International Conference on Flood Resilience:
Experiences in Asia and Europe
5-7 September 2013, Exeter, United Kingdom

CORRELATIONS BETWEEN RAINFALL DATA AND INSURANCE DAMAGE DATA RELATED TO SEWER FLOODING FOR THE CASE OF AARHUS, DENMARK

Matthieu Spekkers[1,*], Qianqian Zhou[2], Karsten Arnbjerg-Nielsen[2], Francois Clemens[1], Marie-claire ten Veldhuis[1]

[1] Delft University of Technology, the Netherlands; [2] Technical University of Denmark, Denmark
[*] *PO Box 5048, 2600 GA, Delft, the Netherlands; E-mail: M.H.Spekkers@tudelft.nl*

ABSTRACT

Sewer flooding due to extreme rainfall may result in considerable damage. Damage data to quantify costs of cleaning, drying, and replacing materials and goods are rare in literature. In this study, insurance claim data related to property damages were analysed for the municipality of Aarhus, Denmark. The aim of this paper was to study the extent to which rainfall data can be used to explain variations in insurance claim data. In particular, the paper addresses the issue of time-lag between claim date and time of the damaging rainfall event, which may, if not taken into account, lead to underestimations of correlations between rainfall and damage variables. Rainfall data from two rain gauges were used to extract rainfall characteristics. From cross correlations between time series of rainfall and claim data, it can be concluded that rainfall events induce claims mostly on the same day, but also on the three days after. A linear model that takes into account rainfall data from previous days slightly improves correlations between rainfall and damage variables compared to a simple linear model. Best correlation coefficients were found between maximum hourly rainfall intensity and daily number of claims (0.47-0.57) and daily total damage (0.43-0.53).

KEYWORDS

Sewer flooding; insurance damage data; regression analysis

1. INTRODUCTION

Flood damage assessment has gained increasing importance in recent years in support of decision-making on climate change adaptation and mitigation planning (European Commission, 2012; Merz et al., 2010). As a result of technological improvements, a large variety of models are now available to assess the economic loss incurred by floods for cost-benefit estimations. The calibration and validation of those models, however, have always been challenging due to a lack of damage data and poor data quality (Freni et al., 2010).

Insurance databases can be considered as important means to develop damage functions or to validate whether results obtained from damage assessment models are consistent with damage observations. Ideally, insurance damage databases should cover years of claimed costs with information on, for example damage cause, flood location, depth and extent.

In the context of sewer flooding, it is of interest to understand the processes of how rainfall results in flood damage. A few studies have investigated relationships between rainfall characteristics and statistics of damage data related to sewer flooding by means of insurance data. For instance, Spekkers et al. (2013), Einfalt et al. (2012) and Zhou et al. (in press) looked into the feasibility of deriving damage functions related to property and content using insurance claim data and rainfall data from rain gauges and weather radars. Zhou et al. (in press) used insurance data to validate a flood damage model that includes an inundation model in combination with simple depth-damage functions. These studies concluded that there is a high potential to establish a relationship between rainfall and damage statistics; however, a significant amount of the variance in damage variables is still left unexplained. This may be due to the fact that resolutions of rainfall and damage data and/or selected statistical approaches are not sufficient to establish rainfall-damage relationships.

This study builds on earlier work by Zhou et al. (in press), who analysed an insurance damage database related to sewer flooding for the municipality of Aarhus, Denmark. The claim database contains 1044 geo-referenced records related to property and content damage for the period of 2005-2011. This paper aims to investigate the extent to which rainfall data can be used to explain recorded damage claims. In particular, it addresses the issue of time-lag between claim date and time of the damaging rainfall event, which may, if not taken into account, lead to underestimations of correlations between rainfall and damage variables.

2. METHODS

2.1 Case study description

The case study area, displayed in Figure 1, covers the entire municipality of Aarhus, which is the second largest city of Denmark. It is located on the east side of the Jutland peninsula with a population about 315000 inhabitants. The area is about 47000 hectares and equipped with separate and combined sewer systems in different regions. The mean annual precipitation varies between 600 to 650 mm in the area and the elevation varies from 0 to 107 m above mean sea level. As a result of climate change (Madsen et al., 2009; Arnbjerg-Nielsen, 2006) and urban growth more pressure is put on the sewer systems in Aarhus nowadays. One of the most notable flood events in the last decade is that of 3-4 May 2005; Aarhus was hit by an extreme rainfall event with 50 mm of rainfall in 140

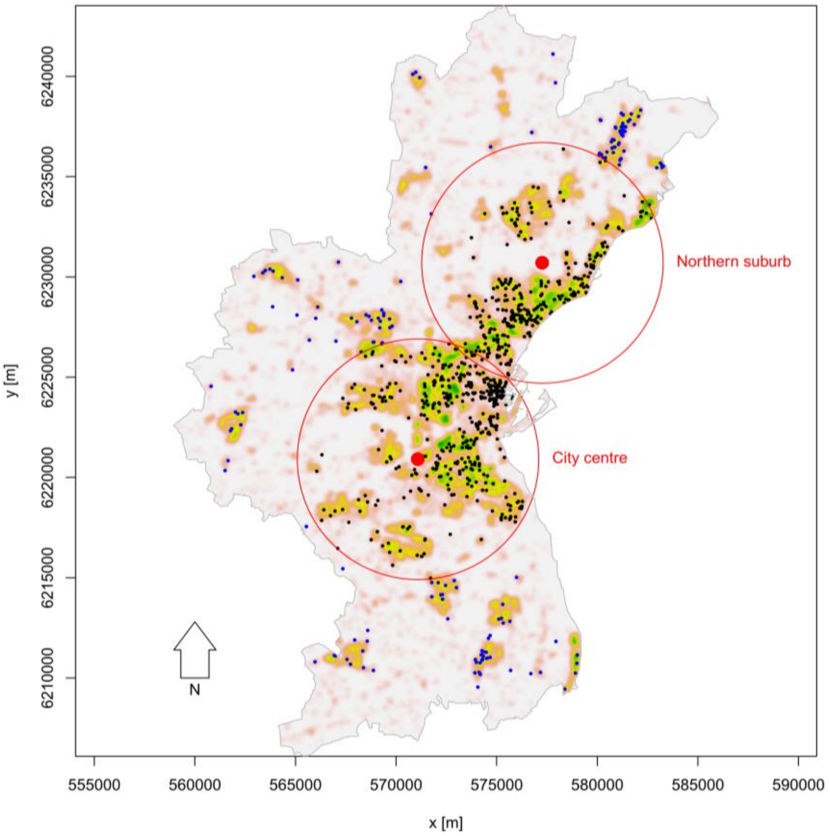

Figure 1. Municipality of Aarhus, Denmark. The colour map on the background depicts building density (0-2500 buildings/km^2). The two large red dots represent the location of rain gauges. The two red circles (labelled “Northern suburb” and “City centre”) have a radius of 6 km and were used to select claim data. The small black dots are locations of claims recorded in the period of 2005-2009 which were selected for analysis (n=813). The blue dots are the ones that were excluded (n=180).

Table 1. Summary of rainfall and insurance data.

Data source	Period	Temporal resolution	Spatial resolution	Availability
Rain gauge radar data	2005-2011	1 minute	two rain gauges	100%
Insurance database	2005-2009	daily records	point data	993 records

minutes and 43 insurance claims were reported afterwards. The area has experienced several rainfall extremes in the following years that resulted in considerable claim numbers, e.g. on 1-2 August 2006, 25 June 2007 and 10 August 2009.

2.2 Data

Table 1 summarizes the rainfall and damage data used for this study. Rainfall data from two nearby tipping buckets were used, which were converted into time series with a 1-minute temporal resolution. The locations of the rain gauges are marked in Figure 1 with two large red dots.

Damage claim data are available for the period of 2005-2011 provided by the Danish Association of Insurers and are related to sewer floods. Damage data in the period of 2005-2009 are based on data from five large companies; one of those five companies also provided data for the period of 2010-2011. However, the years 2010-2011 are discarded because of the low data coverage. For the period of 2005-2009, the database consists of 993 daily records related to individual addresses. A record means that damage was compensated by the insurer. Each record includes the claimed expenses related to property and content and a date on which the claim was received by the insurer. It is assumed that insurance companies have a 24/7 service. This assumption is supported by the fact that the numbers of claims received on any day of the week are not significantly different from the expected value.

Rain gauge data are generally assumed to be representative within a range of several kilometres. Therefore, only data within a 6-km range of one of the two rain gauges, displayed in Figure 1 with red circles, were selected. These two areas cover 813 records in total; 363 claims in the northern suburb, 450 claims in the city centre. Claims were assigned to their nearest rain gauge and aggregated by day. Table 2 lists the variables that were considered. Days with no claim data were assigned zero values.

2.3 Linear time-invariant system

The complexity in finding correlations between rainfall and insurance data stems from the fact that there is a lagged response of a damage claim to the time of rainfall event that caused the damage. The claim date insurers record is not the date on which damage occurs, but the date on which insurers receive and file a claim, which may be from the day of the event up to a few days after. As a consequence, relationships between rainfall and claim series cannot adequately be characterized by correlating them aligned in time.

A way to overcome this is by considering a linear relationship between input x_t, i.e. the value of a rainfall-related variable at day t, and output y_t, i.e. the value of a damage-related variable at day t, in

Table 2. Definitions of rainfall and damage variables.

Variable abbr.	Description	Unit
Damage-related		
dcounts	Number of claims = number of claims per day within a 6-km range of a rain gauge	-
dtot	Total damage = total amount of property and content damage per day within a 6-km range of a rain gauge	DKK/day
Rainfall-related		
rmax	Maximum hourly rainfall intensity = maximum rainfall intensity on a day, based on an 1-hour moving time window	mm/h

the form of (Box et al., 1994, Ch. 10):

$$y_t = \sum_{j=0}^{\infty} v_j x_{t-j} + \varepsilon_t \tag{1}$$

where v_j is the impulse response function at lag j and ε_t the error term at day t. The impulse response function is proportional to the theoretical cross correlation function $\rho_{xy}(j)$:

$$v_j = \frac{\rho_{xy}(j)\sigma_y}{\sigma_x} \tag{2}$$

where σ_x and σ_y are the standard deviations of the series. Because the theoretical cross correlation function is unknown, estimates from data must be used to calculate the impulse response function:

$$\hat{v}_j = \frac{r_{xy}(j)s_y}{s_x} \tag{3}$$

where $r_{xy}(j)$ is the sample cross correlation function between the series and s_x and s_y the sample standard deviations. A square root transformation of the damage-related variables was applied to stabilize variance. The square root transformation was selected as it can handle zero values, which were present in the insurance claim series for most days of the year.

To properly estimate cross correlations, any autocorrelation that may be present in the input series should be removed by prewhitening the input series first. In this study, the prewhitening procedure is followed as is explained by Chatfield (2004, p. 158) and Box et al. (1994, p. 417-418): a higher order

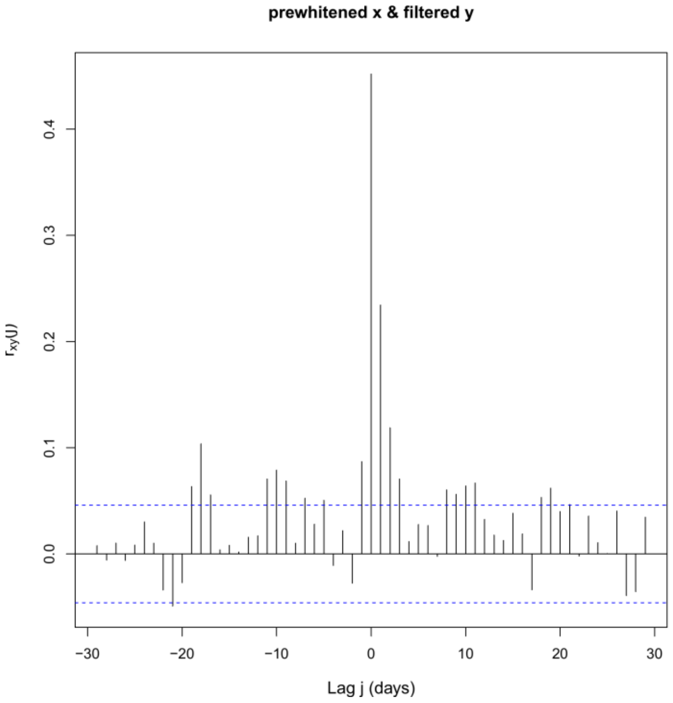

Figure 2. The cross correlation function between maximum hourly rainfall intensity and number of claims is plotted after prewhitening the *x* series and filtering the *y* series (northern area). The cross correlation function is the correlation between the series shifted against one another as a function of number of lags (days). A positive lag is a correlation between the rainfall variable (*x*) at a time before day *t* and the damage variable (*y*) at day *t*. The blue dashed lines are the 1% significance levels.

autoregressive-moving-average (ARMA) model is fitted to the input series such that residuals are uncorrelated. The same ARMA model should than be used to filter output series before computing the cross correlations between the series.

3. RESULTS AND DISCUSSION

As an example, the cross correlation function between maximum hourly rainfall intensity and number of claims is plotted in Figure 2 after prewhitening the rainfall series with an AR(5) model and filtering the insurance claim series with the same model. Rainfall series were prewhitened because they show weak autocorrelation at lag 1-5 and mostly at lag 1 (figures with autocorrelation functions not shown in this paper). Because of the weak autocorrelation, the choice of the order of AR model did not affect the estimation of cross correlations much. An AR model was selected as a first attempt to remove autocorrelation from the time series; this study could be improved by selecting the most suitable ARMA model in the future.

Significant cross correlations are found between lag *j*=-1 and *j*=3 (Figure 2). A few lags not close to lag 0 also show significant cross correlations, but these do not have a clear pattern and are therefore ignored. Highest correlation is observed at lag *j*=0 and correlations taper to zero for lags *j*>0. Similar results are found for cross correlations between other combinations of rainfall-related and damage-related variables. In the context of this study, significant cross correlation at lags with *j*<0 should not be possible. The spike at lag *j*=-1 may be explained by the fact that maximum rainfall intensity is based on 1-hour time window, which may fell between 23:00 and 24:00 the day before.

To summarize, a rainfall event induces claims mostly on the same day, but also significantly on the three days after. Therefore, a linear model is considered that takes into account lags *j*=0 to *j*=3; Eq. 1 reduces to:

$$y_t = \hat{v}_0 x_t + \hat{v}_1 x_{t-1} + \hat{v}_2 x_{t-2} + \hat{v}_3 x_{t-3} \tag{4}$$

Figure 3 compares fitted and observed values for a simple linear model (left figure), $y_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$, and the linear time-invariant model (right figure) as formulated in Eq. 4, for the relationship between maximum hourly rainfall intensity and square root transformed number of claims. The correlation

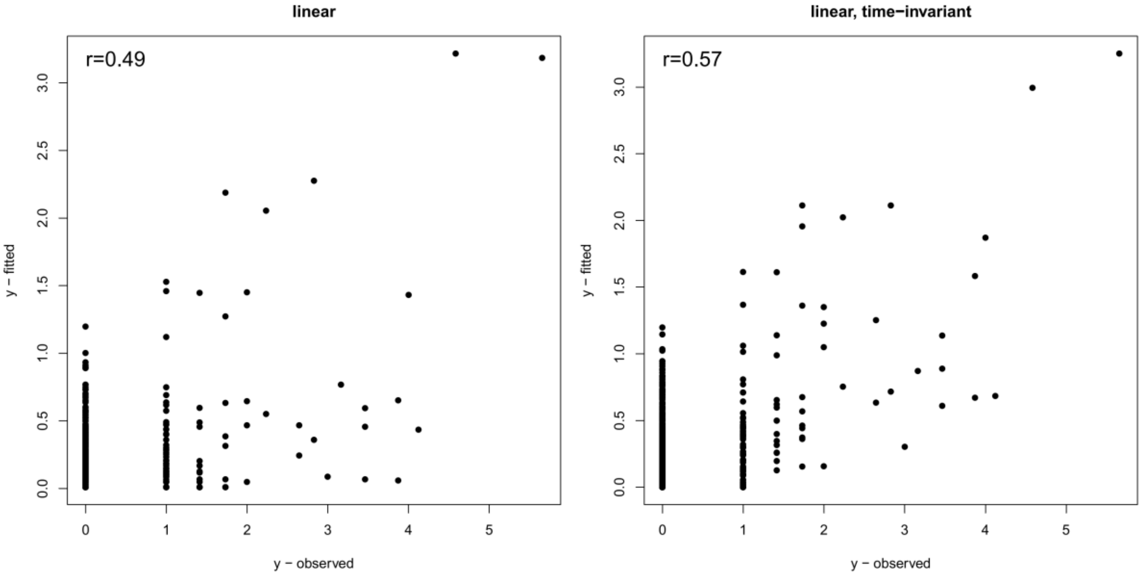

Figure 3. Fitted *y* is plotted versus observed *y* for a simple linear model (left) and a linear time-invariant model (right), for the relationship between maximum hourly rainfall intensity and square root transformed number of claims (*y*). Plots are related to the data of the northern suburb.

Table 3. Overview of Pearson correlation coefficients for various models. The relationships were found to be statistically significant (p<0.01).

		Northern suburb		City centre	
	Model	dcounts	dtot	dcounts	dtot
rmax	Linear	0.49	0.45	0.40	0.37
	Linear, time-invariant	0.57	0.53	0.47	0.43
rvol	Linear	0.45	0.40	0.38	0.34
	Linear, time-invariant	0.54	0.48	0.44	0.38
Avg. distance claim-rain gauge		3.1 km		4.0 km	
Total number of claims		363		450	

slightly improves if rainfall data from previous days are taken into account. This is best visible for the data points near the horizontal axis in the left figure, which are shifted upwards in the right figure.

Table 3 summarizes the Pearson correlation coefficients for various linear combinations of rainfall and square root transformed damage variables. In general, a linear time-invariant model performs slightly better than a simple linear model in terms of correlation coefficients. The maximum hourly rainfall intensity is a better predictor than rainfall volume, although the differences are small. Best correlations coefficients were found between maximum hourly rainfall intensity and number of claims per day (0.47-0.57) and daily total damage (0.43-0.53). The northern area shows overall better correlations, which may be explained by the smaller average distance between claim addresses and rain gauge: 3.1 km for the northern suburb and 4.0 km for the city centre.

4. CONCLUSIONS

The aim of this study was to investigate the extent to which rainfall data can be used to explain statistics of insurance damage claims related to sewer flooding. Results are based on around 800 insurance claims related to property and content flood damage for the case of Aarhus, Denmark, in the period of 2005-2009 and rainfall data from two nearby rain gauges.

From cross correlations between daily time series of rainfall and insurance claim data, it can be concluded that rainfall events induce claims mostly on the same day, but also significantly on the three days after. A linear model that takes into account rainfall input from previous days slightly improves correlations compared to a simple linear model. Best correlation coefficients were found between maximum rainfall intensity and number of claims (0.47-0.57) and total damage (0.43-0.53).

The results show that rainfall only explains a part of the variations in insurance claim data. This is, for example, relevant for the development of flood damage models, which should also take other damage-influencing factors than rainfall into account. Potential factors for sewer flood damage that will be studied in future research include local topography, building and household characteristics, urban drainage properties and spatial variability of extreme rainfall.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge the help from Forsikring & Pension in providing the insurance data.

6. REFERENCES

Arnbjerg-Nielsen, K. (2006). Significant climate change of extreme rainfall in Denmark. *Water Science & Technology*, 54(6-7):1.

Box, G., Jenkins, G., and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*. Prentice-Hall International, Inc., New Jersey.

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*, Sixth Edition. Chapman and Hall/CRC Press LLC.

Einfalt, T., Pfeifer, S., and Burghoff, O. (2012). Feasibility of deriving damage functions from radar measurements. In 9th International Workshop on Precipitation in Urban Areas, 245–249, St. Moritz (Switzerland).

European Commission (2012). A new EU Flood Directive (accessed on 7-3-2012 from http://ec.europa.eu/environment/water/flood_risk).

Freni, G., La Loggia, G., and Notaro, V. (2010). Uncertainty in urban flood damage assessment due to urban drainage modelling and depth-damage curve estimation. *Water Science and Technology*, 61(12):2979–2993.

Madsen, H., Arnbjerg-Nielsen, K., and Mikkelsen, P. S. (2009). Update of regional intensity-duration-frequency curves in Denmark: Tendency towards increased storm intensities. *Atmospheric Research*, 92(3):343–349.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, a. (2010). Review article “Assessment of economic flood damage”. *Natural Hazards and Earth System Science*, 10(8):1697–1724.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2013). A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences*, 17(3):913–922.

Zhou, Q., Panduro, T. E., Thorsen, B. J., and Arnbjerg-nielsen, K. (in press). Verification of flood damage modelling using insurance data. *Water Science & Technology*.