

Bayesian modelling of recurrent pipe failures in urban water systems using non-homogeneous Poisson processes with latent structure

Submitted by

Theodoros Economou

to the University of Exeter as a thesis for the degree
of Doctor of Philosophy in Mathematics, September 2010.

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred upon me.

.....
Theodoros Economou

Abstract

Recurrent events are very common in a wide range of scientific disciplines. The majority of statistical models developed to characterise recurrent events are derived from either reliability theory or survival analysis. This thesis concentrates on applications that arise from reliability, which in general involve the study about components or devices where the recurring event is failure.

Specifically, interest lies in repairable components that experience a number of failures during their lifetime. The goal is to develop statistical models in order to gain a good understanding about the driving force behind the failures. A particular counting process is adopted, the non-homogenous Poisson process (NHPP), where the rate of occurrence (failure rate) depends on time. The primary application considered in the thesis is the prediction of underground water pipe bursts although the methods described have more general scope.

First, a Bayesian mixed effects NHPP model is developed and applied to a network of water pipes using MCMC. The model is then extended to a mixture of NHPPs. Further, a special mixture case, the zero-inflated NHPP model is developed to cope with data involving a large number of pipes that have never failed. The zero-inflated model is applied to the same pipe network.

Quite often, data involving recurrent failures over time, are aggregated where for instance the times of failures are unknown and only the total number of failures are available. Aggregated versions of the NHPP model and its zero-inflated version are developed to accommodate aggregated data and these are applied to the aggregated

version of the earlier data set.

Complex devices in random environments often exhibit what may be termed as state changes in their behaviour. These state changes may be caused by unobserved and possibly non-stationary processes such as severe weather changes. A hidden semi-Markov NHPP model is formulated, which is a NHPP process modulated by an unobserved semi-Markov process. An algorithm is developed to evaluate the likelihood of this model and a Metropolis-Hastings sampler is constructed for parameter estimation. Simulation studies are performed to test implementation and finally an illustrative application of the model is presented.

The thesis concludes with a general discussion and a list of possible generalisations and extensions as well as possible applications other than the ones considered.

Acknowledgements

It was a privilege and honour to have studied at a department which consisted of people whose support was only matched by their expertise. In particular I am indebted to both my supervisors Trevor Bailey and Zoran Kapelan for their trust, guidance and most importantly for their sense of humour.

To Yehuda Kleiner for kindly providing the data for the thesis.

Also to Renato whose friendship and advice kept me sane most of the time and to Rachel for her support and understanding. I would also like to thank Yiwei for making me smile. Also Pete, Graham and Matt for their computing support.

Special thanks to Les Grossman without whom the last few years would be significantly more miserable.

To Dimitris who was always there for me and to the many people and friends who made this experience enjoyable and unforgettable.

Finally, I wish to dedicate this to my parents and sister for their unconditional support and love, and to Victoria for giving me strength and for completing me.

Contents

Acknowledgements	4
Contents	5
List of Figures	11
List of Tables	14
1 Introduction	16
1.1 Motivation	16
1.2 Aims	19
1.3 Structure of Thesis	20
1.4 Summary	21
2 Background	22
2.1 Recurrent Event Modelling	22
2.1.1 Counting Processes	23

2.1.2	Reliability Theory and Survival Analysis	26
2.2	Counting Processes in Reliability	29
2.2.1	Renewal Processes	29
2.2.2	Non-Homogeneous Processes	30
2.2.3	Processes with Dependency Structures	31
2.3	Recurrent Failures in Water Pipe Networks	32
2.4	Summary	37
3	Data	38
3.1	Basic Summary Analysis of the Data	38
3.2	Tests for temporal trends	42
3.3	Summary	43
4	Mixed Effect NHPP Models	45
4.1	Non-Homogeneous Poisson Process Models	45
4.1.1	Failure Rate	46
4.1.2	Likelihood Function	50
4.1.3	Left Truncation	52
4.1.4	Random Effects	53
4.1.5	Bayesian Framework and MCMC	55
4.2	Mixed Effects NHPP model	56

4.2.1	Simulation Experiments	56
4.2.2	Model Application	61
4.3	Summary	68
5	Mixtures of NHPP models	70
5.1	Motivation	70
5.2	Mixture Models	71
5.2.1	Likelihood Function	72
5.2.2	Label Switching	73
5.2.3	Zero-Inflated Poisson Models	74
5.3	Extending the NHPP for Zero-Inflation	76
5.3.1	Simulation Experiments	76
5.3.2	Model Application	81
5.4	Summary	87
6	NHPP for Aggregated Data	90
6.1	NHPP model for Aggregated Data	90
6.1.1	Model Formulation	91
6.1.2	Simulation Experiments	92
6.1.3	Model Application	95
6.2	Aggregated Zero-Inflated NHPP	100

6.2.1	Model Formulation	101
6.2.2	Simulation Experiments	102
6.2.3	Model Application	105
6.3	Summary	112
7	Hidden Semi-Markov NHPP	114
7.1	Motivation	114
7.2	Hidden Markov Models	117
7.3	Markov Modulated Poisson Processes	119
7.4	Hidden Markov NHPP Model	121
7.4.1	Model Formulation	121
7.4.2	Likelihood	123
7.4.3	Recursive Algorithms: Forward	124
7.4.4	Recursive Algorithms: Backward	127
7.4.5	E-M Algorithm	128
7.4.6	State Holding Times	129
7.5	Review of Hidden Semi-Markov Models	130
7.6	Hidden Semi-Markov NHPP Model	133
7.6.1	Semi-Markov Chains	133
7.6.2	Likelihood Formulation	136

7.6.3	Forward Algorithm	137
7.6.4	Backward Algorithm	144
7.7	MCMC Model Implementation	145
7.7.1	Metropolis-Hastings	145
7.7.2	Proposal Distribution	147
7.7.3	Label Switching	149
7.7.4	Prior distributions	151
7.8	Simulation Experiments	152
7.9	Model Application	153
7.10	Summary	159
8	Conclusions	160
8.1	Thesis Summary	160
8.2	Practical Issues	162
8.3	Future Considerations	163
A	Bayesian Framework and MCMC	166
A.1	Outline	166
A.1.1	Metropolis-Hastings Algorithm	168
A.1.2	Gibbs Sampling	169
A.1.3	Prior Distributions	171

A.1.4	Convergence	172
A.1.5	Posterior Predictive Diagnostics	175
A.1.6	Model Comparison	177
B	R and WinBUGS code	180
B.1	WinBUGS Code	180
B.2	R Code	183
	References	189

List of Figures

2.1	Daily flow levels for river Severn in 2005	24
3.1	Pipe age vs pipes	39
3.2	Histogram of total number of failures per pipe	40
3.3	Failure counts vs pipe length	40
3.4	Monthly number of failures vs month (1962-2003)	41
3.5	Monthly number of failures vs pipes (1-1349)	42
3.6	Laplace test for 865 pipes	43
4.1	Bathtub-shaped failure rate function	47
4.2	Examples of log-linear failure rate	48
4.3	Examples of power-law failure rate	49
4.4	Actual vs predicted no. of failures	60
4.5	Samples of the log-posterior	63
4.6	Deviance samples - actual (red) and simulated data (black)	64
4.7	Posterior means and Cr.I. for θ_i	65

4.8	Estimated ranks vs actual ranks	66
4.9	Estimated number of failures vs individual pipes	68
5.1	Rank of p_i vs rank of \hat{p}_i	79
5.2	Actual vs predicted no. of failures	80
5.3	\hat{R} for 1355 model parameters	82
5.4	Samples of the log-posterior	83
5.5	Deviance samples - actual (red) and simulated data (black)	83
5.6	Posterior means and Cr.I. for θ_i and p_i	85
5.7	Estimated ranks vs actual ranks	86
5.8	Estimated number of failures vs individual pipes	86
6.1	Actual vs predicted no. of failures	95
6.2	Samples of the log-posterior	96
6.3	Deviance samples - actual (red) and simulated data (black)	97
6.4	Posterior means and Cr.I. for θ_i	98
6.5	Estimated ranks vs actual ranks	99
6.6	Estimated number of failures vs individual pipes	99
6.7	Rank of p_i vs rank of \hat{p}_i	105
6.8	Actual vs predicted number of failures	105
6.9	\hat{R} for 1355 model parameters	106

6.10	Samples of the log-posterior	107
6.11	Deviance samples - actual (red) and simulated data (black)	107
6.12	Posterior means and Cr.I. for θ_i and p_i	109
6.13	Estimated ranks vs actual ranks	110
6.14	Estimated number of failures vs individual pipes	110
7.1	Exponentially increasing intensity function	115
7.2	Clustered/Repellent failures	116
7.3	Hidden Markov NHPP model	122
7.4	Hidden Markov NHPP model	130
7.5	A semi-Markov chain	134
7.6	Semi-Markov modulated NHPP	136
7.7	Monthly failure counts vs month	156
7.8	Samples of the log-posterior	157
7.9	Deviance samples - actual (red) and simulated data (black)	158
7.10	Observed and estimated cumulative number of failures	158

List of Tables

3.1	Network description	39
3.2	Summary statistics for monthly failures	41
4.1	Simulated data	58
4.2	Simulated NHPP results	59
4.3	NHPP parameter estimates	64
4.4	NHPP confusion matrices	67
5.1	Simulated ZINHPP results	78
5.2	Simulated ZINHPP results	78
5.3	ZINHPP parameter estimates	84
5.4	ZINHPP confusion matrices - observation period	87
5.5	ZINHPP confusion matrices - prediction period	88
6.1	Simulated aggNHPP results	94
6.2	\hat{R} values	96

List of Tables

6.3	aggNHPP parameter estimates	97
6.4	aggNHPP confusion matrices - observation period	100
6.5	aggNHPP confusion matrices - prediction period	101
6.6	Simulated aggZINHPP results	104
6.7	Simulated aggZINHPP results	104
6.8	aggZINHPP parameter estimates	108
6.9	aggZINHPP confusion matrices - observation period	111
6.10	aggZINHPP confusion matrices - prediction period	112
7.1	Priors, input values and estimates	154
7.2	Pipe information	155
7.3	Priors and estimates	157