

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# Non-Parametric Estimation of Geographical Relative Risk Functions

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Palmerston North

New Zealand

W. T. P. Sarojinie Fernando

December 2012

# Abstract

The geographical relative risk function is a useful tool for investigating the spatial distribution of disease based on case and control data. The most common way of estimating this function is using the ratio of spatial kernel density estimates constructed from the locations of cases and controls respectively. This technique is known as the density ratio method. The performance of kernel density estimators depends on the choice of kernel and the smoothing parameter (bandwidth). The choice of kernel is not critical to the statistical performance of the method but the bandwidth is crucial. Different bandwidth selectors such as least squares cross validation (LSCV) and likelihood cross validation (LCV) are chosen to control the degree of smoothing during the computation of the density ratio estimator.

An alternative way of estimating this relative risk function is local linear regression approach. This deserves consideration since the density ratio estimator can be less natural when the relative risk has a global trend, as one might expect to see when there is a line source of risk such as a polluted river or a road. The use of local linear regression for estimation of log relative risk functions *per se* has not been examined in any detail in the literature, so our work on this methodology is a novel contribution. A detailed account of local linear approach in the estimation of log relative risk function is provided, consisting of an analysis of asymptotic properties and a method for computing tolerance contours to emphasize the regions of significantly high risk. Data driven bandwidth selectors for the local linear method including a novel plug-in methodology is examined.

A simulation study to compare the performance of density ratio and local linear estimators using a range of data-driven bandwidth selectors is presented. The analysis of two specific data sets is examined.

The estimation of the spatial relative risk function is extended to spatio-temporal estimation through the use of suitable temporal kernel functions, since time-scale is an important consideration when estimating disease risk. The extended version of the kernel density estimation is applied here to compute the unknown densities of the spatio-temporal relative risk function. Next we investigate the time derivatives of the space-time relative risk function to see how the disease change with time. This discussion provides novel contributions with the introduction to time derivatives of the relative risk function as well as asymptotic methods for the computation of tolerance contours to highlight subregions of significantly elevated risk. LSCV and subjective bandwidths are used to compute these estimators since it performs well in density ratio method. The analysis on a real application to foot and mouth disease (FMD) of 1967 outbreak is employed to illustrate these estimators.

The relative risk function is investigated when the data include a spatially varying covariate. The discussion produces the introduction to generalized relative risk function in two ways and also asymptotic properties of estimators for both cases as novel works. Generalized kernel density estimation is used to replace the unknown densities in the relative risk function. Asymptotic theories are used to compute tolerance contours to identify the areas which show high risk. LSCV bandwidth selector is described in this estimation process providing the implicit formulae. We illustrate this methodology on data from the 2001 outbreak of FMD in the UK, examining the effect of farm size as a covariate.

# Acknowledgements

I am sincerely and heartily grateful to my supervisor, Martin Hazelton, for his continuous support and guidance throughout this research. I am sure this would have not been accomplished without his help. I would also like to thank Martin for being patient specially when I was making less progress. Finally I would like to express my heartfelt gratitude to Martin for his valuable comments and suggestions to direct this dissertation a successful one.

I am thankful to Ganes Ganesalingam for his continuous guidance in numerous ways as being my co-supervisor in the early years of my research. My thanks go to Jonathan Marshall for being my co-supervisor in my final year due to the replacement of Ganes.

I would like to acknowledge all the members in the Statistics group for their valuable comments, and also colleagues in the Institute office of Fundamental Sciences. My special thanks go to IFS for providing me a scholarship at the end of Graduate assistant position and also to Steve Haslett for providing me funds to pay tuition fee afterwards. I should thank my postgraduate colleagues at the Statistics group, Massey University for their encouragement and providing me a pleasing environment.

My special thanks go to two important people. Sarath Kulatunga, a Senior Professor at the University of Kelaniya in Sri Lanka, helped me to acquire a thorough statistical knowledge while I was working towards my M.Phil. with him. I would also like to express my gratitude to Priyantha Wijayatunga, a lecturer at Umea University

in Sweden, who encouraged me to apply to this Graduate assistant Ph.D. position. I take this opportunity to thank to all staff members at the Department of Mathematics, University of Kelaniya. My thanks also go to staff members at the faculty of Applied sciences, specially to my colleagues at the Department of Mathematics, Wayamba University of Sri Lanka for making a friendly environment while I was working there as a lecturer.

Of course, I am indebted to my lovely husband Kithsiri Fernando for his love, support, motivation and constant patience which have taught me so much about sacrifice during this journey. Without him this work would never have come into existence. I am also grateful to my son, Kaveesha Fernando and my daughter, Gihara Fernando for their profound understanding during this period. I missed them so much. I am extremely sorry for the time we spent apart.

I would like to acknowledge my lovely mum, Janet for loving, encouraging me always, believing in me, in all my efforts. I also express my gratitude to my late dad, Julian for making a dream in my mind to be a successful academic when I was a child. I also like to acknowledge my brother and sister for their continuous support. My special remind goes to my mother-in-law, Mary who unfortunately passed away while I was reading for my Ph.D.

Last but not least, I would like to thank my ever loving God for helping and guiding me to be successful throughout my life.

# Table of Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>Table of Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>xii</b>
<b>Notation</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation and problem description . . . . .	1
1.2 Organization of the thesis . . . . .	5
<b>2 Non-parametric estimation of spatial relative risk function</b>	<b>9</b>
2.1 Introduction . . . . .	9
2.2 Kernel density estimation . . . . .	12
2.2.1 Univariate case . . . . .	13
2.2.2 Bivariate case . . . . .	17
2.3 Technical problems arising in relative risk estimation . . . . .	20
2.3.1 Data scarcity . . . . .	20
2.3.2 Edge correction . . . . .	22
2.4 Asymptotic properties . . . . .	24
2.4.1 Univariate case . . . . .	24
2.4.2 Multivariate case . . . . .	26
2.5 Tolerance contours of density ratio estimators . . . . .	30
2.6 Real applications . . . . .	31

2.6.1	Chorley-Ribble data . . . . .	32
2.6.2	Myrtle tree data . . . . .	34
2.7	Conclusion . . . . .	35
<b>3</b>	<b>Bandwidth selection for the density ratio estimator</b>	<b>38</b>
3.1	Introduction . . . . .	38
3.2	Error criteria . . . . .	40
3.3	Cross validation bandwidth selectors . . . . .	42
3.3.1	Least squares cross validation . . . . .	42
3.3.2	Likelihood cross validation . . . . .	46
3.4	Simulation study to compare LSCV over LCV in $\hat{\rho}$ estimation . . . . .	48
3.5	Real application: Cancers in South Lancashire . . . . .	52
3.6	Conclusion . . . . .	53
<b>4</b>	<b>Local linear estimation of the relative risk function</b>	<b>56</b>
4.1	Introduction . . . . .	56
4.2	Local linear estimator of the relative risk function . . . . .	60
4.3	Local scoring procedure . . . . .	63
4.4	Asymptotic properties . . . . .	65
4.5	Methods of bandwidth selection . . . . .	68
4.5.1	Plug-in bandwidth selector . . . . .	69
4.6	Simulation study to compare local linear against density ratio estimator	72
4.6.1	Simulation results: with optimal smoothing . . . . .	72
4.6.2	Simulation results: with data-driven bandwidths . . . . .	77
4.7	Tolerance contours of local linear estimators . . . . .	82
4.8	Real applications . . . . .	85
4.8.1	Myrtle tree data . . . . .	85
4.8.2	Chorley-Ribble cancer data . . . . .	86
4.8.3	Foot and mouth disease (FMD) data . . . . .	87
4.9	Conclusion . . . . .	90
<b>5</b>	<b>Estimation of spatio-temporal relative risk function</b>	<b>91</b>
5.1	Introduction . . . . .	91
5.2	Spatio-temporal relative risk function . . . . .	93
5.3	Spatio-temporal kernel density estimation . . . . .	95
5.4	Edge correction . . . . .	96
5.5	Asymptotic properties of spatio-temporal relative risk estimators . . . . .	98
5.6	Tolerance contours of $\hat{\rho}$ . . . . .	99
5.7	Bandwidth selection . . . . .	101



5.7.1	Least squares cross validation . . . . .	102
5.7.2	Likelihood cross-validation . . . . .	103
5.8	Real application: FMD of 1967 outbreak . . . . .	105
5.9	Time derivative relative risk estimation . . . . .	107
5.9.1	Time derivative density estimation . . . . .	108
5.9.2	Tolerance contours of $\frac{\partial}{\partial t} \hat{\rho}(\mathbf{z}; t)$ . . . . .	111
5.9.3	Revisit to FMD application . . . . .	113
5.10	Conclusion . . . . .	116
<b>6</b>	<b>Non-parametric estimation of relative risk with covariates</b>	<b>124</b>
6.1	Introduction . . . . .	124
6.2	Relative risk function with a covariate . . . . .	126
6.3	Kernel density estimation with a covariate . . . . .	128
6.4	Asymptotic properties . . . . .	128
6.4.1	Bias and variance of $\hat{\rho}(\mathbf{x}, \mathbf{z})$ . . . . .	128
6.4.2	Bias and variance of $\hat{\rho}(\mathbf{x} \mathbf{z})$ . . . . .	130
6.5	Bandwidth Selection . . . . .	132
6.6	Real application: The 2001 outbreak of FMD . . . . .	133
6.7	Conclusion . . . . .	134
<b>7</b>	<b>General Discussion</b>	<b>137</b>
7.1	Summary of my work . . . . .	137
7.2	Suggestions for future work . . . . .	140
	<b>Appendix</b>	<b>142</b>
	<b>Bibliography</b>	<b>166</b>

# List of Figures

1.1	The geographical distribution of larynx (58 cases-●) and lung (978 controls-+) cancer data. Incinerator is displayed as ■. . . . .	4
2.1	Univariate kernel density estimate based on six observations: solid lines - individual kernels, bold line - kernel density estimate . . . . .	14
2.2	Scaled univariate kernel functions. . . . .	15
2.3	Kernel density estimates for 299 observations of a bimodal density. Bandwidths: (a) 1; (b) 8; (c) 4. . . . .	16
2.4	Left panel displays the scatter plot of larynx cancer data and the right panel shows the bivariate kernel density estimate, constructed from these data. The subjective bandwidth, 2 is used. . . . .	19
2.5	Sparse data can be seen mainly in southwest and southeast regions. .	21
2.6	Shaded areas represent the subregions which contribute to the kernel estimate, also is located outside of the region, so needs edge correction. .	23
2.7	P-values surface based on the asymptotic theory describing the excess risk for a given fixed bandwidth, $h = 1$ . White solid lines indicate 95% tolerance contours. . . . .	32
2.8	Estimates of the log-relative risk of larynx cancer in the Chorley-Ribble region of Lancashire, England. The estimate is computed using the density ratio method with subjective bandwidth $h = 1$ . The dashed lines indicate 95% tolerance contours for areas of elevated risk. The red square represents the incinerator. . . . .	33

2.9	Plot of 106 diseased (●) and 221 healthy (+) Myrtle Beech trees in Tasmania. . . . .	35
2.10	Estimates of the log-relative risk of disease from the Myrtle Beech data. This estimate is obtained by using the density ratio method with subjective bandwidth $h = 30$ . The solid line indicates 95% tolerance contours for areas of elevated risk. . . . .	36
3.1	Filled contour plots of the control densities, uniform(left panel) and bivariate normal (right panel) as described in the text. . . . .	50
3.2	Filled contour plots of the log-relative risk functions as described in Table 3.1. Problems 1 and 4 represent the risk surface 1. Problems 2 and 5 represent the risk surface 2. Problems 3 and 6 represent the risk surface 3. . . . .	51
3.3	Boxplots of $\log(\text{ISE})$ of log-relative risk estimates for problems 1-6 from Table 3.1. LCV and LSCV stand for likelihood and least squares cross validation bandwidths respectively. . . . .	54
3.4	Estimates of log-relative risk of larynx cancer data. Left panel shows the estimate using LCV bw (2.74) while the right panel, using the LSCV bw (0.78). . . . .	55
4.1	Contour plots of case density $f$ , control density $g$ , and log relative risk function $\rho$ for the four synthetic problems. . . . .	75
4.2	Boxplots of $\log(\text{ISE})$ for DR and LL estimates of $\rho$ . The suffix 1 indicates sample sizes $n_1 = n_2 = 100$ and similarly for 500. . . . .	76
4.3	Filled contour plots for the test relative risk functions (on the log scale). . . . .	79
4.4	Boxplots of $\log(\text{ISE})$ for estimates of the log-relative risk for test problem 1 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text. . . . .	80

4.5	Boxplots of $\log(\text{ISE})$ for estimates of the log-relative risk for test problem 2 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text. . . . .	82
4.6	Boxplots of $\log(\text{ISE})$ for estimates of the log-relative risk for test problem 3 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text. . . . .	83
4.7	Boxplots of $\log(\text{ISE})$ for estimates of the log-relative risk for test problem 4 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text. . . . .	84
4.8	Estimates of the log-relative risk of disease from the Myrtle Beech data. The left-hand panel shows the estimate using the density ratio method with least-squares cross-validation bandwidth $h = 64$ , and the right-hand one the local linear estimator using our plug-in bandwidth $h = 197.3$ . The dashed lines indicate 95% tolerance contours for areas of elevated risk. . . . .	85
4.9	Estimates of the log-relative risk of larynx cancer in the Chorley-Ribble region of Lancashire, England. The left-hand panel shows the estimate using the density ratio method with least-squares cross-validation bandwidth $h = 0.78$ , and the right-hand one the local linear estimator using our plug-in bandwidth $h = 16.66$ . The dashed lines indicate 95% tolerance contours for areas of elevated risk. . . . .	87
4.10	Spatial distribution of FMD data. 100 cases ( $\bullet$ ), 2129 controls ( $+$ ). . . . .	88

4.11	These Figures show the estimates of log-relative risk of disease from FMD data. The left panel is used DR method with LSCV bandwidth $h = 1.74$ , while the right panel shows the estimate using LL method with PI bandwidth $h = 13.65$ . The dashed lines indicate 95% tolerance contours for areas of elevated risk. . . . .	89
5.1	The distribution of FMD case data. Day is shown in the title. . . . .	105
5.2	Control data distribution of FMD. . . . .	106
5.3	Log relative risk estimates( $\rho$ ) for FMD data in day 1, 6, 16 and 24. Subjective bandwidth is used. 5% tolerance contours are displayed in white. . . . .	108
5.4	Log relative risk estimates ( $\rho$ ) for FMD data in day 1, 6, 16 and 24. LSCV bandwidth is used. 5% tolerance contours are displayed in white.	109
5.5	Derivative density estimation for days 1-9. White lines indicate 95% tolerance contours. Case data are scattered in red. . . . .	114
5.6	Derivative density estimation for days 10-18. White lines indicate 95% tolerance contours. Case data are scattered in red. . . . .	115
5.7	Derivative density estimation for days 19-27. White lines indicate 95% tolerance contours. Case data are scattered in red. . . . .	116
5.8	Derivative density estimation for days 28-36. White lines indicate 95% tolerance contours. Case data are scattered in red. . . . .	117
5.9	Derivative density estimation for days 37-45. White lines indicate 95% tolerance contours. Case data are scattered in red. . . . .	118
5.10	Derivative density estimation for days 1-9. LSCV bandwidth is used.	119
5.11	Derivative density estimation for days 10-18. LSCV bandwidth is used.	120
5.12	Derivative density estimation for days 19-27. LSCV bandwidth is used.	121
5.13	Derivative density estimation for days 28-36. LSCV bandwidth is used.	122
5.14	Derivative density estimation for days 37-45. LSCV bandwidth is used.	123

6.1	Cases and controls for the FMD dataset, including the defined region. Each bullet point represents a farm. . . . .	135
6.2	Heatplots of FMD relative risk surfaces (on log scale), with 5% tolerance contours (solid internal lines). The covariate (total population) is displayed as the title at each plot. LSCV bandwidth is used. . . . .	136

# List of Tables

3.1	Control densities and relative risk functions for six synthetic models. The function $\phi_\sigma$ is a bivariate normal density with zero mean vector and covariance matrix $\sigma^2 I$ , where $I$ is the $2 \times 2$ identity matrix. In addition, $\phi_\sigma^{\mathcal{R}}$ denotes $\phi_\sigma$ truncated to $\mathcal{R}$ . The location parameters are $\mu_1 = [4, 4]^T$ , $\mu_2 = [5, 5]^T$ and $\mu_3 = [6, 6]^T$ . . . . .	49
4.1	The local scoring procedure . . . . .	64
4.2	Case and control densities for four synthetic problems. The function $\phi_\sigma$ is a bivariate normal density with zero mean vector and covariance matrix $\sigma^2 I$ , where $I$ is the $2 \times 2$ identity matrix. In addition, $\phi_\sigma^{\mathcal{R}}$ denotes $\phi_\sigma$ truncated to $\mathcal{R}$ . The location parameters are $\mu_1 = [5, 5]^T$ , $\mu_2 = [8, 6]^T$ , $\mu_3 = [3, 3]^T$ and $\mu_4 = [5, 4]^T$ . . . . .	73
4.3	Optimal bandwidths for the Problems for density ratio (DR) and local linear (LL) estimators. . . . .	74
4.4	MISE estimates in the simulation study to compare DR over LL methods.	74
4.5	Relative risk functions for test problems. . . . .	77
4.6	MISE estimates in the simulation study to compare DR over LL methods. Medians are displayed inside brackets. . . . .	81

# Notation

$\mathbf{I}$	$2 \times 2$ identity matrix.
$f$	Case density function
$g$	Control density function
$r$	Relative risk function
$\rho$	Log relative risk function
$\text{supp}(f)$	Support of $f$
$p$	Degree of the polynomial, conditional probability and size of covariate.
$h$	Fixed bandwidth
$h_1$	Fixed bandwidth for case data.
$h_2$	Fixed bandwidth for control data.
$h_{OS}$	Over smoothing bandwidth.
$q$	Edge correction factor.
$q_1$	Edge correction factor based on case data.
$q_2$	Edge correction factor based on control data.
$\mathbf{H}$	bandwidth matrix
$\Sigma$	Covariance matrix.
$\lambda$	temporal bandwidth
$K$	Unscaled kernel function
$K_h$	Scaled kernel function, scaled with $h$ .
$\mu_2(K)$	Second central moment of $K$
$R(f)$	$\int_{\mathcal{R}} f(\mathbf{x})^2 d\mathbf{x}$
$R(K)$	$\int_{\mathcal{R}} K(\mathbf{z})^2 d\mathbf{z}$
$R(f'')$	$\int_{\mathcal{R}} f''(\mathbf{z})^2 d\mathbf{z}$
$f''$	Second derivative of the density $f$ .
$L$	Univariate probability density function.
$\mathcal{D}_{\{ \mathbf{x} \}}$	First derivative of $f$ with respect to $\mathbf{x}$ .
$\mathcal{H}_{\{ \mathbf{x} \}}$	Hessian of $f$ with respect to $\mathbf{x}$ .
$\phi_{\sigma}(\mathbf{x} - \mu)$	Multivariate normal density.
$\mu$	Mean.



$\sigma^2$	Variance.
$\mathcal{R}$	Geographical region
$ \mathcal{R} $	Area of the geographical region $\mathcal{R}$ .
$\nabla^2$	Laplacian operator.
DR	Density Ratio
LL	Local Linear
RR	Relative risk
ISE	Integrated Squared Error
MSE	Mean Squared Error
AMSE	Asymptotic Mean Squared Error
MISE	Mean Integrated Squared Error
WMISE	Weighted Mean Integrated Squared Error
MIAE	Mean Integrated Absolute Error
AMISE	Asymptotic Mean Integrated Squared Error
PI	Plug-In bandwidth selector
GAM	Generalized additive model.
LSCV	Least Squares Cross Validation
LCV	Likelihood Cross Validation
SE	Standard Error
$x_i$	scalar.
$\mathbf{x}_i$	A vector of dimension 2
$x_1, \dots, x_n$	Random scalar sample of size $n$ .
$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$	a random sample of size $n$
$t$	Time of occurrence
$T$	Time interval
$ T $	Length of time interval
FMD	Foot and mouth disease
KDE	Kernel density estimate
$\hat{f}(x; h)$	Fixed kernel density estimate constructed from univariate data $x$ .
$\hat{f}(\mathbf{x}; h)$	Fixed kernel density estimate constructed from bivariate case data $\mathbf{x}$ .
$\hat{g}(\mathbf{x}; h)$	Fixed kernel density estimate constructed from bivariate control data $\mathbf{x}$ .
$\hat{r}(\mathbf{x}; h)$	Fixed bandwidth relative risk estimate.
$\hat{\rho}(\mathbf{x}; h)$	Fixed bandwidth log relative risk estimate.
$\hat{\rho}_{LL}$	LL estimator of log relative risk function.
$\hat{\rho}_{LC}$	Local constant estimator.
$\hat{f}^{-i}(\mathbf{x}; h)$	Leave-one-out estimate based on case data
$\hat{g}^{-i}(\mathbf{x}; h)$	Leave-one-out estimate based on control data
$h_{opt}$	The optimal bandwidth
$h_{MISE}$	MISE- optimal bandwidth
$\hat{h}_{PI}$	Plug-in bandwidth selector
$\hat{h}_{LSCV}$	LSCV bandwidth selector
$\hat{h}_{LCV}$	LCV bandwidth selector
$\hat{h}_{LL.LCV}$	LCV bandwidth for LL estimator.
$\hat{h}_{LL.PI}$	PI bandwidth for LL estimator.

$\hat{h}_{DR.CV}$	LSCV bandwidth for DR estimator.
$\hat{g}_p$	Pooled estimator.
$n_1$	Case sample size
$n_2$	Control sample size
$n$	Equal to $n_1 + n_2$
$\pi$	Equal to $\frac{n_1}{n_1+n_2}$
$\bar{L}$	The likelihood function
$\bar{L}'$	The first derivative with respect to $\beta$
$\bar{L}''$	The second derivative with respect to $\beta$
$P, Q$	Polynomials.
$p(\mathbf{x})$	Conditional probability for a given point $\mathbf{x}$ .
$y_i$	Binary variable.
Cov	Covariance.
tr	Trace.
$\delta$	A small positive number.
$\delta_0$	A small positive number.
$\epsilon$	A small constant.
$\bar{f}$	Edge corrected kernel density estimate constructed from case data.
$\bar{g}$	Edge corrected kernel density estimate constructed from control data.
$H_0, H_0^1, H_0^2$	Null hypotheses.
$H_1, H_1^1, H_1^2$	Alternative hypotheses.
$\mathbb{E}$	Expectation.
$\theta$	A parameter.
$Z$	Test statistic.
$W_{\mathbf{x}}$	$n \times n$ diagonal matrix at $\mathbf{x}$ .