A Stochastic Infilling Algorithm for Spatial-Temporal Rainfall Data

A thesis presented in partial fulfillment of the requirements
for the Degree of

Master of Science
in
Statistics

at Massey University, Albany, New Zealand

David Russell Munroe

2005

ABSTRACT

The purpose of this thesis is to develop an infilling algorithm for 24-hour (daily) rainfall data. An infilling algorithm replaces missing data within the historical records with sensible estimates, where any appropriate method (prediction from a fitted model, interpolation between points, or random sampling) could be used to select and/or produce the required estimates. The algorithm developed uses simulation data generated using a stochastic point-process model which has been fitted to historical data. In this thesis, the spatial-temporal Neyman-Scott rectangular pulse model as presented in Cowpertwait et al. (2002) is fitted to data provided by Thames Water from 23 sites in the Thames Valley (UK). The model is shown to fit the data reasonably well; however it fails to fit the proportion of dry sites (which is not used in the fitting process). Nevertheless, simulated data is generated using the model and an infilling algorithm is derived. The algorithm is tested by replacing valid historical data with missing values, infilling these missing values, and then comparing relevant statistics for the two samples. Three algorithms are developed in this thesis, of which the final algorithm maintains the statistical characteristics of the historical data, including the proportion of dry sites, while infilling values that are similar to the known historical record.

Preface

In general, only a sample of the plots produced for any given analysis are included in text within this thesis. This sampling is both for brevity and clarity. Further plots are generally included within the appendices at the end of the thesis.

Note for the spatial data analysis, approximately 700 figures were generated. As only ten plots led to a definite conclusion, only these plots have been included in the Appendices.

Furthermore, for consistency, and to make comparison easier, the same months (January and July) were used to represent any seasonal differences where applicable. However, considerable variation occurs between the seasons and the results for the other months included in the supplementary appendices should not be overlooked.

Acknowledgements

I would like to thank my supervisor, Paul Cowpertwait, for his guidance and stimulating discussions throughout this project. In addition, the use of Bruce Mill's bitmap generating code was most appreciated for the production of the simulation movie.

I would also like to thank my family and friends for their support and prayers during this period. This is much appreciated as without them this work would not have been possible.

Finally, I am grateful to Thames Water for permitting the use of their data.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

> The Lord will open the heavens, the storehouse of His bounty, to send rain on your land in season ...
>
> **Deuteronomy 28:12a NIV**

1.1 Background

Accurate modelling of rainfall is critical to the successful design of effective urban drainage and stormwater systems. In order to build such systems, a long historical record is necessary so that the likelihood of extreme events and their relative location can be estimated. However, records of sufficient length and fine resolution are not available. As a result, considerable attention over the last two decades has been placed on developing a model suitable for simulating rainfall.

A fitted model, however, is only as accurate as the source data that the model is based on. Furthermore, any hydrodynamical model (eg: for surface runoff or flood frequency analysis) is heavily dependent on the rainfall modelling component as any inadequacy is directly incorporated into the pipe flow models (Mark and Hosner, 2002).

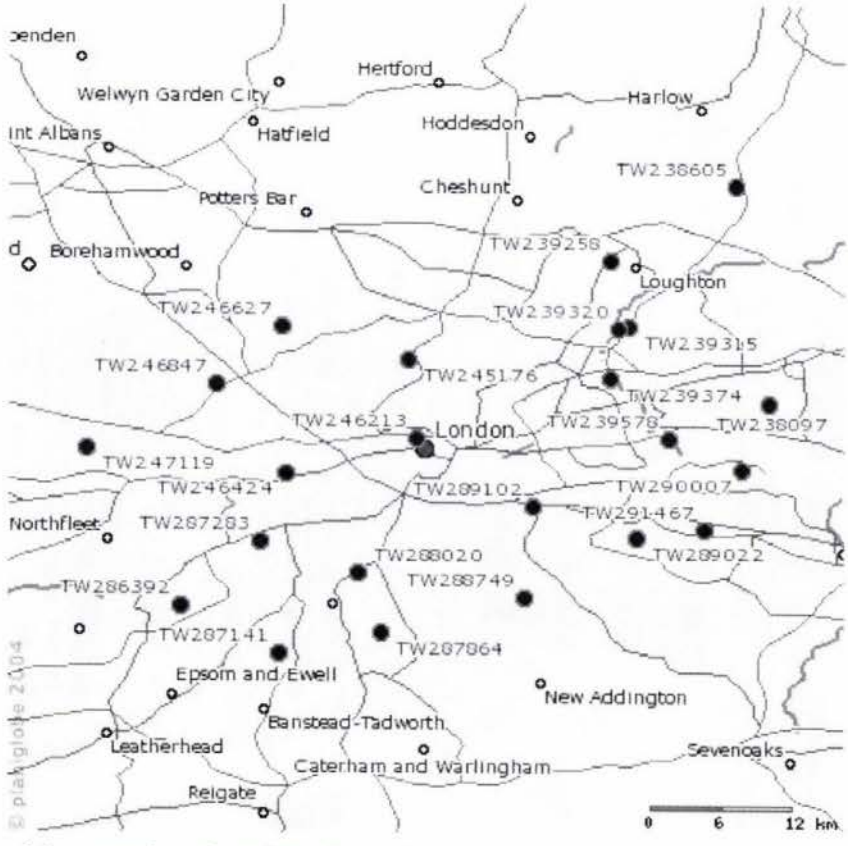
There are two main problems associated with rainfall data. Firstly, the data are sensitive to recording error, especially at fine aggregation levels (see Section 1.2.1). Secondly, the available historical data are generally sparsely populated with valid recordings.

The majority of historical source data available are collected at a 24-hour (daily) resolution. To a lesser extent, 1-hour records are also obtainable. However,

for drainage purposes, it is necessary to have a much finer timescale (for example 1-5 minute resolution). This presents two major hurdles to be overcome before a realistic parametric model can be produced. The data must be fully populated and made available at a useful resolution.

1.2 *Source data*

The data used in this project were collected from rain gauges from twenty-three sites in the Thames catchment from 1970 to 2003. A map of the site locations is shown in Figure 1.1. The site names along with corresponding site numbers are listed in Table 1.1 along with their corresponding Easting/Northing grid coordinates.



Map produced online from `http://www.planiglobe.com/`

Figure 1.1: Map of site locations in the Thames catchment

Table 1.1: Site location coordinates: Easting, Northing, and Altitude

Thames ID	Site number	Easting (0.1)km	Northing (0.1)km	Altitude (m)
TW238097	1	5499	1863	16
TW238605	2	5476	2048	75
TW239258	3	5412	1981	115
TW239315	4	5423	1926	15
TW239320	5	5418	1923	17
TW239374	6	5415	1882	8
TW239578	7	5447	1830	2
TW245176	8	5308	1894	33
TW246213	9	5314	1828	25
TW246424	10	5246	1795	21
TW246627	11	5241	1920	78
TW246847	12	5208	1870	42
TW247119	13	5141	1815	23
TW286392	14	5194	1682	12
TW287141	15	5247	1641	47
TW287283	16	5234	1737	56
TW287864	17	5299	1661	35
TW288020	18	5286	1712	40
TW288749	19	5375	1692	33
TW289022	20	5433	1745	75
TW289102	21	5377	1771	5
TW290007	22	5486	1805	8
TW291467	23	5468	1754	50

In general, a Thames Water site name is preferred to a site number, particularly when analysing the historical data (Chapter 4). However, if it is not necessary to be able to immediately identify a particular site, then the sites are referred to by their corresponding site number.

Two aggregation levels were available for use: a 1-hour record and a 24-hour (daily) record. The 24-hour record was substantially longer and covered the years $1970 - 2003$. The 1-hour record was only available from $1989 - 2003$. All sites had some data, however, for some sites (eg: TW239315, TW289022) the percentage of valid records after the data were cleaned was quite low (see Section 4.4.1).

1.2.1 *Measurement error*

There are a number of devices for recording rainfall measurements. The oldest method is to use a collection of rain gauges, however, more recent developments enable collection of data via radar networks or satellite sensors (Maidment, 1993).

The data used within this project were collected from rain gauges (Section 1.2). Tipping bucket rain gauges, used by the Environment Agency in the United Kingdom (see Tilford et al., 2003, chap. 2) for real-time monitoring of rainfall, are affected by a variety of environmental conditions. The primary sources of measurement error associated with this collection method are well known and include wind speed, the height of the gauge above the ground, and snowfall (Tilford et al., 2003; Maidment, 1993). Note that the errors associated with measuring snowfall are usually larger than rainfall (Maidment, 1993).

1.3 *Thesis outline*

Techniques for fully populating the historical record at the 24-hour aggregation level will be derived within this thesis. The constructed algorithms will use a synthetic record generated using a spatial-temporal point process model of rainfall (see Section 3.1). Once a technique for infilling the historical record at a 24-hour level is developed, methods for *disaggregation* from this fully populated record can be applied (for example Zhiquan and Eltahir, 1994; Glasbey et al., 1995; Güntner et al., 2001; Kottegoda et al., 2003; Cowpertwait et al., 2004). Where *disaggregation* describes an algorithm to generate data at a finer resolution (eg: 1-hour) than the available data (eg: 24-hour). Note that generally the total amount of rainfall at the same site and time interval (24-hour) is expected to match exactly for the available and disaggregated records.

The remainder of this thesis is organised as follows. The next chapter (Chapter 2) presents a review of the literature for both rainfall models and infilling algorithms. In Chapter 3, the model is mathematically described along with the fitting technique. Furthermore, the methods for cleaning the historical data are presented, the infilling algorithms are proposed, and the implementation of the algorithms is discussed. The results of the data analysis and cleaning is covered in Chapter 4. In Chapter 5, the results of the model fitting, validation, and infilling algorithms are presented. Finally, in Chapter 6 the conclusions and directions for future research for the model and infilling are derived.