.

Ochoa-Rodriguez, Susana, Wang, Li-Pen, Gires, Auguste, Pina, Rui Daniel, Reinoso-Rondinel, Ricardo, Bruni, Guendalina, Ichiba, Abdellah, Gaitan, Santiago, Cristiano, Elena, Assel, Johan van, Kroll, Stefan, Murlà-Tuyls, Damian, Tisserand, Bruno, Schertzer, Daniel, Tchiguirinskaia, Ioulia, Onof, Christian, Willems, Patrick, and Veldhuis, Marie-Claire ten. Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation. *Journal of Hydrology*, 531:389–407, December 2015. ISSN 00221694. doi: 10.1016/j.jhydrol.2015.05.035. URL http://linkinghub.elsevier.com/retrieve/pii/S0022169415003856.

Okabe, Atsuyuki and Sugihara, Kokichi. *Spatial Analysis Along Networks : Statistical and Computational Methods*. Wiley, Hoboken, 1 edition, 2012. ISBN 978-1-119-96709-5.

Oksanen, Jari, Blanchet, F. Guillaume, Kindt, Roeland, Legendre, Pierre, Minchin, Peter R., O'Hara, R. B., Simpson, Gavin L., Solymos, Peter, Stevens, M. Henry H., and Wagner, Helene. vegan: Community Ecology Package, 2015. URL https://cran.r-project.org/web/packages/vegan/.

Olivera, F. and Maidment, D. Geographic information systems(GIS)-based spatially distributed model for runoff routing. *Water Resources Research*, 35(4):1155–1164, 1999.

O'Sullivan, David and Unwin, David John. *Geographic information analysis*. John Wiley & Sons, 2nd edition, 2010.

Overeem, A., Buishand, T. A., and Holleman, I. Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar. *Water Resources Research*, 45(10):W10424, October 2009a. ISSN 1944-7973. doi: 10.1029/2009WR007869. URL http://onlinelibrary.wiley.com/doi/10.1029/2009WR007869/abstract.

Overeem, A., Holleman, I., and Buishand, A. Derivation of a 10-year radar-based climatology of rainfall. *Journal of Applied Meteorology and Climatology*, 48(7):1448–1463, 2009b. ISSN 1558-8424. doi: 10.1175/2009JAMC1954.1.

Pistrika, Aimilia, Tsakiris, George, and Nalbantis, Ioannis. Flood Depth-Damage Functions for Built Environment. *Environmental Processes*, 1(4):553–572, December 2014. ISSN 2198-7491, 2198-7505. doi: 10.1007/s40710-014-0038-2. URL http://link.springer.com/article/10.1007/s40710-014-0038-2.

Pistrika, A.K. and Jonkman, S.N. Damage to residential buildings due to flooding of New Orleans after hurricane Katrina. *Natural Hazards*, 54(2):413–434, 2010. ISSN 0921-030X. doi: 10.1007/s11069-009-9476-y.

Publieke Dienstverlening Op de Kaart Loket. BAG Geocodeerservice, 2013. URL https://www.pdok.nl/nl/service/openls-bag-geocodeerservice. 00000.

Publieke Dienstverlening Op de Kaart Loket. Actueel Hoogtebestand Nederland voor iedereen vrij toegankelijk | Publieke Dienstverlening Op de Kaart Loket, 2014. URL https://www.pdok.nl/nl/actueel/nieuws/artikel/06mrt14-actueel-hoogtebestand-nederland-voor-iedereen-vrij-toegankelijk.

Publieke Dienstverlening Op de Kaart Loket. TOP10nl, 2015. URL https://www.pdok.nl/nl/producten/pdok-downloads/basis-registratie-topografie/topnl/topnl-actueel/top10nl.

Python Software Foundation. Python Programming Language, 2014. URL python.org.

QGIS Development Team. QGIS, 2014. URL qgis.osgeo.org.

Ramette, Alban. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, 62(2):142–160, November 2007. ISSN 1574-6941. doi: 10.1111/j.1574-6941.2007.00375.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1574-6941.2007.00375.x/abstract.

Ravazzani, Giovanni, Gianoli, Paride, Meucci, Stefania, and Mancini, Marco. Assessing Downstream Impacts of Detention Basins in Urbanized River Basins Using a Distributed Hydrological Model. *Water Resources Management*, 28(4):1033–1044, March 2014. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-014-0532-3. URL http://link.springer.com/10.1007/s11269-014-0532-3.

RIONED Foundation. *Leidraad Riolering, Module C2100*. Stichting RIONED, Ede, The Netherlands, 2004. ISBN SBN 978-90-73645-68-4.

Romero, Yanina L., Bessembinder, J., Giesen, N.C. van de, and Ven, F. H. M. van de. A relation between extreme daily precipitation and extreme short term precipitation. *Climatic Change*, 106(3):393–405, 2011. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-010-9955-x. URL http://www.springerlink.com/index/10.1007/s10584-010-9955-x.

Roozekrans, Hans and Holleman, Iwan. KNMI HDF5 Data Format Specification, v3.5. Technical report, Netherlands Royal Meteorological Institute, 2003. URL http://www.knmi.nl/kennis-en-datacentrum/publicatie/knmi-hdf5-data-format-specification-v3-5.

Rother, Carsten, Kolmogorov, Vladimir, and Blake, Andrew. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004. URL http://dl.acm.org/citation.cfm?id=1015720.

Sales-Ortells, Helena, Agostini, Giulia, and Medema, Gertjan. Quantification of Waterborne Pathogens and Associated Health Risks in Urban Water. *Environmental Science & Technology*, 49(11):6943–6952, June 2015. ISSN 0013-936X, 1520-5851. doi: 10.1021/acs.est.5b00625. URL http://pubs.acs.org/doi/abs/10.1021/acs.est.5b00625.

Sarwal, Alok, Nett, Jeremy, and Simon, David. Detection of Small Water-Bodies. Technical report, PERCEPTEK INC LITTLETON CO, December 2004.

Shibata, Keiji, Takeuch, Kazuya, Kawai, Shohei, and Horita, Yuukou. Detection of Road Surface Conditions in Winter using Road Surveillance Cameras at Daytime, Night-time and Twilight. *International Journal of Computer Science and Network Security (IJCSNS)*, 14(11):21, 2014. URL http://paper.ijcsns.org/07_book/201411/20141104.pdf.

Sinclair, Dennis F. On Tests of Spatial Randomness Using Mean Nearest Neighbor Distance. *Ecology*, 66(3):1084–1085, June 1985. ISSN 0012-9658. doi: 10.2307/1940568. URL http://www.jstor.org/stable/1940568.

Solem, Jan Erik. *Programming Computer Vision with Python: Tools and algorithms for analyzing images*. " O'Reilly Media, Inc.", first edition edition, 2012. ISBN 978-1-4493-1654-9.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Veldhuis, J. A. E. ten. A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences Discussions*, 9(10):11615–11640, October 2012. ISSN 1812-2116. doi: 10.5194/hessd-9-11615-2012. URL http://www.hydrol-earth-syst-sci-discuss.net/9/11615/2012/hessd-9-11615-2012-discussion.html.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Veldhuis, J.A.E. ten. A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences*, 17(3):913–922, March 2013. ISSN 1607-7938. doi: 10.5194/hess-17-913-2013. URL http://www.hydrol-earth-syst-sci.net/17/913/2013/.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Veldhuis, J. A. E. ten. Decision-tree analysis of factors influencing rainfall-related building structure and content damage. *Natural Hazards and Earth System Science*, 14(9):2531–2547, September 2014. ISSN 1684-9981. doi: 10.5194/nhess-14-2531-2014. URL http://www.nat-hazards-earth-syst-sci.net/14/2531/2014/.

Szeliski, Richard. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

Tarboton, D.G., Bras, R.L., and Rodriguez-Iturbe, I. On the extraction of channel networks from digital elevation data. *Hydrological Processes*, 5 (1):81–100, 1991.

Tongeren, O. F. R. van. Chapter 6 - Cluster analysis. In Jongman, R. H. G., Braak, C. J. F. Ter, and Tongeren, O. F. R. van, editors, *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, March 1995. ISBN 978-0-521-47574-7.

Tsakiris, G. Flood risk assessment: concepts, modelling, applications. *Nat. Hazards Earth Syst. Sci.*, 14(5):1361–1369, May 2014. ISSN 1684-9981. doi: 10.5194/nhess-14-1361-2014. URL http://www.nat-hazards-earth-syst-sci.net/14/1361/2014/.

Tsakiris, George and Bellos, Vasilis. A Numerical Model for Two-Dimensional Flood Routing in Complex Terrains. *Water Resources Management*, 28(5):1277–1291, March 2014. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-014-0540-3. URL http://link.springer.com/article/10.1007/s11269-014-0540-3.

Veldhuis, J.A.E. ten. How the choice of flood damage metrics influences urban flood risk assessment: Urban flood risk assessment. *Journal of Flood Risk Management*, 4(4):281–287, December 2011. ISSN 1753318X. doi: 10.1111/j.1753-318X.2011.01112.x. URL http://doi.wiley.com/10.1111/j.1753-318X.2011.01112.x.

Veldhuis, J.A.E. ten and Clemens, FHLR. Uncertainty in risk analysis of urban pluvial flooding: a case study. *Water Practice & Technology*, 4(1), 2009. URL http://www.iwaponline.com/wpt/004/wpt0040018.htm.

Veldhuis, J.A.E. ten and Clemens, F.H.L.R. Flood risk modelling based on tangible and intangible urban flood damage quantification. *Water Science & Technology*, 62(1):189, July 2010. ISSN 0273-1223. doi: 10.2166/wst.2010.243. URL http://www.iwaponline.com/wst/06201/wst062010189.htm.

Veldhuis, J.A.E. ten and Clemens, F.H.L.R. The efficiency of asset management strategies to reduce urban flood risk. *Water Science & Technology*, 64(6):1317, September 2011. ISSN 0273-1223. doi: 10.2166/wst.2011.715. URL http://www.iwaponline.com/wst/06406/wst064061317.htm.

Veldhuis, J.A.E. ten, Clemens, F.H.L.R., Sterk, G., and Berends, B.R. Microbial risks associated with exposure to pathogens in contaminated urban flood water. *Water Research*, 44(9):2910–2918, May 2010. ISSN

00431354. doi: 10.1016/j.watres.2010.02.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0043135410000989.

Veldhuis, J.A.E. ten, Clemens, François H.L.R., and Gelder, Pieter H.A.J.M. van. Quantitative fault tree analysis for urban water infrastructure flooding. *Structure and Infrastructure Engineering*, 7(11):809–821, 2011. ISSN 1573-2479. doi: 10.1080/15732470902985876. URL http://www.tandfonline.com/doi/abs/10.1080/15732470902985876.

Vitolo, Claudia, Elkhatib, Yehia, Reusser, Dominik, Macleod, Christopher J. A., and Buytaert, Wouter. Web technologies for environmental Big Data. *Environmental Modelling & Software*, 63:185–198, January 2015. ISSN 1364-8152. doi: 10.1016/j.envsoft.2014.10.007. URL http://www.sciencedirect.com/science/article/pii/S1364815214002965.

Wade, T.J., Lin, C.J., Jagai, J.S., and Hilborn, E.D. Flooding and emergency room visits for gastrointestinal illness in Massachusetts: A case-crossover study. *PLoS ONE*, 9(10), 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0110474.

Waternet. Extreme wateroverlast in Amsterdam? Ook deze zomer een reëel scenario., 2015. URL https://www.waternet.nl/actueel/nieuwsberichten/2015/extreme-wateroverlast-in-amsterdam-ook-deze-zomer-een-reeel-scenario/.

Whitaker, Jeff. pyproj, 2014. URL https://pypi.python.org/pypi/pyproj/.

Wong, T. H. F. and Brown, R. R. The water sensitive city: principles for practice. *Water Science & Technology*, 60(3):673, July 2009. ISSN 0273-1223. doi: 10.2166/wst.2009.436. URL http://www.iwaponline.com/wst/06003/wst060030673.htm.

Zon, N. van der. Kwaliteitsdocument AHN-2. Technical Report 1.1, Rijkswaterstaat & Waterschappen, 2011.

Acknowledgements

First I want to express sincere and deepest gratitude to Nick van de Giesen. His vision, approach, human quality, and sense of fairness were inspiring and crucial for the completion of my PhD, for understanding what becoming a PhD means. His challenging feedback and contributions were cardinal during the adventure of accomplishing this doctorate. I also kindly thank Marie-claire ten Veldhuis, for her trust, dedication, and prompt feedback throughout my doctorate. Her opportune contributions effectively promoted the lucidity and cogency of yielded scientific output. I hope I can share with you both further knowledge-diving adventures.

This dissertation was possible thanks to the funding and support provided by Climate-KIC, TU Delft, and IBM's Global PhD Fellowship award. Also, I thank the collaboration and data provided by the Municipality of Rotterdam, WaterNet, Netherlands Royal Meteorological Institute, Netherlands Central Bureau for Statistics, and the Netherlands Cadaster. My gratitude also goes to Wing Yan Man, Robert Jan Sips, Manfred Overmeen, and Bram Havers, for making us aware of the pertinence of our work in the context of IBM's Global PhD Fellowship award, for their support, and for hosting me during my time as a research fellow.

I also want to thank my TU Delft colleagues: Anna Solcerova, Cesar Jimenez, Moshiur Rahman, Sandra Junier, Ronald van Nooyen, Luz Ton-Estrada, Betty Rothfusz, Petra Jorritsma, Lydia de Hoog, Matthieu Spekkers, Guenda Bruni, Elena Cristiano, Birna van Riemsdijk, Paolo Palmieri, Luca Calderoni, Maaike Beliën, Soren Johnson, Wim Luxemburg, Maurits Erstsen, Martijn Koole, Michelle Loozen, Jelmar Schellingerhout, Ken Arroyo Ohori, Martijn Meijers, Stijn de Jong, Rolf Hut, Genadii Donchyts, Leonardo Alfonso, and Martin Bloemendal. Thank you all for the enriching conversations, coaching, support, and knowledge exchange.

Finally, I want to acknowledge the essential support of my family: my wife, my parents, and my sister. This thesis has been achieved thanks to you.

Curriculum Vitæ

Santiago GAITAN SABOGAL

Born in 11 June 1982, in Planet Earth, Solar System.

Education

2007–2007 **Specialist in Remote Sensing and Geographic Information Systems**
Brazil *National Institute for Space Research*

2001–2006 **Biologist Diploma (MSc equivalent)**
Colombia *National University of Colombia*

Experience (selected)

2015–2016 **Research fellow - IBM PhD fellowship award**
The Netherlands *IBM*
Development of open spatial data analytics in smart urban operations. Exploring the potential of crowd sourced images and surveillance cameras for smart water management.

2012–2016 **PhD researcher**
The Netherlands *Delft University of Technology and Climate KIC*
Modeling of urban pluvial flooding using heterogeneous sources of spatial information.

2009–2011 **Research intern**
Argentina *Cartog. & Land Surv., Water Res. Departments, Nat. Univ. of the Littoral*
Remote sensing and GIS for water management.

2007–2007 **Graduate trainee**
Brazil *National Institute for Space Research*
Remote Sensing, multi-temporal analysis, and prediction model of changes on vegetation cover and land use in northwestern Colombian Amazon.

2006–2006 **Research intern**
Amazonas, Colom. *Alexander von Humboldt Biological Resources Research Institute*
Floristic comparison between two transects on dryland forests in Amacayacu National Park: A Contribution to the Understanding of the Relationship Between the Biosphere and Atmosphere in the Amazonian Forests.

2003–2010 **Co-founder**
Bogota, Colombia *NGO Mimesis*
Consultancy projects for different organizations: Planetarium of Bogota, Institute for Culture and Tourism, Botanical Garden, Secretary of Culture and Secretary of Education of Bogotá City Hall; Ministry of Education, Colombia; UNESCO, United Nations, Universe Awareness for Children.

1999–1999 **Elected Territorial Counselor.**
Chia, Colombia *Municipal Land Planning Counsel*
Land Use Participatory Planning. Representative of youth and schools.

Distinctions

2015 Recipient of the IBM PhD Fellowship Award for exceptional students.

2012 PhD Fellowship. European Institute of Innovation and Technology - Climate Knowledge and Innovation Communities.

2007 United Nations University Grant for Graduate Studies on Remote Sensing an GIS applied on Natural Resources.

List of Publications

9. **Gaitan, S.**, Veldhuis, J. t., and Giesen, N. v. d. *Proof of concept: detecting urban flooding in crowdsourced images and surveillance video.* Submited to ACM-Dev, 2016b.

8. **Gaitan, S.**, Giesen, N. v. d., and Veldhuis, J. t. *Can urban pluvial flooding be predicted by open spatial data and weather data?.* Environmental Modelling & Software 85, 156–171. doi:10.1016/j.envsoft.2016.08.007, 2016a.

7. **Gaitan, S.** and ten Veldhuis, J. *Opportunities for multivariate analysis of open spatial datasets to characterize urban flooding risks.* Proceedings of the International Association of Hydrological Sciences, 370:9–14, June 2015. ISSN 2199-899X. doi: 10.5194/piahs-370-9-2015. URL http://www.proc-iahs. net/370/9/2015/.

6. Ochoa-Rodriguez, S., Wang, L.-P., Gires, A., Pina, R. D., Reinoso-Rondinel, R., Bruni, G., Ichiba, A., **Gaitan, S.**, Cristiano, E., van Assel, J., Kroll, S., Murlà-Tuyls, D., Tisserand, B., Schertzer, D., Tchigu- irinskaia, I., Onof, C., Willems, P., and ten Veldhuis, M.-C. *Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation.* Journal of Hydrology, 531, Part 2:389–407, December 2015. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2015.05.035. URL http://www.sciencedirect.com/science/article/pii/S0022169415003856.

5. **Gaitan, S.**, ten Veldhuis, J., and van de Giesen, N. *Spatial Distribution of Flood Incidents Along Urban Overland Flow-Paths.* Water Resour Manage, pages 1–13, May 2015. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-015-1006-y. URL http://link.springer.com/article/10.1007/ s11269-015-1006-y.

4. **Gaitan, S.**, Calderoni, L., Palmieri, P., ten Veldhuis, J. A. E., Maio, D., and van Riemsdijk, M. *From Sensing to Action: Quick and Reliable Access to Information in Cities Vulnerable to Heavy Rain.* IEEE Sensors Journal, 14(12):4175–4184, December 2014. ISSN 1530-437X. doi: 10.1109/JSEN.2014. 2354980.

3. **Gaitan, S.**, ten Veldhuis, J., Spekkers, M., and van de Giesen, N. C. *Urban vulnerability to pluvial flooding: complaints location on overland flow routes.* In Proceedings of the 2nd European Conference on Flood Risk Management FLOODrisk2012, Rotterdam, The Netherlands, 19-23 November 2012, pages 338–339, The Netherlands, 2012. CRC Press.

2. Graciani, S., **Gaitan, S.**, and Bas, G. *Soil Moisture Content Determination from SAR Images: Case of Study of the Salado River Low Basin (Santa Fe – Argentina).* In Annals of the VI Brazilian Colloquium of Geodesic Sciences, Brazil, 2009. Federal University of Paraná.

1. **Gaitan, S.**, Pardi-LaCruz, S., and Santos, J. R. d. *Multi-temporal Analysis and Change Prediction of Vegetation Cover and Land Use in the North-West of Colombian Amazonia.* In Annals of the VIII Seminar for Actualization on Remote Sensing and Geographic Information Systems Applied to Forestry, Brazil, 2008. Foundation for Forestry Research of Paraná.