2011

Approaches to the spatial modelling of cancer incidence and mortality in metropolitan Perth, Western Australia, 1990 -2005

Changying Shao
Edith Cowan University

Edith Cowan University

Copyright Warning

APPROACHES TO THE SPATIAL MODELLING OF CANCER INCIDENCE AND MORTALITY IN METROPOLITAN PERTH, WESTERN AUSTRALIA, 1990-2005

DOCTOR OF PHILOSOPHY (MATHEMATICS)

CHANGYING SHAO

FACULTY OF COMPUTING, HEALTH AND SCIENCE

EDITH COWAN UNIVERSITY

WESTERN AUSTRALIA

June 2011

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

STATEMENT OF CONFIDENTIAL INFORMATION

As a user of information at the Department of Health Western Australia (DOHWA), I complied with the terms stated by DOHWA on the approval of data usage.

Terms and conditions are:

DOHWA Human Research Ethics Committee (HREC) holds the Principal Investigator responsible for the ethical conduct of the project and the security of the personal health information therefore he/she must

Report anything which might warrant review of ethical approval of the project in the specified format including:
Any serious or unexpected adverse events.
Unforeseen events that might affect the continued ethical acceptability of the project.
Submit for approval any changes or amendments to the research protocol, including methodology, data required, duration of the project and any changes to the approved data storage arrangements.
Advise if the project is discontinued or withdrawn before the expected date of competion and give reasons for this action.
Provide an annual progress report to the HREC and a final report at the completion of the project.
Advise any changes of personnel in the research team, and provide a DOHWA Confidentiality Agreement/Confidentiality Acknowledgement form for any addition to the research team.

Signature:…….....
Name: Changying Shao

ABSTRACT

Cancer is one of potentially preventable and treatable diseases. Cancer analysis from different perspectives is necessary to provide the information for health research and the initiation of prevention and treatment programs. The purpose of this study was to analyse five top cancers in the Perth metropolitan area, including lung, melanoma, breast, prostate and colorectal cancers, using two methodologies: Area-to-Point Poisson (ATP) kriging and fitting an inhomogeneous Poisson process model using the Berman-Turner algorithm.

ATP Poisson kriging was used to undertake the analysis on the spatial distribution of cancer rates per 100,000 person-years for Perth Statistical Local Areas during the period 1990-2005. This analysis is based on age-adjusted and age-sex-adjusted rates (non-sex specific cancers only). For lung, colorectal and melanoma cancer incidence, it is also evaluated for the period 1990-2005 by sex to investigate how these cancers affect males and females differently. For mortality data, three contiguous periods: 1990-1995, 1996-2000 and 2001-2005 are studied to investigate how cancers vary over the time. It is demonstrated that female cancer incidence varies more continuously than males and male incidence rates are relatively high. There is difference in estimates and spatial variation between age-adjusted rates and age-sex-adjusted rates. The cancer risk calculated from age-sex-adjusted rates shows less variability than the risk only from age-adjusted rates. Lung cancer incidence rates are high in Kwinana industrial area and melanoma cancer incidence rates are relatively high along west coast of Perth. For breast cancer and prostate cancer, the analysis during three contiguous periods is explored. Breast cancer incidence increased gradually with time while prostate cancer mortality decreased with time. Prostate cancer mortality risk from the age is possibly more continuous than the incidence risk from the age. It is believed that cancer mortality rate varies more continuously than the corresponding cancer incidence and that spatial correlation is relatively strong for mortality compared with the incidence.

Spatial point pattern theory focuses on the analysis of the incidence (mortality) rates in the year 2005. For each individual cancer adequately fitting models were determined and validated. The model fitting is based on the inhomogeneous Poisson nature of the cancer data. The spatial trends based on the population distribution in Perth metropolitan area are

important in obtaining an adequate fit. Except for the spatial trends, the fitted intensity function for the point process is also related to the spatially susceptible population. It is demonstrated that the distance from Kwinana industrial area (KIA) is an appropriate covariate for lung cancer incidence and the distance from the coast accounts for melanoma incidence. Lung cancer incidence increases with proximity to the KIA and similarly melanoma incidence increases with proximity to the coast. The incidence point pattern in 2005 is quite similar to the point pattern in other individual year (or years) when population density is included. Compared with incidence, one covariate (susceptible population intensity) is enough to achieve an adequate fit in most cases for mortality data fitting. The percentage of selected age groups is a reasonable covariate for prostate cancer and lung cancer while population density is a good covariate for other cancers since age has more influence on these two cancer incidences than other cancers. The work shown here presents spatial maps of incidence and mortality risk for the five major causes in Perth, potentially aiding planning.

COPYRIGHT AND ACCESS DECLARATION

I certify that this thesis does not, to the best of my knowledge and belief:

incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education;
contain any material previously published or written by another person except where due referenc is made in the text; or
contain any defamatory material.

Signed:…….
Date:…….

ACKNOWLEDGMENT

First of all, I would like to thank my principal supervisor, Prof. Ute Mueller, for the kindness, guidance and support given me over the period of my PhD study. I really appreciate her consistent advice, both personal and professional, which has helped me finish the thesis. My thanks also go to my associate supervisor, Prof. James Cross, for valuable discussions and professional guidance.

The data for this study were provided by the Department of Health of Western Australia. I would like to thank Dr Timothy Threllfall of the Cancer Registry who extracted the data used in this thesis and who made himself available for discussion.

Last but not least, I appreciate the support of my family, my parents, my parents-in-law, my husband and my two boys: Vincent and Victor. It would have been impossible to finish my thesis without their immense help.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acronyms

ABS: Australian Bureau of Statistics

OK: Ordinary kriging

PK: Poisson kriging

PWA: Population-weighted average

SK: Simple kriging

SLA: Statistical local area

TCCWA: The Cancer Council Western Australia

UTM: Universal Transverse Mercator

WHO: World Health Organisation

WSS: Weighted sum of squares

Notation and Abbreviations

A: study area

$C(0)$: covariance value at separation distance $|\mathbf{h}| = 0$. It is also the stationary variance of the RV $Z(\mathbf{u})$

$N(\mathrm{h})$: the number of data pairs separated by the vector $\mathbf{h}$

$Z^*_{OK}(\mathbf{u})$: ordinary kriging estimator of $Z(\mathbf{u})$

$Z^*_{SK}(\mathbf{u})$: simple kriging estimator of $Z(\mathbf{u})$

$\hat{r}_{PWA}(\mathbf{u}_\alpha)$: population-weighted average estimator of $\mathrm{R}(\mathbf{u}_\alpha)$. The same type of notation applies to other algorithms, for example, *GBS* and *LBS*.

$C(\mathbf{h})$: stationary covariance of the RF $Z(\mathbf{u})$ for lag vector $\mathbf{h}$

D: domain

$\sigma^2_E(\mathbf{u})$: error variance of RV $Z(\mathbf{u})$

$\sigma^2_{SK}(\mathbf{u})$: simple kriging variance associated with the simple kriging estimate $Z^*_{SK}(\mathbf{u})$ at location $\mathbf{u}$. The same type of notation applies to other algorithms, for example, OK.

$\sigma^2_{PK}(\mathbf{u}_\alpha)$: Poisson kriging variance associated with the Poisson kriging estimate $\hat{\mathrm{R}}_{PK}(\mathbf{u}_\alpha)$ The same type of notation applies to other algorithms, for example, PWA, GBS and LBS.

$\gamma(\mathbf{h})$: stationary semivariogram of the RF $Z(\mathbf{u})$ for lag vector $\mathbf{h}$

$\mathbf{u}$: coordinate vector at a location

$\mathbf{u}_\alpha$: the recorded cancer incidence (or mortality) location

$\mathrm{d}(\mathbf{u}_\alpha)$: the recorded cancer incidence (or mortality) count at location $\mathbf{u}_\alpha$

$\mathrm{D}(\mathbf{u}_\alpha)$: the recorded cancer incidence (or mortality) count variable at location $\mathbf{u}_\alpha$

$\mathrm{n}(\mathbf{u}_\alpha)$: the size of the population at risk at location $\mathbf{u}_\alpha$

$\mathrm{z}(\mathbf{u}_\alpha)$: the observed incidence (or mortality) rates at location $\mathbf{u}_\alpha$

$\mathrm{Z}(\mathbf{u}_\alpha)$: the observed incidence (or mortality) rates variable at location $\mathbf{u}_\alpha$

$\mathrm{R}(\mathbf{u}_\alpha)$: the local cancer incidence (or mortality) risk

$Z(\mathbf{u})$: random variable at location $\mathbf{u}$

$\Re^N$: N-dimensional Euclidean space

$\mathbf{h}$: separation vector

$E\{\cdot\}$: expected value

$Var\{\cdot\}$: variance

m: stationary mean of the RF $\mathrm{R}(\mathbf{u}_\alpha)$

m^*: the population-weighted sample mean

$m(\mathbf{u})$: expected value of the RV $Z(\mathbf{u})$

s^2: variance of the rates

n_{tot}: the sum of the population at risk across N areas

$s^2(\mathbf{u}_\alpha)$: local variance around location $\mathbf{u}_\alpha$

$\overline{n}(\mathbf{u}_\alpha)$: the average population size across the local area.

$\mu(\mathbf{u})$: Lagrange multiplier location $\mathbf{u}$

δ_{ij}: when i = j it is 1 and 0 otherwise

σ^2_{R}: variance of $\mathrm{R}(\mathbf{u}_\alpha)$

$\mathrm{C}_{\mathrm{R}}(\mathbf{u}_\alpha - \mathbf{u}_\beta)$: stationary covariance of the RF $\mathrm{R}(\mathbf{u}_\alpha)$ for lag vector $\mathbf{u}_\alpha - \mathbf{u}_\beta$

$\hat{R}_{PK}(\mathbf{u}_\alpha)$: Poisson kriging cancer incidence (or mortality) risk estimator of $R(\mathbf{u}_\alpha)$

$\lambda_i(\mathbf{u}_\alpha)$: the weight assigned to the observed rate $Z(\mathbf{u}_i)$ when estimating the risk at location $\mathbf{u}_\alpha$

$\lambda(\mathbf{u}_\alpha)$: Bayes shrinkage factor

$\hat{R}_\alpha$: denotes the Poisson kriging risk estimator of $\hat{R}_{PK}(\mathbf{u}_\alpha)$

n_α: denotes the population size $n(\mathbf{u}_\alpha)$

R_α : denotes local risk $R(\mathbf{u}_\alpha)$

z_α: denotes $z(\mathbf{u}_\alpha)$

$\overline{n}$: denotes $\overline{n}(\mathbf{u}_\alpha)$

C_{ij}: denotes the covariance function $C_R(\mathbf{u}_i - \mathbf{u}_j)$

p_i: the percentage of age group i in WA 2000

r_i: the age-specific rate for the population of age group

N_i: the population of age group i in SLA

d: the total number of incidence cases over the period 1990-2005

$\gamma_e(h_\ell)$: experimental semivariogram value

$\gamma_m(h_\ell)$: model semivariogram value

$\mathbf{x}$: the observed point pattern

X: a random point process

1 INTRODUCTION

1.1 Background

Cancer is a major public health concern, as one of the world's leading causes of death with 12% of the 56 million deaths in 2000 being due to cancer (WHO 2003). In some countries more than 25% of deaths were attributed to cancer. In 2005 the percentage increased to 13%. Cancer killed more than 7 million people worldwide in 2006, which was more than AIDS, malaria, diabetes, tuberculosis, malnutrition, violence and war combined (WHO 2007). There were 7.6 million cancer deaths in the world in 2008 and the cancer deaths account for around 13% of all deaths (WHO 2011).

Cancer risk increases with age. With the population aging in some countries, cancer deaths are predicted to rise to an estimated 9 million in 2015 and to 11.4 million in 2030 (WHO 2007).

Cancer is the leading cause of death in Australia, - more than 36,000 people die from cancer annually, and the risk of developing some form of cancer is one in four for women and one in three for men (TCCWA 2007). At least 88,000 new cases of cancer are diagnosed each year.

In Western Australia (WA), at least 52,000 people are living with cancer. More than 8,000 new diagnoses are reported every year, an average of more than 20 per day (Threlfall and Thompson 2004). Western Australia covers approximately one-third of the Australian continent, but has less than one-tenth of the nation's population, with the majority of the population living in the Perth metropolitan area. The estimated population in the Perth metropolitan area is 1,659,000 according to Australian Bureau of Statistics (2009).

According to WHO (2011), more than 30% of cancer deaths can be prevented. This study aims to analyse cancer incidence and mortality recorded in the Perth metropolitan area during the period 1990-2005. There are five cancer types considered: lung, melanoma, breast, prostate and colorectal cancers. The analysis of cancer data will be presented at two

different spatial scales: Statistical Local Areas (SLA) scale and a point pattern analysis at finer scale.

1.2 Significance

According to the report by the Western Australian Cancer Registry (WACR 2008), the most common cancers in males in 2008 were prostate cancer, colorectal cancer, melanoma and lung cancer. For females, they are breast cancer, colorectal cancer, melanoma and lung cancer. In 2008 one in seven men could be expected to have a diagnosis of prostate cancer before the age of 75 and one in eleven women could be expected to develop breast cancer.

Information about the spatial distribution of cancers can be used for etiologic research and the guidance for control activities in community health. The data used in this study are cancer count data. The number of cancer cases can be regarded as the realisation of Poisson random variable. Poisson kriging is applied to cancer data analysis to take into account the Poisson nature. In addition, the data are recorded at domicile locations and so can be treated using point pattern analysis. The kriging analysis at Statistical Local Area (SLA) scale focuses on age-adjusted rates and age-sex-adjusted rates. In addition to age, sex is another important cancer risk. For example, males have a much higher lung cancer risk than females (WACR 2008), and this is also shown in Figure 4. As the population distribution by sex may vary between areas, it is necessary to take into account sex for lung cancer estimation. To the best of our knowledge this is the first attempt to include the sex composition of the susceptible population in cancer data analysis aiming at a better understanding of the correlation between cancer data and population composition. In addition, this is the first application of Area-to-point (ATP) Poisson kriging to cancer data in Perth based on the geographical characteristics of Perth.

Geostatistical modelling of cancer data takes place within areas (e.g. SLA) based on adjusted rates during the entire study period 1990-2005 or three contiguous periods: 1990-1995, 1996-2000 and 2001-2005. If the aggregate data are too rough (e.g. within areas greater than SLA), the fine view of geographical patterns of cancer data will be missed. If the aggregate data are too refined (e.g. suburb), the risk estimation will be unreliable, as

area might be small or sparsely populated. In fact, the relationship between cancer data and covariates is not always constant over space. To take into account the heterogeneous nature of the covariates, spatial point pattern analysis will be applied to investigate the data at a finer scale for the year 2005. We will explore the application of spatial point pattern analysis to cancer data by introducing Berman-Turner algorithm for modelling cancer data, aiming at finer understanding of the influence of covariates on cancer rates and exploring the intensity of a point pattern.

1.3 Literature Review

Geostatistical techniques have been applied to the modelling of cancer risk since 1998. Oliver et al. (1998) studied cases of cancer in children under fifteen years of age, Prou et al. (2003) provided an analysis of leukemia for children aged less than fifteen years, and Goovaerts and collaborators considered lung cancer (Goovaerts and Jacquez 2004; Goovaerts 2005a; Goovaerts 2005d), breast cancer (Goovaerts 2005b; Goovaerts 2005c; Goovaerts 2006b), brain and bladder cancer (Goovaerts et al. 2005), prostate cancer (Goovaerts 2005a), cervix cancer (Goovaerts 2005c; Goovaerts 2006a), pancreatic cancer (Goovaerts 2006b) . In addition, Lai (2004) studied sixteen kinds of cancers in some rural counties in China. Apart from geostatistical modelling, lung cancer mortality rates were studied using time series analysis (Shibuya et al. 2005), and breast cancer incidence was analyzed through spatiotemporal random fields (Christakos and Lai 1997).

Geostatistics has its origins in the prediction of ore grades in the analysis of mining data (Goovaerts 1997). Currently it is applied in various fields including ecology, agriculture, remote sensing, fisheries research and medical geography. Kriging is a group of geostatistical techniques for finding an unbiased estimate at unsampled locations using values obtained from nearby sample locations and it is in essence a spatial linear least-squares regression.

Simple kriging and ordinary kriging (Goovaerts 1997) are the most commonly used kriging algorithms. Simple kriging is mathematically the simplest. It assumes that the mean of the random function is constant throughout the entire study region. However, in most

applications that the mean may vary significantly over the study region. Ordinary kriging assumes that mean is locally constant but unknown throughout the entire study region.

The geostatistical analysis of disease data has received increasing attention with kriging becoming more popular. Lai (2004) performed ordinary kriging on Chinese cancer mortality data of 63 rural counties. To produce a set of contour maps, the spatial structure of the cancer mortality rates was studied but other possible covariates were not incorporated.

Classical methods of kriging, such as simple kriging and ordinary kriging, are used to estimate the value of a continuous attribute. In fact, the application of geostatistics in the medical field is relatively new. As disease count data are discrete, we need to use the Poisson kriging technique to take account of the discrete nature of the data.

A first attempt to take into account the discrete nature of cancer data was the use of binomial cokriging which was employed to produce a map of childhood cancer risk in the West Midlands Health Authority Region (WMHAR) of England (Oliver et al. 1998). The application of this technique to Long Island (USA) data led to negative variogram estimates. To avoid this problem, binomial cokriging was extended to the case when the variance of observed rates is smaller than expected under the binomial model. One geostatistical filtering approach used is modified binomial cokriging which was applied to estimate breast cancer incidence in Long Island, New York (Goovaerts 2005b) and lung cancer mortality recorded in the New England counties and the US State Economic Areas (Goovaerts 2005d). The modified technique was shown to be more flexible and robust concerning the underlying hypothesis that all counties have the same spatial support, and the simulation studies have demonstrated its more accurate estimates (Goovaerts 2005d). The technique was also used in cluster analysis to detect unreliable spatial outliers (Goovaerts 2005b).

Another geostatistical technique, Poisson kriging, was recently developed to filter noise from the data by accounting for spatially varying population sizes and spatial patterns. The methodology for estimating a spatial Poisson distribution was first introduced by Kaiser et al.(1997). They developed the spatial “auto-models” based on the Poisson distribution to be used to incorporate spatial dependencies among the variables. However, their model is not

well suited for irregularly sampled data and interpolation. Monestiez et al. (2004; 2006) introduced Poisson kriging to model spatially heterogeneous observations. The approach applied by Monestiez et al. (2004; 2006) is similar to binomial cokriging proposed by Oliver et al. (1998) except that the count data are assumed to follow a Poisson distribution.

Poisson riging was then generalised to estimate prostate cancer mortality risk in the United States (Goovaerts 2005a), breast and cervix cancer mortality in New England States (Goovaerts 2005c) and cholera and dysentery incidence risk in Bangladesh (Ali et al. 2006) by incorporating varying population sizes in the processing of cancer data. When the risk values were spatially correlated, simulation studies showed that in most cases, Poisson kriging outperformed other smoothers such as population-weighted estimators and empirical Bayes smoothers (Goovaerts 2005c). These studies also claimed that when the risk is spatially random, although it might not be the common situation, global empirical Bayes smoothers provide more accurate predictions.

It is not practical to represent each geographic unit by its centroid, especially when geographic units vary greatly in size and shape. The geographical characteristics need to be incorporated for data analysis together with spatially varying population. The framework for Area-to-area (ATA) or Area-to-point (ATP) kriging was first introduced by Kyriakidis (2004) for interpolating point values from available areal data. Goovaerts (2006a) modified the ATP estimator to a Poisson estimator and applied ATP Poisson kriging to lung and cervix cancer mortality in counties of the United States. In that study Goovaerts only used age-adjusted rates to filter noise and created a continuous map of mortality risk by taking into account the size and shape of study areas, as well as the population density, and so it is more flexible than Poisson kriging and the cancer risk estimation is more reliable. In addition, simulation studies showed that the use of ATP Poisson kriging results in more accurate estimation than point kriging, especially when the geographical units are heterogeneous. To date, geostatistical techniques have been applied to the data analysis for geographical units including County, ZIP code and census block group. This study will apply ATP Poisson kriging to cancer data in Perth at a different scale: SLA. Nearly all geostatistical analysis is within age-adjusted rates. This study takes into account not only age but also sex to remove the bias from the population structure in different SLAs.

An early example of work in spatial epidemiology is the study by Diggle (1990) who considered point patterns of larynx and lung cancers. The aim in this paper was to investigate whether there is evidence of raised larynx cancer incidence near the location of a now-disused industrial incinerator. The lung cancer cases were used as control for the spatial variation in susceptible population intensity. Since then spatial point pattern analysis has been applied to data from many different areas including incidence of disease (Gatrell et al. 1996), prediction of forest fires in Ontario (Podur et al. 2003) and a study of rainforest spatial point patterns with incomplete covariate data (Waagepetersen 2008).

Before fitting any models the nature of the point process needs to be determined partly to ensure that the data are not completely spatially random (CSR). CSR process is also known as homogeneous Poisson process, whose points are independent of each other and have the same propensity to be found at any location (Baddeley 2008). There are two ways in which a point pattern can depart from randomness: One is clustering where points tend to be close together than would be expected if their locations were CSR. The other is regularity where points tend to avoid each other. The point process modelling approach used by Diggle (1990) can detect clustering but cannot validate a clustered point process. Baddeley and his collaborators (2005) developed techniques to validate a point process model. To assess the departure from CSR Ripley's *K*-function can be applied but it is not appropriate when a spatial trend exists in the data. To take account of the inhomogeneity the inhomogeneous *K*-function (Baddeley et al. 2000) needs to be used.

Turner (2009) illustrated the techniques for analysing New Brunswick fire data by cleaning data, conducting exploratory data analysis, detecting the dependence between points, fitting models and validating models. However, no actual conclusions on the modelling were drawn. This study will explore the intensity of the point patterns using the Berman-Turner technique and interpret the outputs to draw the conclusions based on the modelling.

1.4 Objective

This study aims to analyse the top five cancers in Perth using two methodologies to provide estimates for cancer incidence and mortality rates at different spatial scales. One method is Area-to-Point Poisson (ATP) kriging using software SpaceStat which will provide

estimates at an areal scale. The other method is fitting an inhomogeneous Poisson process model using the Berman-Turner algorithm via customised R-packages which will provide rate estimates at point scale.

The innovative application of these two techniques will address questions on spatial distribution of cancers from different perspectives. Age-sex-adjusted rates are used in this study to remove bias resulting from population composition. The point process technique is introduced to cancer data to explore the correlation between the cancer incidence, mortality data and covariates. The spatial estimation risk maps are used to display incidence and mortality rates geographically, to potentially inform public health officials about planning decisions and provide evidencs concerning susceptible population distribution in Perth.

ATP Poisson kriging is a technique to estimate the attribute values at a point using observed values at SLAs by taking into account the Poisson nature of the disease data. This technique is within aggregate data. Point pattern analysis is carried out within the locations of cancer case events and detailed spatial trends for Perth cancers are achieved. The risk maps obtained using these two techniques provide information from different perspectives which presents further understanding to researchers and health officials.

For ATP Poisson kriging, the analysis of lung cancer, melanoma cancer and colorectal cancer is based on age-adjusted rates and age-sex-adjusted rates at SLA level. For the analysis of sex specific cancers (breast and prostate cancers), the spatial distributions are compared based on age-adjusted rates. To understand the spatial continuity and variability of cancer incidence and mortality for five cancers, the variance map and semivariogram plots are investigated. The study will also identify the difference between males and females in spatial distribution of cancer incidence. The analysis will be split into two parts: first, the data over the entire period will be considered, and second analysis for shorter time periods will be done to see how the distribution develops over time: three contiguous periods 1990-1995, 1996-2000 and 2001-2005. The following questions will be addressed:

What is the difference between ATP Poisson kriging estimates based on age-adjusted rates and age-sex-adjusted rates? What is the difference in spatial variability between incidence

and mortality? What is the difference in spatial variability between different cancers? How are different cancer mortality rates distributed across the Perth area in different periods?

The second methodology is spatial point pattern analysis. The object of this analysis is to explore the trend within the point process density and validate the fitted point process model using the techniques in (Baddeley et al. 2005). Some of the covariates which might affect cancer incidence and mortality will be examined in this study: age, sex, population density and susceptible population intensity which is computed based on cancer incidence. These potential covariates will be evaluated. The following major questions will be addressed within this analysis:

How to validate the model for Poisson processes and see whether it is appropriate?
What is an appropriate way to treat missing values?
Given a particular type of cancer which covariate is relatively strong?
Is there difference between incidence model fitting and mortality model fitting in terms of covariates?
What is the spatial intensity of the point pattern for different cancers?

The risk maps at different scales will be used for each cancer to explore and quantify any differences in the distribution of cancer risk.

1.5 Thesis outline

This thesis will present the ATP Poisson kriging analysis and spatial point pattern analysis for cancer incidence and mortality data for lung, melanoma, colorectal, breast and prostate cancers for the time period from 1990 to 2005.

Chapter 2 covers the background theories used in the thesis, including geostatistical theory and spatial point pattern theory. Geostatistical theory includes a discussion of the random function model, Poisson model, Poisson kriging and ATP Poisson kriging. In addition, the proof of theories used for data analysis is provided in detail. Spatial point pattern theory comprises the test for complete spatial randomness, model validation via the

inhomogeneous *K*-function, a generalisation of *K*-function, model fitting via the Berman-Turner algorithm, kernel smoothing, marked point patterns, empty space distances and Papangelou conditional intensity function.

Chapter 3 describes the methodology used in the thesis. In this chapter we discuss how the geostatistical analysis is carried out for different cancers by incorporating age or sex. Also we describe how the coordinates are converted for two types of methodologies. Then we focus on how the spatial point theory will be applied to five cancers.

Chapter 4 involves the data description. The first section focuses on the general background of the data used in the thesis and their preparation for analysis. This section includes an analysis of the survival rates for the different cancers. Raw cancer data are described in the second section. The fourth section introduces the census data to be used as covariates for the data modelling. The calculation of age-adjusted rates and age-sex-adjusted rates is presented in the fifth section. The last section is exploratory data analysis. We provide the adjusted rates and population at risk for five cancer types on which ATP Poisson kriging estimation are based. Lastly we introduce the covariates to be used in spatial point pattern analysis.

Chapter 5 and chapter 6 present geostatistical and spatial point pattern analysis respectively for lung, melanoma, colorectal, breast and prostate cancers.

Conclusions and recommendations for future work are presented in Chapter 7.

2 THEORY

This chapter presents geostatistical theory and spatial point pattern theory which provides the framework for the modelling of disease data.

2.1 The theory of Poisson kriging and its variants

The development of the theory of geostatistics started in the 1950s in the field of mining deposits evaluation (Sparks et al. 2011). Currently geostatistical techniques are successfully applied in diverse fields such as soil science, oceanography, geography, forestry and environmental control. Geostatistics allows the estimation of attribute values at unsampled locations using available data over a study area. Geostatistical techniques mainly depend on a statistical model that is based on the random function. A random function is a mapping from the study region to a family of spatially dependent random variables. The random function concept is introduced in section 2.1.1. In section 2.1.2, the kriging algorithm is presented. The Poisson model is introduced in section 2.1.3. In this section, we also derive the formulae for the first-order moment and second-order moment (variance) of rates variable. Section 2.1.4 presents the theory of Poisson kriging. In section 2.1.4.1 area-to-area (ATA) Poisson kriging is brought in. Area-to-point (ATP) Poisson kriging is a particular case of ATA and is introduced in the last section 2.1.4.2. Unless otherwise stated, the geostatistical theory given in this chapter is adapted mainly from Goovaerts (1997) and (2006a).

2.1.1 The Random Function Model

In geostatistical applications, the data are assumed to be a partial realisation of a random function (RF) because it is impossible to get the observed values at all locations of the study area. A RF Z is a collection of random variables on the study region A to which the laws of probability apply

$$\left\{Z(\mathbf{u}): \mathbf{u} \in A \subset \Re^N\right\} \tag{2.1}$$

where $\Re^N$ is N-dimensional Euclidean space, and the spatial index $\mathbf{u}$ varies continuously throughout the study area A. The value of an attribute at location $\mathbf{u}$ is denoted by $z(\mathbf{u})$ which can be regarded as a particular realisation of a random variable $Z(\mathbf{u})$ at the point $\mathbf{u}$. When $N = 2$, the RF Z is defined on a plane, and the Cartesian coordinates are often written as $\mathbf{u} = [x, y]^T$. The RF is characterised by its cumulative distribution function and the corresponding moments:

N-point cdf:

$$F(\mathbf{u}_1, \cdots, \mathbf{u}_N; z_1, \cdots, z_N) = \Pr ob\{Z(\mathbf{u}_1) \le z_1, \cdots, Z(\mathbf{u}_N) \le z_N\} \tag{2.2}$$

The expected value:

$$m(\mathbf{u}) = E(Z(\mathbf{u})) \tag{2.3}$$

The variance:

$$Var[Z(\mathbf{u})] = E\{[Z(\mathbf{u}) - m(\mathbf{u})]^2\} \tag{2.4}$$

The (two-point) covariance function:

$$C(\mathbf{u}, \mathbf{u}+\mathbf{h}) = E\{Z(\mathbf{u}) \cdot Z(\mathbf{u}+\mathbf{h})\} - E\{Z(\mathbf{u})\} \cdot E\{Z(\mathbf{u}+\mathbf{h})\} \tag{2.5}$$

The semivariogram function:

$$2\gamma(\mathbf{u}, \mathbf{u}+\mathbf{h}) = Var\{Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\} \tag{2.6}$$

The (two-point) covariance or variogram, is a random variable related to the same attribute z at two different locations. In general, the values of first-order moment: $E(Z(\mathbf{u}))$ and second-order moments: $Var[Z(\mathbf{u})]$, $C(\mathbf{u}, \mathbf{u}+\mathbf{h})$ and $\gamma(\mathbf{u}, \mathbf{u}+\mathbf{h})$ are location-dependent. The separation vector $\mathbf{h}$ accounts for distance and direction. The two-point covariance or semivariogram function is said to be isotropic if it only depends on the absolute distance between points, and the function is said to be anisotropic when it depends on both separation distance and direction.

A RF Z is said to be *strictly (or strongly) stationary* if the spatial distribution is invariant under translation of the coordinates, that is:

$$\begin{aligned}&\text{Prob}\{Z(\mathbf{u}_1)\le z_1, Z(\mathbf{u}_2)\le z_2,\cdots,Z(\mathbf{u}_k)\le z_k\}=\\&\quad\text{Prob}\{Z(\mathbf{u}_1+\mathbf{h})\le z_1, Z(\mathbf{u}_2+\mathbf{h})\le z_2,\cdots,Z(\mathbf{u}_k+\mathbf{h})\le z_k\}\end{aligned} \tag{2.7}$$

for all k and **h**. In practice it is hard to confirm that a RF is strictly stationary from limited sample data so this assumption is normally relaxed.

A RF Z is said to be *second-order (or weakly) stationary* if the expected value (mean) $E(Z(\mathbf{u}))$ exists and is invariant within A; and the covariance exists and depends only on the separation vector (lag) **h**.

Second-order stationarity implies that $E(Z(\mathbf{u}))=m$ and $C(\mathbf{h})=E\{Z(\mathbf{u})\cdot Z(\mathbf{u}+\mathbf{h})\}-m^2$ for all $\mathbf{u}\in A$. In other words, a RF with such characteristics has constant expected value and finite covariance which depends only on the separation vector **h**. If $Z(\mathbf{u})$ is not second-order stationary, the increments $Z(\mathbf{u}+\mathbf{h})-Z(\mathbf{u})$ might be. A RF that has this characteristic is said to be *intrinsic stationary.* That is, an intrinsic stationary RF has a constant mean and a semivariogram function which does not depend on absolute coordinates: $E(Z(\mathbf{u}))=m$ and $\gamma(\mathbf{h})=\frac{1}{2}Var\{Z(\mathbf{u})-Z(\mathbf{u}+\mathbf{h})\}$, but might not have a finite variance. Strict stationarity implies second-order stationarity but the converse is not true, and second-order stationarity implies intrinsic stationarity but the converse is not true. The covariance function $C(\mathbf{h})$ may not exist when the random function is intrinsic rather than second-order stationary. Many geostatisticians prefer the semivariogram to covariances when estimating these functions based on the observed data (Schabenberger and Gotway 2005, §4.2).

In geostatistics, the experimental semivariogram $\gamma(\mathbf{h})$ reflects one of the spatial data attributes: autocorrelation. The experimental semivariogram is computed as:

$$\gamma(\mathbf{h})=\frac{1}{2N(\mathbf{h})}\sum_{\alpha=1}^{N(\mathbf{h})}\left[z(\mathbf{u}_\alpha)-z(\mathbf{u}_\alpha+\mathbf{h})\right]^2 \tag{2.8}$$

where $z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})$ is an **h**-increment of attribute z and $z(\mathbf{u}_\alpha)$ is the known sample value of the attribute z at location $\mathbf{u}_\alpha$. The semivariogram defined in equation (2.8) is also called the traditional semivariogram.

Kriging estimation is based on a covariance model inferred from the sample data via the experimental semivariograms. The models used in this study are the the nugget, spherical, exponential and cubic models. Their definitions are as follows:

Nugget effect model. This function models a discontinuity at the origin when the variogram is not differentiable. The nugget effect may indicate measurement error, micro-scale variation or purely random behaviour.

$$g(h) = \begin{cases} 0 & \text{if } h = 0 \\ \text{b} & \text{otherwise} \end{cases}$$

where b is positive. This model is bounded and the bound is referred to as the "sill".

Spherical model. This function models linear behaviour at the origin and is bounded by the sill b. The model reaches the sill at actual range a.

$$g(h) = \text{Sph}\left(\frac{h}{a}\right) = \begin{cases} \text{b}\left(1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right) & \text{if } h \le a \\ \text{b} & \text{otherwise} \end{cases}$$

Exponential model. This function is bounded by the sill b. This model reaches the sill asymptotically. The practical range a is the distance at which the semivariance is equal to 95% of the sill.

$$g(h) = \text{b}\left(1 - \exp\left(\frac{-3h}{a}\right)\right)$$

Cubic model. This function models parabolic behaviour at the origin and is bounded by the sill b.

$$g(h)=\begin{cases} \mathrm{b}\left(7\left(\frac{h}{a}\right)^2-\frac{35}{4}\left(\frac{h}{a}\right)^3+\frac{7}{2}\left(\frac{h}{a}\right)^5-\frac{3}{4}\left(\frac{h}{a}\right)^7\right) & \text{if } 0\le h<a \\ \mathrm{b} & \text{otherwise} \end{cases}$$

2.1.2 Kriging

Kriging is geostatistical techniques to estimate the value of a RF at a unsampled location by using observed values at contiguous sampled locations. Kriging is a linear regression estimation algorithm. Simple kriging (SK) and ordinary kriging (OK) are classical types of kriging. Ordinary kriging considers the mean unknown but constant in the local neighbourhood. It is used more often than simple kriging which assumes the mean is known and constant throughout the study area. Ordinary kriging and simple kriging can estimate the attribute values at unsampled locations. The simple kriging unbiased estimator is defined as:

$$Z^*_{SK}(\mathbf{u})=\sum_{\alpha=1}^{n(\mathbf{u})}\lambda_\alpha^{SK}(\mathbf{u})[Z(\mathbf{u}_\alpha)-m]+m=\sum_{\alpha=1}^{n(\mathbf{u})}\lambda_\alpha^{SK}(\mathbf{u})Z(\mathbf{u}_\alpha)+\left[1-\sum_{\alpha=1}^{n(\mathbf{u})}\lambda_\alpha^{SK}(\mathbf{u})\right]m \qquad (2.9)$$

where m is the known stationary mean and $\lambda_\alpha^{SK}(\mathbf{u})$ are the simple kriging weights to be determined. Let $R(\mathbf{u})=Z(\mathbf{u})-m$ (and hence $R(\mathbf{u}_\alpha)=Z(\mathbf{u}_\alpha)-m$), then the error variance can be expressed as:

$$\begin{aligned}\sigma_E^2(\mathbf{u}) &= Var\{Z^*_{SK}(\mathbf{u})-Z(\mathbf{u})\} \\ &= Var\{[Z^*_{SK}(\mathbf{u})-m]-[Z(\mathbf{u})-m]\} \\ &= Var\left\{\sum_{\alpha=1}^{n(\mathbf{u})}\lambda_\alpha^{SK}(\mathbf{u})R(\mathbf{u}_\alpha)-R(\mathbf{u})\right\} \\ &= Var\{R^*_{SK}(\mathbf{u})-R(\mathbf{u})\} \\ &= \sum_{\alpha=1}^{n(\mathbf{u})}\sum_{\beta=1}^{n(\mathbf{u})}\lambda_\alpha^{SK}(\mathbf{u})\lambda_\beta^{SK}(\mathbf{u})C_R(\mathbf{u}_\alpha-\mathbf{u}_\beta)+C_R(0)-2\sum_{\alpha=1}^{n(\mathbf{u})}\lambda_\alpha^{SK}(\mathbf{u})C_R(\mathbf{u}_\alpha-\mathbf{u})\end{aligned} \qquad (2.10)$$

To determine the weights which minimize the error variance, we have to solve:

$$\frac{\partial \sigma_E^2(\mathbf{u})}{\partial \lambda_\alpha^{SK}(\mathbf{u})} = 2\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta^{SK}(\mathbf{u}) C_R(\mathbf{u}_\alpha - \mathbf{u}_\beta) - 2C_R(\mathbf{u}_\alpha - \mathbf{u}) = 0, \qquad \alpha = 1, 2, ..., n(\mathbf{u}) \tag{2.11}$$

and the simple kriging variance (the minimum error variance) is deduced as (by replacing (2.11) into (2.10)):

$$\sigma_{SK}^2(\mathbf{u}) = C(0) - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{SK}(\mathbf{u}) C(\mathbf{u}_\alpha - \mathbf{u}) \tag{2.12}$$

The ordinary kriging estimator is defined similarly to the simple kriging estimator except that the mean is locally constant but unknown across the study area:

$$Z_{OK}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OK}(\mathbf{u}) Z(\mathbf{u}_\alpha) \text{ with } \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OK}(\mathbf{u}) = 1 \tag{2.13}$$

The minimization of the error variance needs a function of the data weights $\lambda_\alpha^{OK}(\mathbf{u})$ and a Lagrange parameter $2\mu_{OK}(\mathbf{u})$:

$$L\left(\lambda_\alpha^{OK}(\mathbf{u}), \alpha = 1, ..., n(\mathbf{u}); 2\mu_{OK}(\mathbf{u})\right) = \sigma_E^2(\mathbf{u}) + 2\mu_{OK}(\mathbf{u})\left[\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OK}(\mathbf{u}) - 1\right] \tag{2.14}$$

The optimal weights are obtained by setting each of the partial first derivatives to zero. In a similar way, the ordinary kriging system is given by:

$$\frac{\partial \sigma_E^2(\mathbf{u})}{\partial \lambda_\alpha^{OK}(\mathbf{u})} = 2\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta^{OK}(\mathbf{u}) C\ (\mathbf{u}_\alpha - \mathbf{u}_\beta) - 2C(\mathbf{u}_\alpha - \mathbf{u}) + 2\mu_{OK}(\mathbf{u}) = 0, \qquad \alpha = 1, 2, ..., n(\mathbf{u})$$

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OK}(\mathbf{u}) = 1$$

and the ordinary kriging variance is calculated as:

$$\sigma^2_{OK}(\mathbf{u}) = C(0) - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OK}(\mathbf{u}) C(\mathbf{u}_\alpha - \mathbf{u}) - \mu_{OK}(\mathbf{u}) \quad (2.15)$$

where $\mu_{OK}(\mathbf{u})$ is the Lagrange multiplier.

2.1.3 Poisson model

2.1.3.1 Model definition

The Poisson distribution plays an important role in the study of discrete (or categorical) data. Disease data are always available by case count. In the geostatistical study, the given units (e.g. suburb, counties, states, SLA) have a common geographic reference such as their centroids $\mathbf{u} = (x, y)$. At each unit with centroid $\mathbf{u}_\alpha$, the recorded disease count $\mathrm{d}(\mathbf{u}_\alpha)$ can be regarded as a particular realisation of a random variable $\mathrm{D}(\mathbf{u}_\alpha)$. For a given number of K locations $\mathbf{u}_\alpha$ belonging to the study region $\mathcal{D}$, the model for the random variable $\mathrm{D}(\mathbf{u}_\alpha)$ is defined by

$$\mathrm{D}(\mathbf{u}_\alpha) \,|\, \mathrm{R}(\mathbf{u}_\alpha) = \mathrm{Poisson}\left(\mathrm{n}(\mathbf{u}_\alpha)\mathrm{R}(\mathbf{u}_\alpha)\right) \quad \alpha = 1,..., \mathrm{K} \quad (2.16)$$

where $\mathrm{D}(\mathbf{u}_\alpha) \,|\, \mathrm{R}(\mathbf{u}_\alpha)$ is Poisson distributed with a parameter $\mathrm{n}(\mathbf{u}_\alpha)\ \mathrm{R}(\mathbf{u}_\alpha)$. The random variable $\mathrm{R}(\mathbf{u}_\alpha)$ and $\mathrm{n}(\mathbf{u}_\alpha)$ represent the local disease risk and the size of the population at risk at location $\mathbf{u}_\alpha$ respectively. The observed disease rates are denoted as $\mathrm{z}(\mathbf{u}_\alpha) = \mathrm{d}(\mathbf{u}_\alpha) \big/ \mathrm{n}(\mathbf{u}_\alpha)$. Conditionally on $\mathrm{R}(\mathbf{u}_\alpha)$, the rates variables $\mathrm{Z}(\mathbf{u}_\alpha)$ are assumed to be independent since the correlation between observed rates is normally dependent on population composition or local individual risks. The random function R is modelled as a positive stationary random function with mean *m*, variance σ^2_{R} and covariance $\mathrm{C_R}$. To simplify notation, $\hat{\mathrm{R}}_{PK}(\mathbf{u}_\alpha), \mathrm{R}(\mathbf{u}_\alpha), \mathrm{Z}(\mathbf{u}_\alpha), \mathrm{z}(\mathbf{u}_\alpha)$ and $\mathrm{n}(\mathbf{u}_\alpha)$ in this study will be abbreviated in the following as $\hat{\mathrm{R}}_\alpha$, R_α, Z_α, z_α and n_α respectively. In any kriging system, C_{ij} denotes the covariance function $C_R(\mathbf{u}_i - \mathbf{u}_j)$.

2.1.3.2 Expectation and variances of rates variables $Z(\mathbf{u}_\alpha)$

The conditional mean and variance of rates variable Z_α are defined in (Goovaerts 2005c):

$$E[Z_\alpha | R_\alpha] = R_\alpha \tag{2.17}$$

$$Var[Z_\alpha | R_\alpha] = R_\alpha / n_\alpha \tag{2.18}$$

Since Z_α and R_α are two random variables, and the expectations exist, we have

$$E(Z_\alpha) = E(E(Z_\alpha | R_\alpha)) \tag{2.19}$$

and

$$Var[Z_\alpha] = E[Var(Z_\alpha | R_\alpha)] + Var[E(Z_\alpha | R_\alpha)] \tag{2.20}$$

Proof of (2.19): Let $f(z_\alpha, r_\alpha)$ be the joint pdf of Z_α and R_α. According to the definition, we have

$$\begin{aligned} E[Z_\alpha] &= \int\int z_\alpha f(z_\alpha, r_\alpha) dz_\alpha d r_\alpha \\ &= \int \left[\int z_\alpha f(z_\alpha | r_\alpha) dz_\alpha \right] f_{R_\alpha}(r_\alpha) d r_\alpha \\ &= \int E[Z_\alpha | r_\alpha] f_{R_\alpha}(r_\alpha) d r_\alpha \\ &= E(E(Z_\alpha | R_\alpha)) \end{aligned}$$

Proof of (2.20): With (2.19), we have

$$\begin{aligned} Var[Z_\alpha] &= E(Z_\alpha^2) - [E(Z_\alpha)]^2 \\ &= E\left[E(Z_\alpha^2 | R_\alpha)\right] - [E(E(Z_\alpha | R_\alpha))]^2 \\ &= E\left[E(Z_\alpha^2 | R_\alpha)\right] - E\left(\left[E(Z_\alpha | R_\alpha)\right]^2\right) + E\left(\left[E(Z_\alpha | R_\alpha)\right]^2\right) - \left[E(E(Z_\alpha | R_\alpha))\right]^2 \\ &= E[Var(Z_\alpha | R_\alpha)] + Var[E(Z_\alpha | R_\alpha)] \end{aligned}$$

According to equations (2.19) and (2.20), the unconditional mean and variance can be obtained directly from (2.17) and (2.18):

$$E[Z_\alpha] = E[R_\alpha] = m \tag{2.21}$$

$$\begin{aligned} Var[Z_\alpha] &= E\left[Var\left(Z_\alpha \middle| R_\alpha\right)\right] + Var\left[E\left(Z_\alpha \middle| R_\alpha\right)\right] \\ &= E[R_\alpha / n_\alpha] + Var[R_\alpha] = \sigma_R^2 + m/n_\alpha \end{aligned} \tag{2.22}$$

As for the conditional mean value of observed rates at independent locations $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$, it is calculated as:

$$\begin{aligned} E\left[Z_\alpha Z_\beta \middle| R\right] &= Cov\left(Z_\alpha, Z_\beta | R\right) + E\left[Z_\alpha \middle| R_\alpha\right] E\left[Z_\beta \middle| R_\beta\right] \\ &= \delta_{\alpha\beta} Var\left[Z_\alpha \middle| R_\alpha\right] + R_\alpha R_\beta = \delta_{\alpha\beta} R_\alpha / n_\alpha + R_\alpha R_\beta \end{aligned} \tag{2.23}$$

where $\delta_{\alpha\beta}$ is 1 when $\alpha = \beta$ and 0 otherwise.

2.1.4 Poisson kriging

Although standard kriging methods are fairly common, they are not suitable for data having a discrete distribution like disease data. Poisson kriging accounts for the Poisson nature of the count data but ignores the spatial supports of geographic entity. Each entity is assumed to have point support, and be located at its geographic centroid. The Poisson kriging estimator is derived from the OK estimator. The Poisson kriging disease risk is estimated as a linear combination of K neighboring data:

$$\hat{R}_{PK}(\mathbf{u}_\alpha) = \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha) z(\mathbf{u}_i) \tag{2.24}$$

where $z(\mathbf{u}_i)$ is the observed disease rate. In order to make sure this estimator is unbiased, we compute its expectation by application of expression (2.17):

$$E\left[\hat{R}_\alpha \middle| R_i\right] = \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha) E\left(Z_i \middle| R_i\right) = \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha)\, R_i ,$$

referring back to (2.19), since

$$\mathrm{E}\left[\hat{\mathrm{R}}_{\alpha}\right]=\mathrm{E}\,\mathrm{E}\left[\hat{\mathrm{R}}_{\alpha}\middle|\mathrm{R}_{\mathrm{i}}\right]=\sum_{i=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\mathrm{E}\left(\mathrm{R}_{\mathrm{i}}\right)=\mathrm{m}\sum_{i=1}^{K}\lambda_i(\mathbf{u}_{\alpha}),$$

the condition for the estimator to be unbiased is:

$$\sum_{i=1}^{K}\lambda_i(\mathbf{u}_{\alpha})=1 \tag{2.25}$$

Similarly, to minimise the mean square error of prediction under the unbiasedness condition (2.25), we have to compute $\mathrm{E}\left[\left(\hat{\mathrm{R}}_{\alpha}-\mathrm{R}_{\alpha}\right)^2\right]$, since

$$\mathrm{Var}\left(\hat{\mathrm{R}}_{\alpha}-\mathrm{R}_{\alpha}\right)=\mathrm{E}\left[\left(\hat{\mathrm{R}}_{\alpha}-\mathrm{R}_{\alpha}\right)^2\right]$$

The following calculation is derived by application of equations (2.23) and (2.25):

$$\begin{aligned}
&\mathrm{E}\left[\left(\hat{\mathrm{R}}_{\alpha}-\mathrm{R}_{\alpha}\right)^2\middle|\mathrm{R}\right]\\
&=\mathrm{E}\left[\left(\sum_{i=1}^{K}\lambda_i(\mathbf{u}_{\alpha})z_{\mathrm{i}}-\mathrm{R}_{\alpha}\right)^2\middle|\mathrm{R}\right]\\
&=\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\lambda_j(\mathbf{u}_{\alpha})\mathrm{E}\left(z_{\mathrm{i}}z_{\mathrm{j}}\middle|\mathrm{R}\right)+\mathrm{R}_{\alpha}{}^2-2\mathrm{R}_{\alpha}\sum_{i=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\mathrm{E}\left(z_{\mathrm{i}}\middle|\mathrm{R}\right)\\
&=\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\lambda_j(\mathbf{u}_{\alpha})\left[\delta_{ij}\times\mathrm{R}_{\mathrm{i}}\middle|\mathrm{n}_{\mathrm{i}}+\mathrm{R}_i\mathrm{R}_j\right]+\mathrm{R}_{\alpha}{}^2-2\mathrm{R}_{\alpha}\sum_{i=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\mathrm{R}_i
\end{aligned}$$

Then, by deconditioning:

$$\begin{aligned}
\mathrm{E}\left[\left(\hat{\mathrm{R}}_{\alpha}-\mathrm{R}_{\alpha}\right)^2\right]&=\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\lambda_j(\mathbf{u}_{\alpha})\delta_{ij}\,\mathrm{m}\big/\mathrm{n}_{\mathrm{i}}+\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\lambda_j(\mathbf{u}_{\alpha})\left(\mathrm{m}^2+\mathrm{C}_{\mathrm{ij}}\right)\\
&\quad+\sigma_{\mathrm{R}}^2+\mathrm{m}^2-2\sum_{i-1}^{K}\lambda_i(\mathbf{u}_{\alpha})\left(\mathrm{m}^2+\mathrm{C}_{\mathrm{i}\alpha}\right)\\
&=\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\lambda_j(\mathbf{u}_{\alpha})\delta_{ij}\,\mathrm{m}\big/\mathrm{n}_{\mathrm{i}}+\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i(\mathbf{u}_{\alpha})\lambda_j(\mathbf{u}_{\alpha})\mathrm{C}_{\mathrm{ij}}\\
&\quad+\sigma_{\mathrm{R}}^2-2\sum_{i-1}^{K}\lambda_i(\mathbf{u}_{\alpha})\,\mathrm{C}_{\mathrm{i}\alpha}
\end{aligned} \tag{2.26}$$

By minimising the expression (2.26) on λ_i's with the unbiasedness condition (2.25), the following $(K+1)$ equations are obtained:

$$\begin{aligned} &\sum_{j=1}^{K} \lambda_j(\mathbf{u}_\alpha)\left[C_{ij} + \delta_{ij}\frac{\mathrm{m}}{\mathrm{n_i}}\right] + \mu(\mathbf{u}_\alpha) = C_{i\alpha} \quad i = 1,...,K \\ &\sum_{j=1}^{K} \lambda_j(\mathbf{u}_\alpha) = 1 \end{aligned} \tag{2.27}$$

where $\mu(\mathbf{u}_\alpha)$ is a Lagrange multiplier. The term, $\frac{\mathrm{m}}{\mathrm{n_i}}$, accounts for variability resulting from the population size. The prediction variance associated with the estimate (2.24) is computed using the traditional formula for the ordinary kriging variance (2.15):

$$\sigma^2_{PK}(\mathbf{u}_\alpha) = C_R(0) - \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha) C_R(\mathbf{u}_i - \mathbf{u}_\alpha) - \mu(\mathbf{u}_\alpha) \tag{2.28}$$

The covariance of the unknown risk, $C_R(\mathrm{h})$, or its equivalent semivariogram $\gamma_R(\mathbf{h}) = C_R(0) - C_R(\mathbf{h})$, is estimated according to Monestiez et al. (2004; 2006) based on the traditional semivariogram (2.8):

$$\hat{\gamma}_{\mathrm{R}}(\mathbf{h}) = \frac{1}{2\sum_{\alpha=1}^{\mathrm{N(h)}} \frac{\mathrm{n}(\mathbf{u}_\alpha)\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})}{\mathrm{n}(\mathbf{u}_\alpha)+\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})}} \sum_{\alpha=1}^{\mathrm{N(h)}} \left\{ \frac{\mathrm{n}(\mathbf{u}_\alpha)\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})}{\mathrm{n}(\mathbf{u}_\alpha)+\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})} \left[\mathrm{z}(\mathbf{u}_\alpha) - \mathrm{z}(\mathbf{u}_\alpha+\mathbf{h})\right]^2 - \mathrm{m} \right\} \tag{2.29}$$

The calculations of semivariogram in the geostatistical analysis are carried out using this equation within software SpaceStat.

2.1.4.1 Area-to-Area (ATA) Poisson kriging

When geographic units vary greatly by size and shape, it is important to take into account their spatial supports. Area-to-Area (ATA) kriging was first introduced by Kyriakidis

(2004). Goovaerts (2006a) generalised this technique and introduced ATA Poisson kriging for the modelling of disease data. This terminology refers to the situation when the measurement and prediction supports are areas rather than points. Area-to-area spatial interpolation is used to predict an areal value $z(v_\alpha)$ using K areal data $\{z(v_i), i = 1,..., K\}$:

$$z^*(v_\alpha) = \sum_{i=1}^{K} \lambda_i(v_\alpha) z(v_i) \tag{2.30}$$

where the areal supports v_i are disjoint. The weights $\lambda_i(v_\alpha)$ are computed to ensure the minimization of prediction mean square error under the condition of the unbiasedness of $z^*(v_\alpha)$, and they are the solution of the following equations:

$$\begin{aligned} &\sum_{j=1}^{K} \lambda_j(v_\alpha)\left[\overline{C}(v_i, v_j) + \delta_{ij}\frac{m}{n(v_i)}\right] + \mu(v_\alpha) = \overline{C}(v_i, v_\alpha) \quad i = 1,..., K \\ &\sum_{i=1}^{K} \lambda_i(v_\alpha) = 1 \end{aligned} \tag{2.31}$$

where $\overline{C}(v_i, v_j)$ is the covariance between areal suppots v_i and v_j. It is the average of point covariance values $C(\mathbf{s}-\mathbf{t})$ corresponding to vectors $\mathbf{s}-\mathbf{t}$ formed by all possible pairs of points $\mathbf{s} \in v_i$ and $\mathbf{t} \in v_j$ respectively (Kyriakidis 2004). The difference between the system of linear equations in (2.27) and (2.31) is that the prediction and measurement supports is a point support in (2.27) while in (2.31) it is an areal support.

2.1.4.2 Area-to-Point (ATP) Poisson kriging

ATP kriging is the special case of ATA kriging where the prediction support is reduced to a point $\mathbf{u}_\alpha$. ATP Poisson kriging will be applied to cancer data in this study. Area-to-point spatial interpolation is to predict any point value $z(\mathbf{u}_\alpha)$ using K areal data $\{z(v_i), i = 1,...,K\}$:

$$z^*(\mathbf{u}_\alpha) = \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha) z(v_i) \tag{2.32}$$

The weights $\lambda_i(\mathbf{u}_\alpha)$ are computed to ensure that the prediction mean square error is minimised under the condition of the unbiasedness of $z^*(\mathbf{u}_\alpha)$, and they are the solution of the following equations:

$$\begin{aligned} &\sum_{j=1}^{K} \lambda_j(\mathbf{u}_\alpha)\left[\overline{C}(v_i, v_j) + \delta_{ij}\frac{m}{n(v_i)}\right] + \mu(\mathbf{u}_\alpha) = \overline{C}(v_i, \mathbf{u}_\alpha) \quad i = 1,...,K \\ &\sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha) = 1 \end{aligned} \tag{2.33}$$

where the covariance $\overline{C}(v_i, \mathbf{u}_\alpha)$ between areal supports v_i and the point $\mathbf{u}_\alpha$ is the average of point covariance values $C(s - \mathbf{u}_\alpha)$ corresponding to vectors $s - \mathbf{u}_\alpha$ formed by all possible pairs of points $\mathbf{s} \in v_i$ and $\mu(\mathbf{u}_\alpha)$ is the Lagrange multiplier, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The size of the population at risk in area v_i is denoted by $n(v_i)$. The term $\frac{m}{n_i}$ accounts for variability resulting from the population size. That is, the more reliable the data (i.e. measured over larger populations) are, the more weight they get. The variance is calculated as:

$$\sigma^2_{PK}(\mathbf{u}_\alpha) = C(0) - \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha)\overline{C}(v_i, \mathbf{u}_\alpha) - \mu(\mathbf{u}_\alpha) \tag{2.34}$$

where $C(0) = \mathrm{Var}(Z(\mathbf{u}_\alpha))$. Since the population is not uniformly distributed across the study area and disease rates computed from sparsely populated geographic units tend to be less reliable. This effect, known as "small number problem" (Waller and Gotway 2004). To filter the noise caused by small number problem, it is important to assign a weight to an area's observed rate based on the size of the population at risk. When the area is sparsely populated the weight is small, when the area is densely populated the corresponding weight is large. The calculation of weights $\lambda_i(\mathbf{u}_\alpha)$ based on equation (2.33) can be corrected for small number problem.

To apply ATP Poisson kriging to the data, it is important to know point-support semivariogram. Derivation of a point-support semivariogram from the regularized experimental semivariogram computed from areal data is called "deconvolution" (Goovaerts 2008). Using iterative procedure, it seeks the point support model that minimizes the difference between the theoretically regularized model and the model fitted to data. The deconvolution is based on the following relationship between the regularized semivariogram model $\gamma_v(\mathbf{h})$ and $\bar{\gamma}(v, v_h)$ which is a function of the unknown point-support model $\gamma(\mathbf{h})$:

$$\gamma_v(\mathbf{h}) = \bar{\gamma}(v, v_h) - \bar{\gamma}_h(v, v) \tag{2.35}$$

where $\bar{\gamma}_h(v, v)$ is the within-area semivariogram as a function of the separation distance *h*.

$\bar{\gamma}(v, v_h)$ is the mean value of the point-support semivariogram between an arbitrary point in the support *v* and another in the translated support v_h and this area-to-area semivariogram value is estimated as:

$$\bar{\gamma}_h(v, v_h) = \frac{1}{N(h)} \sum_{\alpha=1}^{N(h)} \bar{\gamma}_h(v_\alpha, v_{\alpha+h}) \tag{2.36}$$

where $N(h)$ is the number of pairs of areas separated by the distance h. The semivariogram value $\bar{\gamma}_h(v_\alpha, v_{\alpha+h})$ between any two areas v_α and $v_{\alpha+h}$ is calculated as:

$$\bar{\gamma}_h(v_\alpha, v_{\alpha+h}) = \frac{1}{P_\alpha P_{\alpha+h}} \sum_{s=1}^{P_\alpha} \sum_{s'=1}^{P_{\alpha+h}} \gamma(\mathbf{u}_s, \mathbf{u}_{s'}) \quad (2.37)$$

where P_α and $P_{\alpha+h}$ are the number of points $\mathbf{u}_s$ and $\mathbf{u}_{s'}$ used to discretize the two areas respectively. The within-area semivariogram $\bar{\gamma}_h(v,v)$ is calculated using equation (2.37) with h=0.

2.2 Spatial point pattern theory

Spatial data, also known as geospatial data, are associated with three components: features, supports and attributes (Waller and Gotway 2004, p.38). In this study, the spatial feature is the point, that is, the location of cancer incidence and mortality characterised by latitude/longitude. Points have the smallest support. Unless otherwise stated, the spatial point pattern theory presented in this chapter is mainly from Baddeley (2008).

2.2.1 The spatial point pattern

The observed point pattern is denoted by **x** and a random point process is denoted by **X**. Every point in a spatial point pattern may carry extra information called a "mark" and this point pattern is called a marked point pattern. In spatial point pattern analysis, the observed point pattern **x** is treated as a realisation of a random point process **X** in two-dimensional space. The data comprise a spatial point pattern **x** observed in a bounded region *W* of the plane $\Re^2$. The spatial point pattern is denoted by

$$\mathbf{x} = \{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\},$$

where the number of points, $n \geq 0$, is not fixed and each $\mathbf{u}_i$ is an observed point in *W*.

Each point $\mathbf{u}$ corresponds to its coordinates in the plane $\Re^2$. Let $N(B)$ be the number of points in any given sub-region *B* of *W*. The (first-order) intensity of a point process, denoted by $\lambda(\mathbf{u})$, is defined (Marcon and Puech 2003) as:

$$\lambda(\mathbf{u}) = \lim_{dB \to 0} \frac{E[N(dB)]}{dB}$$

where dB is the area of an infinitesimal disk centred at point $\mathbf{u}$ and E denotes the mathematical expectation. The point process is homogeneous when $\lambda(\mathbf{u}) = \lambda$ for all $\mathbf{u} \in W$. The point process is inhomogeneous when its intensity $\lambda(\mathbf{u})$ is not constant. To capture the spatial dependency of event locations, (second-order) density is defined (Schabenberger and Gotway 2005):

$$\lambda_2(\mathbf{u}_i, \mathbf{u}_j) = \lim_{|dx_i| \to \mathbf{0}, |dx_j| \to \mathbf{0}} \frac{E[N(d\mathbf{u}_i)\, N(d\mathbf{u}_j)]}{|d\mathbf{u}_i|\, |d\mathbf{u}_j|}$$

When the point pattern is a homogeneous Poisson process, $\lambda_2(\mathbf{u}_i, \mathbf{u}_j)$ only depends on distance, $\lambda_2(\mathbf{u}_i, \mathbf{u}_j) = \lambda_2^*(\|\mathbf{u}_i - \mathbf{u}_j\|) = \lambda_2^*(h)$.

2.2.2 Inhomogeneous K-function

A point process is a uniform Poisson point process in the plane when its intensity is constant. The uniform Poisson point process is also called Complete Spatial Random (CSR). In point pattern analysis, the assumption of CSR is normally taken as the 'null' hypothesis. The classical test for CSR is the chi-square approximation test based on quadrat counts which is described by Diggle (2003). In this approximation test, the domain W is assumed to be bounded by a rectangle, which is partitioned into equal-sized non-overlapping subregions ('quadrats') with r rows and c columns: B_{11}, $B_{12}, \cdots, B_{rc}$. The number of observed points falling in each subregion is denoted by $n_{ij} = n(\mathbf{x} \cap B_{ij})$ for $1 \le i \le r, 1 \le j \le c$. Under the null hypothesis of CSR the counts n_{ij} are independent and identically distributed (i.i.d) Poisson random variables with the same expected value in each quadrat. The Pearson chi-square goodness-of-fit test performs very well when the expected number of events per quadrat exceeds 1 and $rc > 6$ (Diggle 1983 p.33).

Ripley's *K*-function (1977) is used to describe spatial structures of a homogeneous point process. The *K*-function is defined as:

$$K(r) = \int_0^r g(h) 2\pi h \, dh$$

where $g(h) = \frac{\lambda_2(h)}{\lambda^2}$. For a homogeneous Poisson process, $\lambda_2(h) = \lambda^2$ for all values of *h*, so $K(r) = \pi r^2$. A clustered pattern and an inhibited pattern are defined by Schabenberger and Gotway (2005) as: a point pattern is clustered when the average distance between an event and its nearest-neighbour event is smaller than the same average distance in a CSR pattern; a point pattern is called inhibited when the average distance between an event and its nearest neighbour is larger than expected under a CSR pattern. For a clustered point pattern, $g(h) > 1$ indicates $K(r) > \pi r^2$. That is, the chance of discovering a point within distance *r* is greater than the chance of discovering a point at any other location with the same area in the domain. For an inhibited point pattern, $g(h) < 1$, and so $K(r) < \pi r^2$. The product $\lambda K(r)$ is the expected number of points other than point **u** within the circle of radius *r*, centred on **u** (Baddeley 2008):

$$\lambda K(r) = \mathbb{E}\left[n(\mathbf{X} \cap \mathrm{b}(\mathbf{u}, r) \setminus \{\mathbf{u}\}) \mid \mathbf{u} \in \mathbf{X} \right]$$

However, Ripley's *K*-function is used for homogeneous point processes and it does not consider the weighting of points in the calculation. Consequently, the *K*-function was generalised by Baddeley (2000) to the inhomogeneous *K*-function. If λ_i is the intensity function of the point process **X** at location $\mathbf{u}_i$ and each point $\mathbf{u}_i$ is weighted by the inverse of the intensity function at this point, the inhomogeneous *K*-function is defined as:

$$K_{inhom}(r) = E\left[\frac{1}{\lambda(\mathbf{u})} \sum_{\mathbf{u}_i \in \mathbf{X}} \frac{1}{\lambda(\mathbf{u}_i)} \mathbf{1}\{0 < \|\mathbf{u} - \mathbf{u}_i\| \le r\} \,\middle|\, \mathbf{u} \in \mathbf{X} \right] \qquad (2.38)$$

where $\mathbf{1}(\cdot)$ is the indicator function. If the process is homogeneous, then $\lambda(\mathbf{u}_i) = \lambda$ and $K_{inhom}(r) = K(r)$, since $K_{inhom}(r) = E\left[\mathbf{1}\{0 < \|\mathbf{u} - \mathbf{u}_i\| \le r\} \,|\, \mathbf{u} \in \mathbf{X}\right] = \pi r^2$. According to (2.38), we have:

$$K_{inhom}(r) = \frac{1}{\text{area}(W)} E\left[\sum_{\mathbf{u}\in\mathbf{X}} \sum_{\mathbf{u}_i\in\mathbf{X}} \frac{\mathbf{1}}{\lambda(\mathbf{u})\lambda(\mathbf{u}_i)} \mathbf{1}\{\mathbf{0} < \|\mathbf{u} - \mathbf{u}_i\| \le r\}\right]$$
$$= \frac{1}{\text{area}(W)} \int_{\mathbf{u}\in W} \int_{0<\|\mathbf{u}-\mathbf{u}_i\|\le r} \frac{\rho^{(2)}(\mathbf{u}_i, \mathbf{u})}{\lambda(\mathbf{u}_i)\,\lambda(\mathbf{u})} d\mathbf{u}_i\, d\mathbf{u}$$

where $g(\mathbf{u}, \mathbf{u}_i) = \frac{\rho^{(2)}(\mathbf{u}_i, \mathbf{u})}{\lambda(\mathbf{u}_i)\,\lambda(\mathbf{u})}$, $\mathbf{u}_i, \mathbf{u} \in W \subset \Re^2$ is the pair correlation function of W and "$\rho^{(2)}(\mathbf{u}_i, \mathbf{u}) d\mathbf{u}_i\, d\mathbf{u}$ is the probability that W has a point in each of two infinitesimally small discs with centers $\mathbf{u}_i$, $\mathbf{u}$ and volumes $d\mathbf{u}_i$, $d\mathbf{u}$ respectively" (Baddeley et al. 2000). As for an inhomogeneous Poisson process function $g \equiv 1$ (Baddeley et al. 2000), so $\mathrm{K}_{\text{inhom}}(r) = \frac{1}{\text{area}(W)} \int_W \int_{\pi r^2} d\mathbf{u}_j\, d\mathbf{u} = \pi r^2$.

2.2.3 Berman-Turner algorithm

The inhomogeneous Poisson model with covariates included (Baddeley 2008) has the parametric form:

$$\lambda_\theta(\mathbf{u}) = \exp(\beta_0 + \beta_1 \times Z(\mathbf{u})) \tag{2.39}$$

where θ, a vector of unknown parameters β_0 and β_1 is to be estimated. In practice $Z(\mathbf{u})$ could be a covariate, spatial coordinates, or any transformation of covariate or spatial coordinates (Baddeley 2008). They all can work as the statistic for the intensity function. The log-likelihood of θ for an inhomogeneous Poisson process with parametric intensity function $\lambda_\theta(\mathbf{u})$ is given by (Berman and Turner 1992):

$$L(\theta; \mathbf{u}_1, \cdots, \mathbf{u}_n; W) = \sum_{i=1}^{n} \log \lambda_\theta(\mathbf{u}_j) - \int_W \lambda_\theta(\mathbf{u}) d\,\mathbf{u} \tag{2.40}$$

where θ denotes the unknown parameters to be estimated. The numerical estimator of the integral $\int_W \lambda_\theta(\mathbf{u}) d\,\mathbf{u}$ is given by:

$$\int_W \lambda_\theta(\mathbf{u}) d\,\mathbf{u} \approx \sum_{j=1}^{M} w_j\, \lambda_\theta(\mathbf{y}_j) \tag{2.41}$$

where $\bigcup_{j=1}^{M} \mathbf{y}_j \equiv \mathbf{y}$ and the quadrature weight w_j is the area of the Dirichlet tile based on Dirichlet tessellation (Baddeley and Turner 2000). The points $\mathbf{y}_j$ are chosen so that $\mathbf{x} \equiv \{\mathbf{u}_1, \cdots, \mathbf{u}_n\} \subset W \subset \{\mathbf{y}_1, \cdots, \mathbf{y}_M\}$. The points in $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{y}$ **-x** are called the design points, the data points and the dummy points respectively.

Substituting equation (2.41) into equation (2.40), it follows that

$$\begin{aligned} L(\theta; \mathbf{u}_1, \cdots, \mathbf{u}_n; W) &= \sum_{i=1}^{n} \log \lambda_\theta(\mathbf{u}_i) - \sum_{j=1}^{M} w_j\, \lambda_\theta(\mathbf{y}_j) \\ &= \sum_{j=1}^{M} w_j \left\{ \frac{s_j}{w_j} \log \lambda_\theta(\mathbf{y}_j) - \lambda_\theta(\mathbf{y}_j) \right\} \end{aligned} \tag{2.42}$$

where $s_j = \begin{cases} 1 & \text{if} \quad \mathbf{y}_j \in \mathbf{x} \\ 0 & \text{if} \quad \mathbf{y}_j \in \mathbf{y} - \mathbf{x} \end{cases}$

Then the expression (2.42) is maximised in which the intensity function $\lambda_\theta(\mathbf{u})$ is of the form

$$\log \lambda_\theta(\mathbf{u}) = \theta \cdot Z(\mathbf{u}) \tag{2.43}$$

where the values of $Z(\mathbf{u})$ are known for all locations $\mathbf{u} \in W$, even when $\mathbf{u}$ is not a data point. In a point pattern process, the observed information does not include only the data

points locations, the absence of other points locations is also important (Baddeley et al. 2005). The form of Z in (2.43) is arbitrary.

2.2.4 Methods to obtain covariate

There are spatial covariates which are important in the analysis and these spatial covariates are treated as explanatory variables. This thesis will study whether the intensity of the spatial point patterns depends on the covariates. It is possible to fit a model with an intensity function which depends on the available covariates. The covariates used in the inhomogeneous Poisson model (2.39) are pixel images. Pixel images in this study are calculated using a kernel function, and also inhomogeneous *K*-function is calculated using kernel function. A kernel function is a weighting function and it is more commonly used in non-parametric estimation. In kernel estimation, bandwidth is a scalar parameter and it determines what range of the nearby data points will be heavily weighted when making an estimate. Popular kernel functions are the Gaussian kernel, the quadratic kernel and the minimum variance kernel. The Gaussian kernel is defined as:

$$k(\mathbf{u}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\mathbf{u}^2/2\right\} \tag{2.44}$$

and $\int \mathbf{u}^2 k(\mathbf{u}) d\,\mathbf{u} = 1$ for the Gaussian kernel where σ is the standard deviation.

In practice, choosing kernel functions is less important than choosing the bandwidth (Schabenberger and Gotway 2005). In R documentation, `smooth.ppp` is described as performing spatial smoothing of numeric values observed at irregular locations. If the observed numeric values are $m_1, \cdots, m_n$ at a set of irregular locations $\mathbf{u}_1, \cdots, \mathbf{u}_n$ respectively, then the smoothed value at a location $\mathbf{u}$ is

$$g(\mathbf{u}) = \sum_{i=1}^{n} k(\mathbf{u} - \mathbf{u}_i)\mathrm{m}_i \Big/ \sum_{i=1}^{n} k(\mathbf{u} - \mathbf{u}_i) \tag{2.45}$$

where edge correction is ignored and $\{\mathbf{u}_1,\cdots,\mathbf{u}_n\}$ must be a marked point pattern. Points which are classified into two or more different types may be regarded as marked points, with a mark which identifies which type they belong to. The mark can be a categorical or a continuous variable.

Distance methods can be used to explore point dependence, and they are measured by calculating the distances between points. The empty space distance is one of these and it reflects how far away each observed point is from the nearest point on a pre-defined regular grid (thus resulting in a pixel image). The empty space distance is the distance from a fixed reference location **u** in the window to the nearest data point in a point pattern **x**. It is defined by

$$\mathrm{d}(\mathbf{u},\mathbf{x}) = min\{\|\mathbf{u}-\mathbf{u}_i\| : \mathbf{u}_i \in \mathbf{x}\} \tag{2.46}$$

The models are fitted by `spatstat` command **ppm** using a formula with regard to Papangelou conditional intensity $\lambda(\mathbf{u},\mathrm{x})$ (Papangelou 1974). This intensity function is dependent on the realisations of the point process at all spatial locations except **u** itself. This conditional intensity is briefly introduced (Baddeley et al. 2005) here for the purpose of data analysis.

Suppose the point process **X** in *W* has a probability density *f*(**x**) which is hereditary, that is

$$\text{If } \mathrm{f}(\mathbf{x})>0 \quad \text{and} \quad \mathbf{y}\subset\mathbf{x} \quad \Rightarrow \mathrm{f}(\mathbf{y})>\mathbf{0} \quad \text{for any} \mathbf{x},\mathbf{y}\subset \mathrm{W} \tag{2.47}$$

For dummy points in the bounded region *W*: $\mathbf{u}\in W$ but $\mathbf{u}\notin\mathbf{x}$, the intensity function is defined

$$\lambda(\mathbf{u},\mathbf{x}) = \begin{cases} \mathrm{f}(\mathbf{x}\cup\{\mathbf{u}\})/\mathrm{f}(\mathbf{x}) & \mathrm{f}(\mathbf{x})>\mathbf{0} \\ \mathbf{0} & \mathrm{f}(\mathbf{x})\le\mathbf{0} \end{cases} \tag{2.48}$$

For data points: $\mathbf{u}\in W$ and $\mathbf{u}\in\mathbf{x}$, the intensity function is defined

$$\lambda(\mathbf{u},\mathbf{x}) = \lambda(\mathbf{u},\mathbf{x}\setminus\{\mathbf{u}\}) \tag{2.49}$$

where $\lambda(\mathbf{u},\mathbf{x})\,d\,\mathbf{u}$ can be interpreted as the conditional probability of finding a point at a specified location provided that the information of the point pattern is known at locations $(\mathbf{x}\setminus\{\mathbf{u}\})$.

2.2.5 Validation of the fitted model

Quantile-quantile plots and diagnostic plots are used to validate the fitted model. The residuals used in these plots are raw residuals. For any set $B \subseteq W$, raw residuals (Baddeley et al. 2005) are defined as the difference between the observed value and the expected value. That is:

$$\mathrm{P}(B) = \mathrm{n}(\mathbf{x}\cap B) - \int_B \lambda_{\hat{\theta}}(\mathbf{u},\mathbf{x})\mathrm{d}\mathbf{u}$$

where $\hat{\theta}$ is the estimator of θ. When the covariates account for point process intensity, it is important to indicate how strongly the point pattern depends on the covariates. A lurking variable plot not only shows the appropriateness of the fit but also illustrates how much the fitted intensity function relies on the covariates. A lurking variable is defined by Box (1966) as a variable that may have important influence and yet is not incorporated among the explanatory variables under consideration. The lurking variable plot is a plot of residuals $B(z)$, denoted by $C(z) = R(B(z))$, against covariate value z where

$$B(\mathrm{z}) = \{\mathbf{u}\in W : Z(\mathbf{u}) \le \mathrm{z}\}$$

is the region of space where the covariate value is less than or equal to z (Baddeley,2008). In the next chapter the approach to the modelling of the data is outlined, including a discussion of the data preparation.

3 METHODOLOGY

In this study, two distinct types of cancer data will be considered: incidence and mortality data at SLA level. These data are given as rates; the other type of data is at individual level where the occurrence of incidence or mortality is recorded and these data are binary. For the first type geostatistical techniques can be used, and for the second point process theory is more appropriate.

3.1 Geostatistical analysis

The incidence and mortality for the five top cancers are considered in this study: lung, melanoma, colorectal, breast and prostate cancers. For the geostatistical analysis, to avoid the use of temporal structure in the model setup, the study period is broken down into three contiguous periods: 1990-1995, 1996-2000 and 2001-2005 to get a relatively refined mortality data analysis, and to see how cancers vary over time. The changes for each period are shown in the temporal plots (Figure 2). For breast cancer and prostate cancer incidence, the analysis during the three contiguous periods is explored. For lung cancer, colorectal cancer and melanoma cancer incidence, we also provide an analysis for the period 1990-2005 by sex and investigate how these cancers affect males and females differently. The analysis of individual contiguous periods for each sex is not carried out within the study since there are not enough data. Spatial continuity and variability of age-adjusted rates and age-sex-adjusted rates will be investigated based on omnidirectional semivariogram plots. For the cancers in this study, omnidirectional semivariograms are well-behaved and show good structure compared with directional semivariograms. What we will obtain from ATP Poisson Kriging are estimates and variance based on the adjusted rates.

3.2 Tools for coordinates conversion

For the coordinates of SLAs, Redfearn's formulae (Redfearn 1948) are used to convert coordinates individually from geographic (latitude & longitude) to Universal Transverse Mercator (UTM) grid coordinates using the GRS80 ellipsoid. This projection is used for

Australia's new coordinate system. It is compatible with the global coordinate system (WGS84). The formulae are accurate to 1mm in any zone. All UTM coordinates are measured in meters.

Geostatistical analysis is carried out for aggregate data within SLA. To get a refined view of geographical patterns, spatial point pattern analysis is applied. Coordinates in spatial point pattern analysis are converted from Latitude/Longitude to UTM using command `convUL` within Jon Schnute's `PBSmapping` package (CRAN 2009). In this command, UTM coordinates in kilometres are used when km is true. The window, a map of Perth, was obtained in the form of GIS "shapefiles". These files were read into `spatstat` using `maptools` package from CRAN (2009). See Appendix 9.2 for the commands.

3.3 The treatment of missing coordinates in spatial point pattern analysis

In the original cancer data set, there is a small percentage of cases where only the Statistical Local Areas (SLAs) at the time of diagnosis (or death) is recorded, but not the exact location. Two methods will be used to treat these missing values in Section 6.1 for spatial point pattern analysis. One method is to remove the case locations without coordinates. The other method is to assign the mean of 16-year period cancer case locations within SLA to them. We will verify whether a slightly different point pattern of data has any influence on the model fitting using lung cancer data as an example, and see which data treatment is better and why.

3.4 Model fitting based on inhomogeneity for spatial point pattern analysis

The quadrat counting test is used to verify whether a point pattern is homogeneous within the Rpackage `spatstat`. Based on the test, the cancer incidence and mortality point processes for 2005 are not completely spatially random (see details in chapter 6). In order to take account of the inhomogeneity, the inhomogeneous *K*-function (Baddeley et al. 2000) is applied in Chapter 6 for each individual point process to test whether the point pattern is

independent. If an inhomogeneous point pattern is independent, then the point pattern is an inhomogeneous Poisson process.

The inhomogeneous *K*-function is a generalisation of the *K*-function to inhomogeneous point patterns. To calculate the inhomogeneous *K*-function for a point process a kernel estimate of the point process intensity is required. We will be using an isotropic Gaussian kernel with standard deviation σ to calculate the estimate. In addition to the expected value of the K_{inhom} 5% confidence limits are calculated via simulation and again, a smoothing bandwidth is required. As the *K*-function of an inhomogeneous Poisson process is πr^2 the smoothing bandwidth sigma needs to be chosen so that the estimated values of the inhomogeneous *K*-function are of the same order of magnitude as πr^2. So, when the r values are in the range 0 to 15, the maximum values of *K*-function should be about 700. In the inhomogeneous *K*-function plot, the range of 'r' values shown is the 'recommended range' according to Ripley's rule, which is [0,R] where R is one quarter of the smallest side of the rectangle containing the study region. The estimate of K_{inhom} may also be available for large values of r, but is not plotted for these larger values by default.

3.5 The inhomogeneous Poisson model

In the inhomogeneous Poisson model (2.44), the covariate can be a dataframe or a pixel image. In this study, the covariates available for different cancers are considered for model fitting. How to find an adequate fit and how to validate the fitted models are studied individually for five cancers.

The treatment options for missing values will be illustrated using the lung cancer point patterns and their potential covariates. The validated treatment method is then applied in the discussion of the other cancers. Prostate cancer is used as an example, to discuss the process to select an appropriate model.

When the fitted model involves spatial covariates, lurking variable plots are important to show how much the fitted model captures the dependence of intensity on the covariates. The plot indicates approximate 5% significance bands for the cumulative residual

(Baddeley 2008). The selected covariate does not account for the point pattern when the plot is totally outside of the significance bands and just stays either over or below the horizontal axis.

What we obtain from point pattern analysis is the spatial trends within the point process intensity and the validation of the fitted point process model using the techniques in Baddeley et al.(2005).

4 DATA DESCRIPTION

4.1 Data preparation

The data used in this study are cancer incidence and mortality data for lung, melanoma, colorectal, breast and prostate cancers for the period 1990-2005 in the Perth metropolitan area. Perth is the capital city of Western Australia, located in the southwest of Western Australia. Western Australia has a total of 2.3 million inhabitants, the majority of whom (1.7 million) live in the Perth metropolitan area. There are 55519 observations for cancer incidence with 1314 missing values and 18043 observations for cancer mortality with 192 missing values. The cancer incidence and mortality data obtained from the Department of Health of WA cover the areas Perth, Bunbury and Geraldton. Since Bunbury and Geraldton are not close to Perth, the data from these two areas are removed for the analysis in this study. The records used in this study include the following variables: Coded ID number; Latitude/longitude of location at incidence and death; age in 5-year brackets at time of diagnosis and death; sex (1 denotes males, and 2 denotes females); year of diagnosis and death; type of cancer; survival time; cause of death if applicable (by these cancer types, other cancer, other cause) and indigenous status.

Acoording to the data, nearly all the cancer patients died of the diagnosed cancer rather than other cancers or other causes. As for the indigenous status, nearly 99% of the cancer patients are not indigeneous and 1% are indigeneous or of unknown status. Therefore, these variables are not informative for this study.

4.2 Raw cancer data

During the period 1990-2005, there is an overall increase in the number of incidence and mortality cases for top five cancers in the Perth metropolitan area and the total number of cancer incidence cases is increasing more rapidly than the number of mortality cases in 1990-2005 (Figure 1 and Figure 2). The cancer rates are higher for men than for women in Perth. This overall increase is apparent in all cancers. However, for prostate cancer incidence there was rapid growth from 1990 to 1994 followed by a sharp decline until 1997

and again a phase of rapid growth from 2000 onwards. The number of mortality cases is much lower than the number of incidence cases for each type of cancer. For melanoma mortality, the number of cases is relatively low compared with other cancers in each individual year. There is slight increase with fluctuations in cancer mortality for breast and prostate cancer during the period 1990-2005.

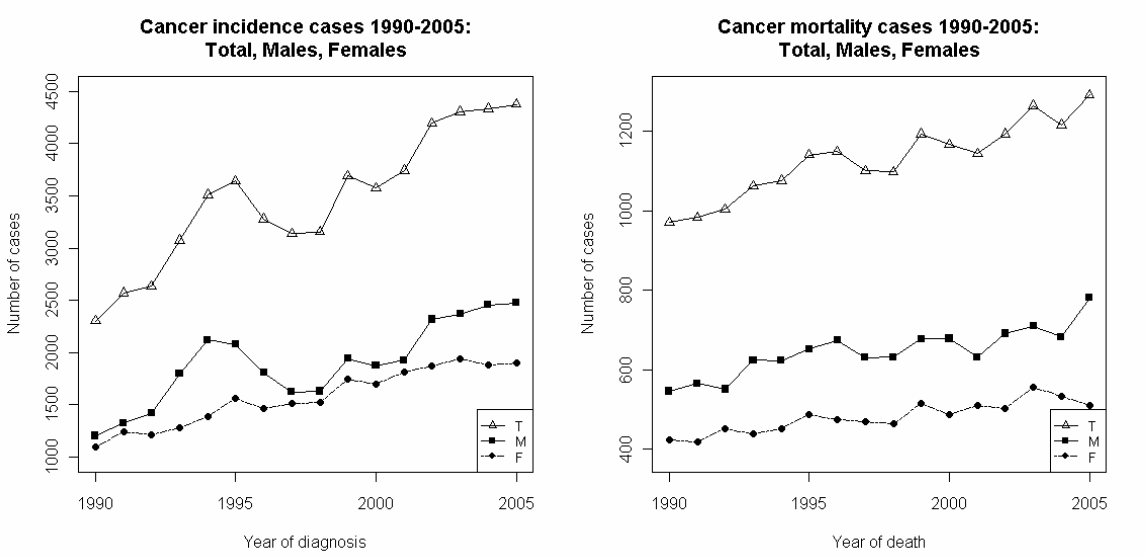


Figure 1 Counts of cancer incidence (left) and mortality (right) recorded in the Perth metropolitan area (T denotes total, M denotes males and F denotes females).

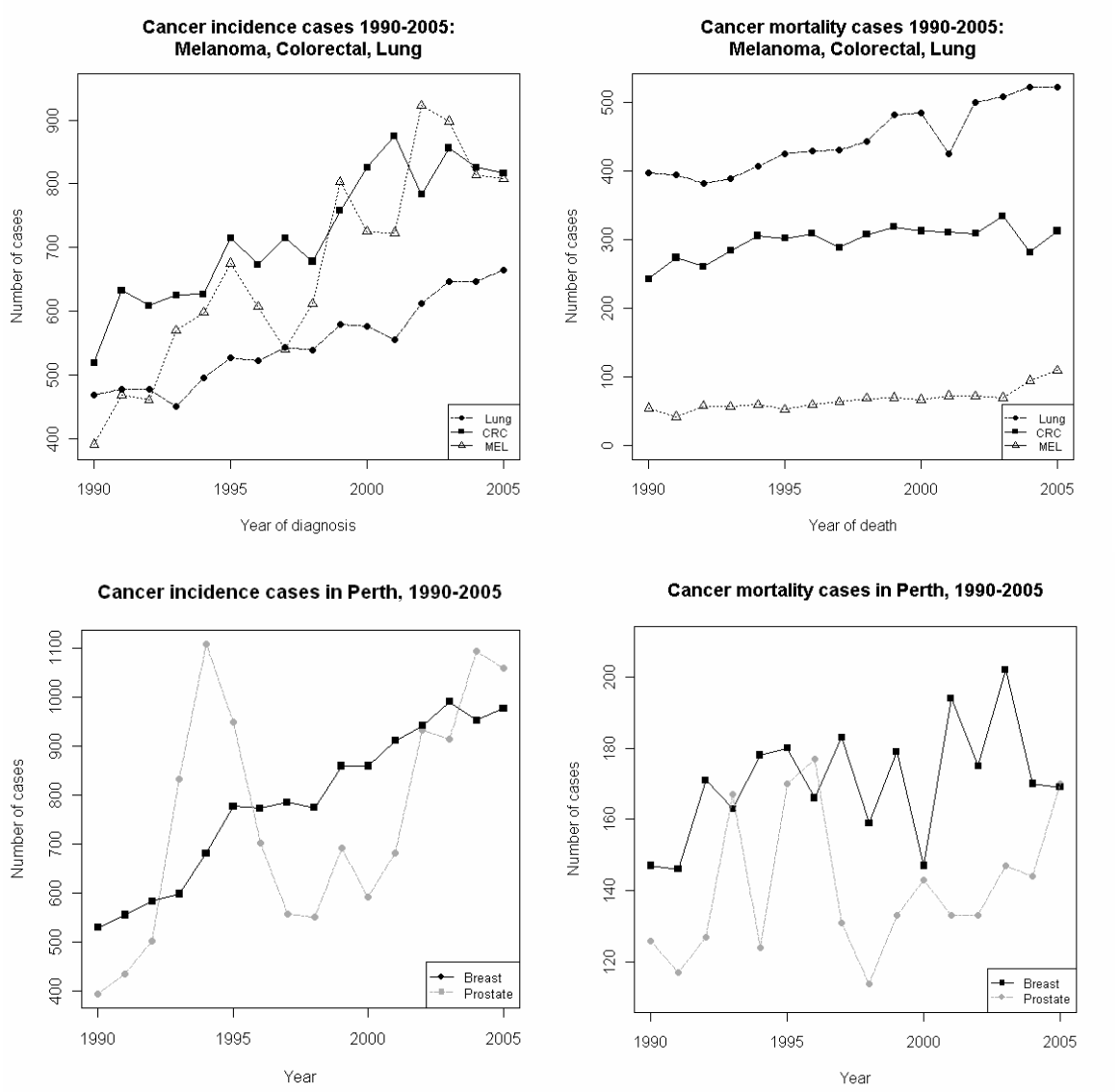


Figure 2 Yearly incidence cases (left) and mortality cases (right) recorded in Perth metropolitan area.

Survival rate is the percentage of people who survive a disease such as cancer for a given period of time after diagnosis. The plots for survival rates (Figure 3) show that about 40% of people with lung cancer were alive one year after diagnosis. In comparison, the survival time of melanoma cancer patients is relatively long: about 50% survived 6 years after diagnosis. The survival rates for breast and prostate cancers are quite similar. The percentages of breast and prostate cancer patients who were alive for 6 years are 43% and 38% respectively. The survival rates for lung cancer are relatively low compared with other cancers. The survival rates are not directly used for the modelling in this study, but they are the reference for the selection of covariates (susceptible population intensity) in spatial point pattern analysis.

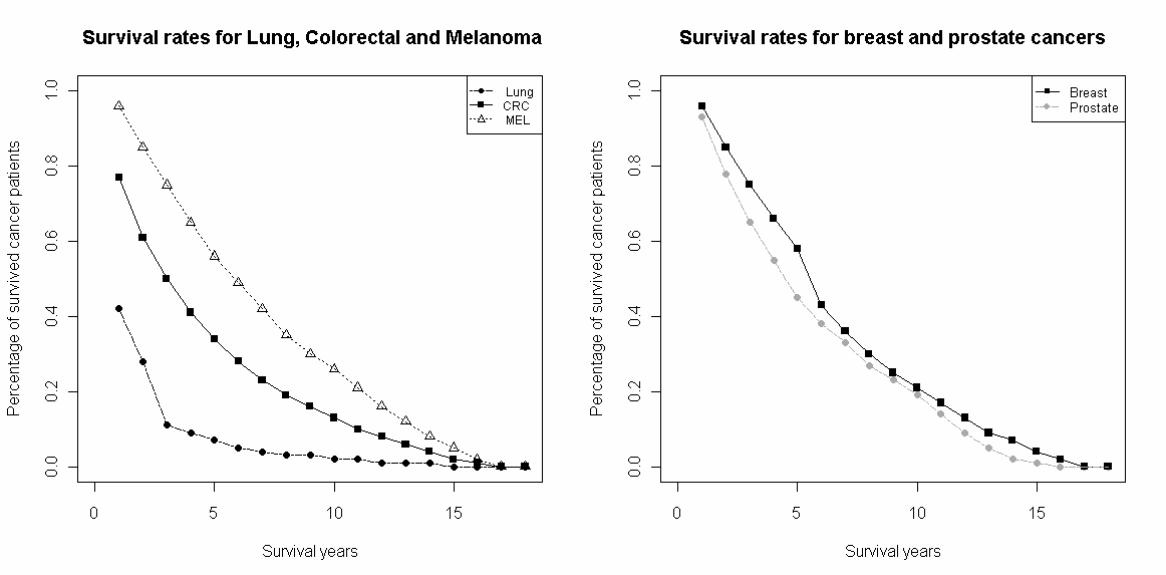


Figure 3 Survival rates for lung, colorectal, melanoma, breast and prostate cancers recorded in 1990-2005.

Barplots shown in Figure 4 and Figure 5 are based on incidence and mortality cancer data respectively recorded in the Perth metropolitan area from 1990 to 2005. They indicate clearly that males and females have different risk from lung, colorectal and melanoma cancer especially when people are of age 50 or above. Moreover, different age groups of people have different cancer risks. Generally speaking, age is a main cancer risk factor. The incidence risk for lung cancer and colorectal cancer increases greatly with age for both sexes and reaches the highest at age 70. While the growth rate for lung cancer incidence with age is higher for males than females, this is not the case for colorectal cancer. For melanoma cancer, people aged 50-60 are the most affected group. Melanoma incidence risk increases with the age for males but this is not the case for females. For sex-specific cancers, breast cancer has the similar pattern with female melanoma cancer: age group 40-60 is the most affected. The risk of having prostate cancer increases with age, especially when males are aged 50 or over. For mortality, the number of cases is much lower than the corresponding number of incidence cases. And also the influence of sex on mortality is not so strong compared with the influence of sex on incidence. Based on these observations, kriging analysis will incorporate age and sex for incidence and mortality data to make the estimation more reliable.

To obtain further information on incidence and mortality data from different perspectives, we will look at the kriging analysis for lung, melanoma and colorectal cancer incidence by

sex and the kriging analysis for the mortality of three cancers will be carried out for three contiguous periods: 1990-1995, 1996-2000 and 2001-2005.

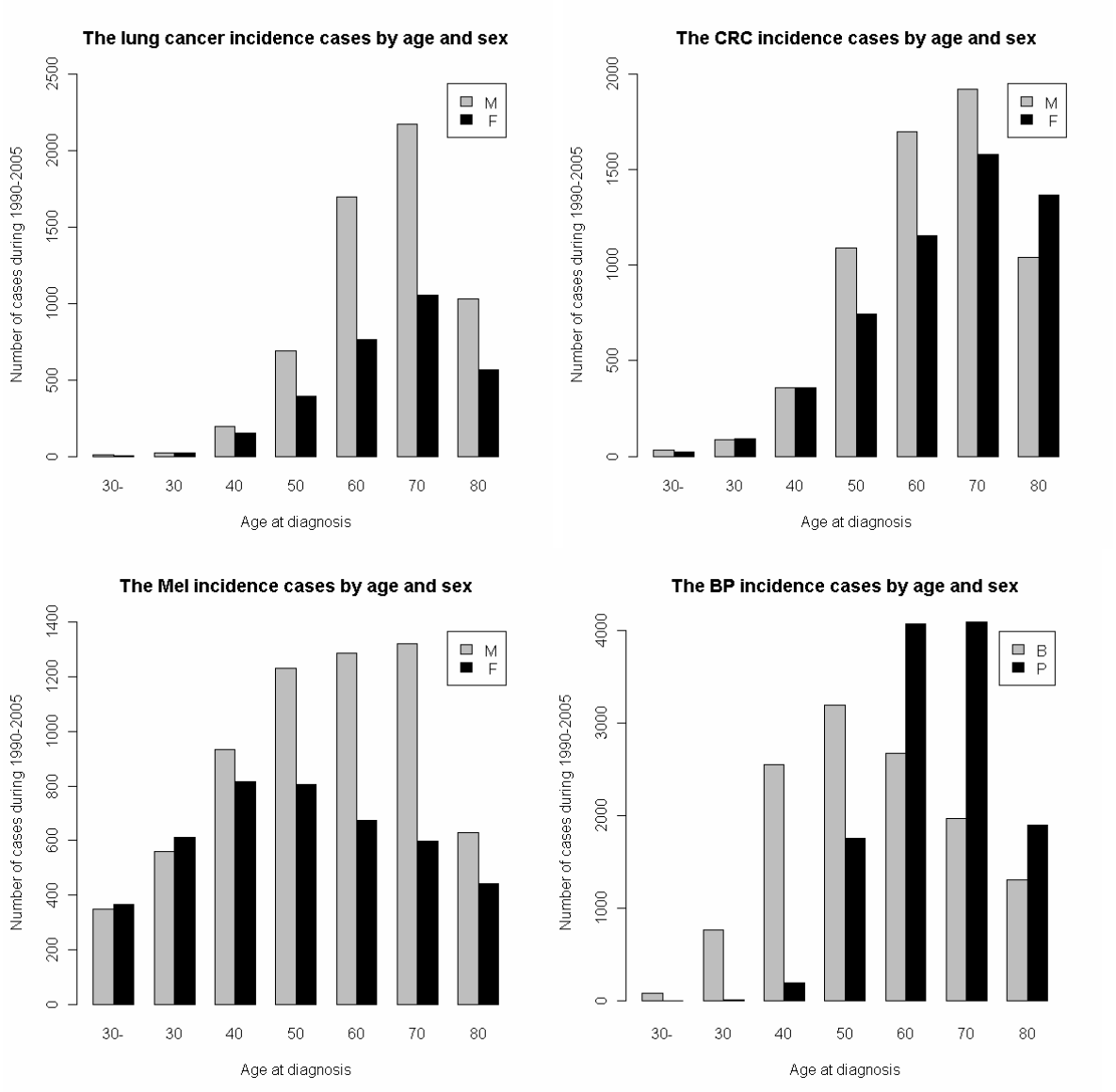


Figure 4 Counts of cancer incidence by age and sex (M denotes males and F denotes females) recorded in Perth during the period 1990-2005 (B denotes breast and P denotes prostate). The age is in 10-year brackets with the lower limit stated and this is not the case for 30-.

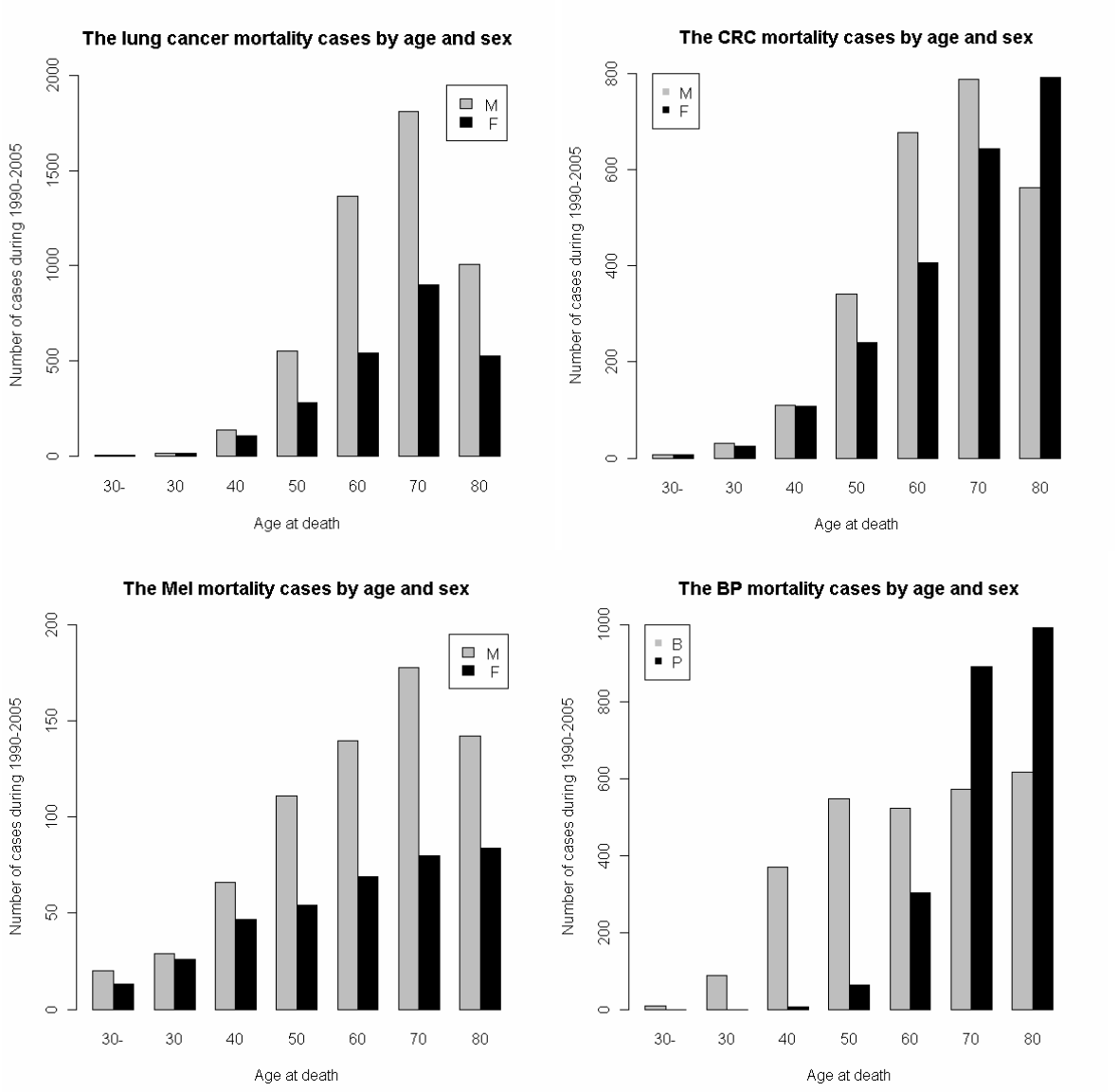


Figure 5 Counts of cancer mortality by age and sex (M denotes males and F denotes females) recorded in Perth during the period 1990-2005 (B denotes breast and P denotes prostate). The age is in 10-year brackets with the lower limit stated.

4.3 Coordinates transformations

For geostatistical modelling, data were aggregated within Statistical Local Areas (SLAs) (see Figure 6 for the SLA map of Perth). For nearly all cancer records the SLA name is recorded. An SLA is a general purpose spatial unit. The SLAs cover the Australia without gaps or overlaps. The SLAs vary in terms of area size and population.

To apply ATP Poisson Kriging the centroid of each SLA is required. The centroid for each SLA is calculated as the mean value of the non-zero coordinates of cancer case locations in

each SLA during the entire study period. For example, to obtain the centroid for Bassendean (T), 602 case locations of the total of 611 records for five cancers were used to compute the mean coordinates of cancer incidences. There are cancer incidence cases between 1990 and 2005. For nine case locations no coordinates were recorded. The coordinates were first converted from Lat/Long to UTM and then the mean location was calculated by simple averaging of the UTM Eastings and Northings. The nine locations in this area without latitude and longitude were then allocated the coordinates of the centroid. The coordinates for other Statistical Local Areas at incidence and mortality with more than one case are calculated in a similar way. In the original data file, there are ten incidence locations without SLAname. Eight of them are located within the Perth metropolitan area. In this study, they are allocated to the SLA Perth (inner). The Cartesian coordinates for the SLAs in Perth obtained in this fashion are shown in Table 1.

Table 1 The Cartesian coordinates of SLA in Perth for cancer incidence. The coordinates were converted from Lat/Long to UTM using Redfearn's formulae.

Perth-SLAname	X-coord	Y-coord	Perth-SLAname	X-coord	Y-coord
Armadale (C)	407322.54	6444166.60	Mundaring (S)	415780.97	6470986.32
Bassendean (T)	400317.66	6469864.53	Nedlands (C)	386260.98	6461326.06
Bayswater (C)	396438.27	6469526.74	Peppermint Grove (S)	383489.96	6459023.52
Belmont (C)	399181.15	6463667.39	Perth (C) - Inner	391980.03	6464115.81
Cambridge (T)	386167.12	6466178.95	Perth (C) - Remainder	391615.85	6463716.63
Canning (C)	396783.75	6455249.60	Rockingham (C)	381723.06	6424365.49
Claremont (T)	384679.25	6460828.65	Serpentine-Jarrahdale (S)	404937.22	6429136.66
Cockburn (C)	387280.63	6447590.60	South Perth (C)	392968.21	6459442.73
Cottesloe (T)	382734.38	6459614.62	Stirling (C) - Central	390694.93	6471749.66
East Fremantle (T)	383705.84	6454632.45	Stirling (C) - Coastal	384360.80	6471370.77
Fremantle (C) - Inner	381680.86	6452813.13	Stirling (C) - South-Eastern	393378.79	6467827.54
Fremantle (C) -Remainder	383769.04	6452235.92	Subiaco (C)	388340.94	6463779.82
Gosnells (C)	403035.30	6452388.93	Swan (C)	403502.47	6475236.37
Joondalup (C) - North	381979.41	6486667.41	Victoria Park (T)	396471.29	6460747.02
Joondalup (C) - South	383964.98	6479576.16	Vincent (T)	391602.65	6466696.34
Kalamunda (S)	408917.19	6461435.01	Wanneroo (C) - North-East	387298.47	6486843.25
Kwinana (T)	388897.37	6432425.66	Wanneroo (C) - North-West	375295.14	6499582.94
Melville (C)	388976.49	6454343.56	Wanneroo (C) - South	390817.14	6477623.66
Mosman Park (T)	383341.10	6457489.73			

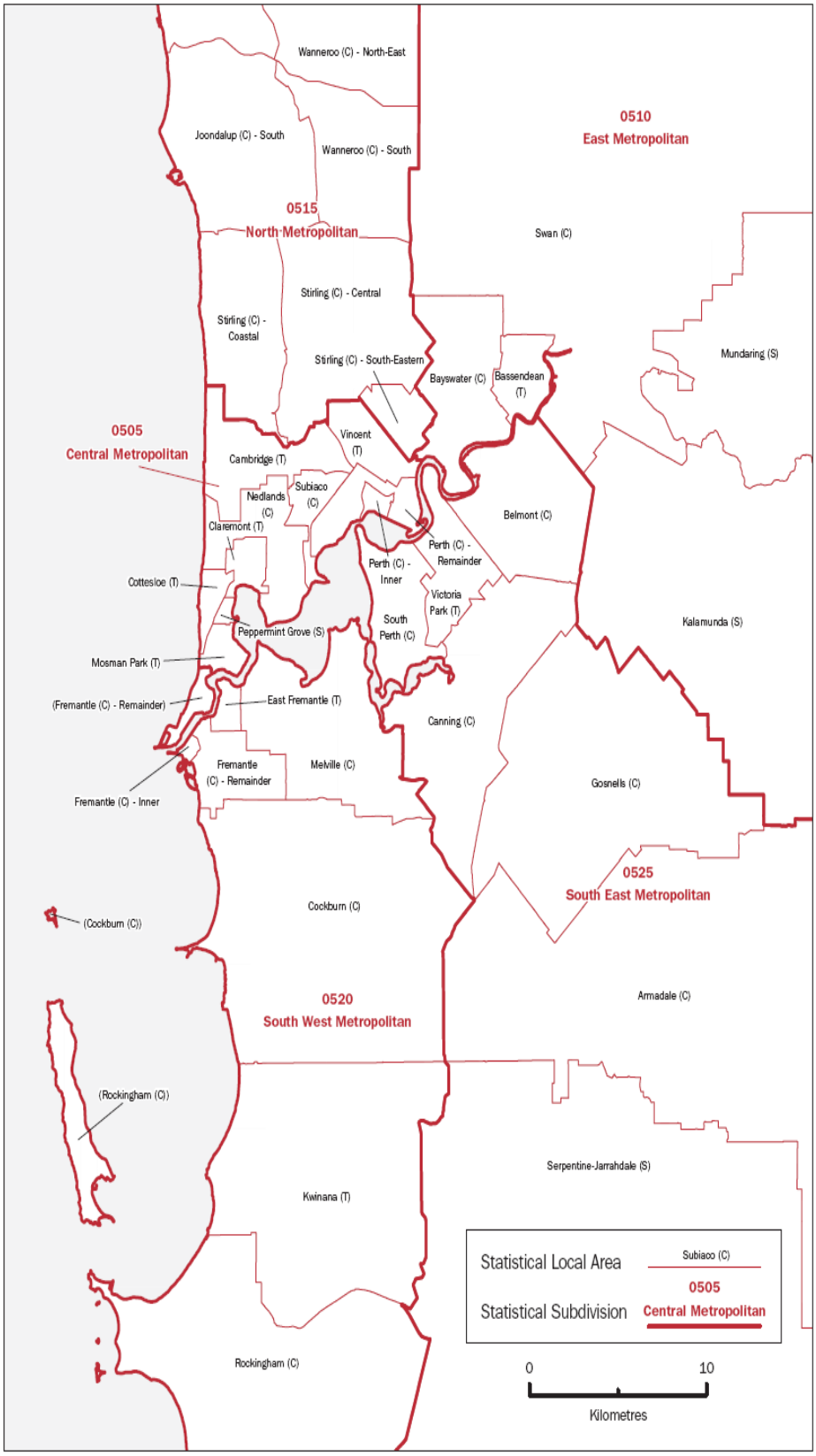


Figure 6 Statistical local area (SLA) map of Perth.

4.4 Census data

The population data were obtained from Australian Bureau of Statistics (ABS) and they are only available for census years. The population data (year 1995, 2000 and 2006) used for kriging analysis are recorded by SLA, age in 5-year brackets and sex. There is a gradual increase in the population with time, so the year 2000 population is appropriate to be used in equation (4.2) for the calculation of rates. The population data by sex in SLAs of the Perth metropolitan area in the year 2000 are shown in Table 2. In this table, the population size of males, females and total in each SLA is denoted by Males, Females and Total respectively. The population data in year 2006 (Table 3) used for point pattern analysis have the variables by suburb: ObjectId, Total (population size), Area (suburb size by square kilometers), XY (coordinate by Lat/Long), Thirtypp (proportion of people aged 30 above) and Fiftypp (proportion of people aged 50 above). The corresponding Perth map in the form of "shapefile" by SLA and by suburb was also obtained from ABS. The population data in 2006 by sex (Table 4) used for point pattern analysis have the variables by suburb: ObjectId, Mpop (male population size), Fpop (female population size), Area (suburb size by square kilometers), XY (coordinate by Lat/Long), Mfortypp (proportion of males aged 40 above) and Ffortypp (proportion of females aged 40 above).

Table 2 The population data by SLA in Perth metropolitan area in year 2000.

SLANAME	Males	Females	Total	SLANAME	Males	Females	Total
Armadale (C)	26594	26788	53382	Mundaring (S)	17431	17584	35015
Bassendean (T)	6948	7015	13963	Nedlands (C)	10454	11051	21505
Bayswater (C)	28382	28495	56877	Peppermint Grove (S)	731	1017	1748
Belmont (C)	14685	14759	29444	Perth (C) - Inner	341	238	579
Cambridge (T)	11643	12356	23999	Perth (C) - Remainder	3526	2680	6206
Canning (C)	38281	38452	76733	Rockingham (C)	35022	35730	70752
Claremont (T)	4494	5028	9522	Serpentine-Jarrahdale (S)	5962	5520	11482
Cockburn (C)	34642	33531	68173	South Perth (C)	18814	18870	37684
Cottesloe (T)	3706	3893	7599	Stirling (C) - Central	48135	49887	98022
East Fremantle (T)	3306	3354	6660	Stirling (C) - Coastal	30472	31282	61754
Fremantle (C) - Inner	434	310	744	Stirling (C) - South-Eastern	8269	8744	17013
Fremantle (C) - Remainder	12664	12596	25260	Subiaco (C)	7671	8092	15763
Gosnells (C)	42021	41359	83380	Swan (C)	41752	40619	82371
Joondalup (C) - North	23295	23448	46743	Victoria Park (T)	13368	14131	27499
Joondalup (C) - South	55795	54893	110688	Vincent (T)	12766	12570	25336
Kalamunda (S)	24514	24965	49479	Wanneroo (C) - North-East	9333	9145	18478
Kwinana (T)	11186	10367	21553	Wanneroo (C) - North-West	12032	12011	24043
Melville (C)	46890	50459	97349	Wanneroo (C) - South	17975	18235	36210
Mosman Park (T)	3779	4340	8119				

Table 3 The extract of the population data by Suburb in Perth metropolitan area in year 2006.

ObjectId	Total	Area	X	Y	Thirtypp	Fiftypp
0	7233	2.83614	115.7382	-31.71801	0.299737	0.167289
1	798	1.55409	115.9976	-31.8775	0.086466	0.357143
2	11090	5.83138	115.7543	-32.31566	0.03183	0.256447
3	3685	3.5533	115.8665	-31.84142	1.56445	0.271913
4	680	0.32683	116.0424	-31.90809	1.938235	0.282353
5	740	1.34342	116.0725	-31.99616	0.468919	0.313514
6	9	6.20102	115.7902	-32.22808	31.55556	0
7	781	12.4735	115.8762	-32.19426	2.957746	0.28041
8	0	0.57674	115.8147	-32.24972	0	0
9	69	5.84969	115.8315	-31.79418	5.811594	0.507246
10	10929	20.1763	115.8161	-31.75743	0.022143	0.338915
11	5633	8.37551	115.8663	-31.80763	0.218889	0.160128
12	11420	13.9844	115.7694	-32.34137	0.08993	0.210946
13	4840	3.21225	115.9418	-32.04081	0.665496	0.270661
14	3794	3.09271	115.8088	-31.83965	0.383237	0.387454
15	2740	1.53811	115.9063	-31.96756	0.708029	0.319343
16	2653	4.90167	115.7963	-32.26056	1.093856	0.166981
17	2054	1.56839	115.8866	-32.01573	0.363681	0.325219
18	1195	1.6777	115.7599	-31.85283	0.119665	0.422594
19	2741	1.4514	115.8411	-31.93191	0.502736	0.245166
20	11549	7.53015	115.8672	-32.0759	0.223742	0.304615

Table 4 The population data by Suburb for males and females in Perth metropolitan area in year 2006.

ObjectId	Mpop	Fpop	Area	X	Y	Mfortypp	Ffortypp
0	3555	3678	2.83614	115.7382	-31.71801	0.355837	0.348287
1	382	416	1.55409	115.9976	-31.8775	0.489529	0.528846
2	5526	5564	5.83138	115.7543	-32.31566	0.411328	0.423077
3	1871	1814	3.5533	115.8665	-31.84142	0.385355	0.421169
4	339	341	0.32683	116.0424	-31.90809	0.410029	0.407625
5	369	371	1.34342	116.0725	-31.99616	0.460705	0.490566
6	9	0	6.20102	115.7902	-32.22808	0	0
7	418	363	12.4735	115.8762	-32.19426	0.449761	0.498623
8	0	0	0.57674	115.8147	-32.24972	0	0
9	33	36	5.84969	115.8315	-31.79418	0.545455	0.555556
10	5447	5482	20.1763	115.8161	-31.75743	0.452176	0.503831
11	2752	2881	8.37551	115.8663	-31.80763	0.349564	0.319334
12	5630	5790	13.9844	115.7694	-32.34137	0.359147	0.377547
13	2469	2371	3.21225	115.9418	-32.04081	0.378696	0.414171
14	1873	1921	3.09271	115.8088	-31.83965	0.486385	0.550234
15	1379	1361	1.53811	115.9063	-31.96756	0.476432	0.493755
16	1280	1373	4.90167	115.7963	-32.26056	0.325781	0.296431
17	1060	994	1.56839	115.8866	-32.01573	0.437736	0.512072
18	585	610	1.6777	115.7599	-31.85283	0.555556	0.590164
19	1346	1395	1.4514	115.8411	-31.93191	0.360327	0.370609
20	5725	5824	7.53015	115.8672	-32.0759	0.469345	0.497596

4.5 Calculation of cancer rates

The crude cancer rate in each SLA during the period 1990-2005 was calculated as the ratio of 100,000 $\times$ the total number of cancer incidences during the 16-year period applied to $16 \times$ the total population in year 2000, in which 16 is the time span between 1990 and 2005. The crude rate in an SLA may be affected by the population structure in the SLA.

The rates can be adjusted for demographic factors such as age or sex to remove the bias from them. In this study, nine age specific groups are used: 0-9, 10-19, …, 70-79 and one 80+ group. The age-adjusted rates (Goovaerts 2005c) per 100,000 person-years for each cancer in each SLA during the period 1990-2005 were computed as:

$$R=100{,}000\times\sum_{i=1}^{9} r_i \times p_i \qquad (4.1)$$

where p_i is the percentage of age group i in WA in the year 2000, and r_i is the age-specific rate for the population of age group i and computed as

$$r_i = {d_i}\big/{16N_i} \qquad (4.2)$$

where d_i is the total number of incidence cases over the period 1990-2005 for age group i and N_i is the population of age group i in SLA in the year 2000. An age-adjusted rate will reflect what the overall rate would be in different Statistical Local Areas if the population in SLA had the same age distribution as the WA age distribution in 2000. For each cancer, the population at risk was computed as:

$$100{,}000\times{d}\big/{R} \qquad (4.3)$$

where d is the total number of incidence cases over the period 1990-2005 and R is the age adjusted rate. The population at risk will be used in the kriging estimation calculation.

The age-sex-adjusted rate is calculated based on the calculation of age-adjusted rates. The age-sex-adjusted rates per 100,000 person-years for each cancer in SLA during the period 1990-2005 were computed as:

$$100{,}000\times\sum_{i=1}^{9}(r_{1i}\times p_{1i}+r_{2i}\times p_{2i})/2 \qquad (4.4)$$

where p_{1i} and p_{2i} are the percentages of male age group i and female age group i respectively in WA 2000. r_{1i} and r_{2i} are the age-specific rates for the male population of age group i and female population of age group i respectively.

Since the age distribution may vary greatly by geographic region, age-adjusted rates are commonly used in cancer data analysis to eliminate the bias resulting from the difference in

the population structure (Klein and Schoenborn 2001). Compared with crude rates, age-adjusted rates can make the groups in diverse regions more comparable. Much of the recent research and discussion on geostatistical analysis of the cancer data are associated with age-adjusted rates. In fact, in addition to age, sex is another significant risk factor. For example, lung cancer rates are higher for males than females in Western Australia and it is more common for old people than young people (Threlfall and Thompson 2007). The age-sex adjusted rate will help to eliminate the bias resulting from differences in age and sex. Also, rates need to be corrected for the small number problem.

4.6 Exploratory data analysis

The ATP Poisson kriging analysis is based on the adjusted rates and population at risk. While the point pattern analysis incorporates the observed covariates. In the first section, we will introduce the age-adjusted rates and age-sex-adjusted rates for cancer incidence and mortality. The population at risk which is based on the adjusted rates is in the second section. For crude rates for different cancers see Appendix 9.1. In the third section, the available covariates used for different cancers are provided.

4.6.1 Adjusted rates for geostatistical analysis

In this section, the adjusted rates for the five types of cancer are considered to identify the difference of cancer distribution in Perth SLAs. First, we look at the maps of the age-adjusted and age-sex-adjusted rates for lung, melanoma and colorectal cancers.

For the overall incidence rates, Figure 7 indicates that lung cancer incidence is relatively low compared with melanoma and colorectal (CRC) cancer and the difference between age-adjusted and age-sex-adjusted rates is slight. Higher melanoma cancer incidence rates arise mainly in the western part of the Perth metropolitan area. Higher colorectal cancer incidence rates occur mainly in the north-west of the Perth metropolitan area and in the suburbs near Perth.

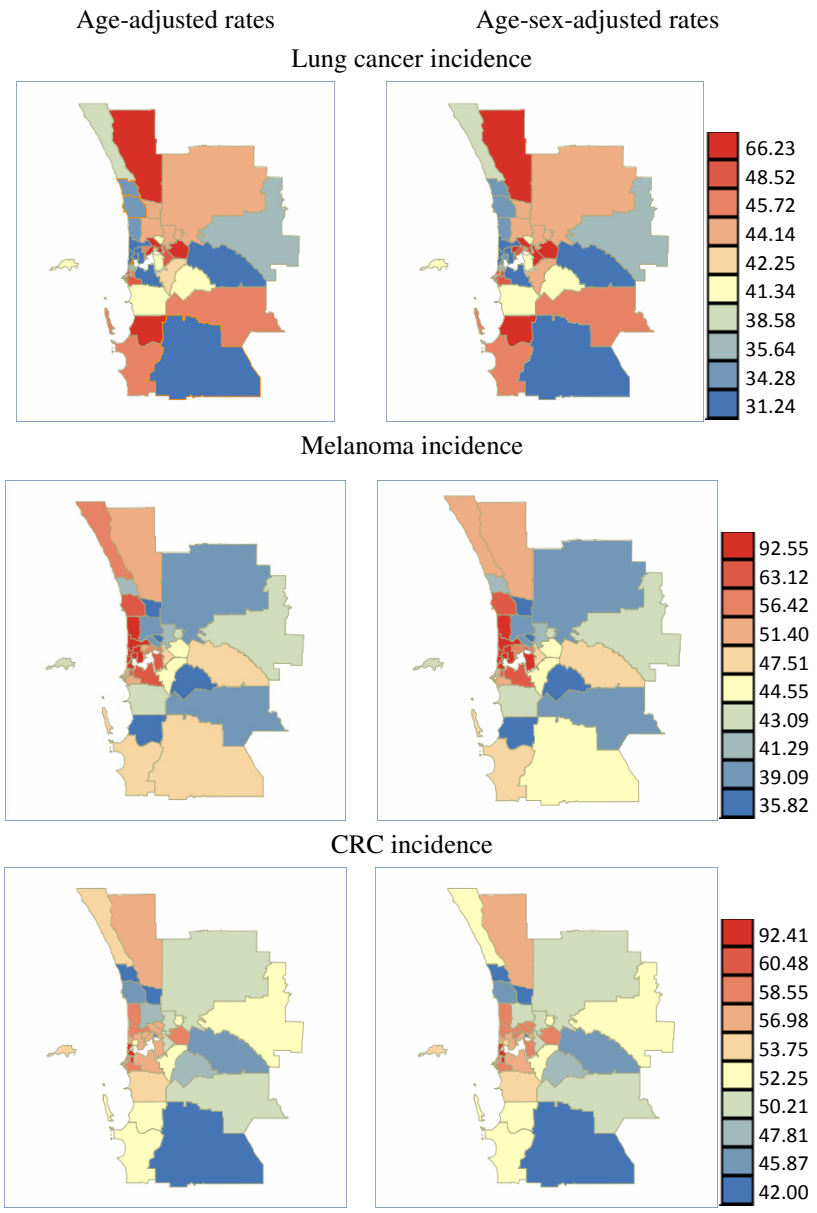


Figure 7 Adjusted rates for cancer incidence.

For the incidence rates by sex, Figure 8 indicates that the difference between males and females is relatively strong for lung cancer compared with melanoma and colorectal cancers. Males and females have relatively low lung cancer risk compared with melanoma and colorectal cancers. Lung cancer adjusted rates of males are much higher than those of females in most areas. For melanoma, the difference between males and females is slight. For colorectal cancer, the highest incidence rate is 130 per 100,000 person-years for males during the period 1990-2005 but only 90 for females.

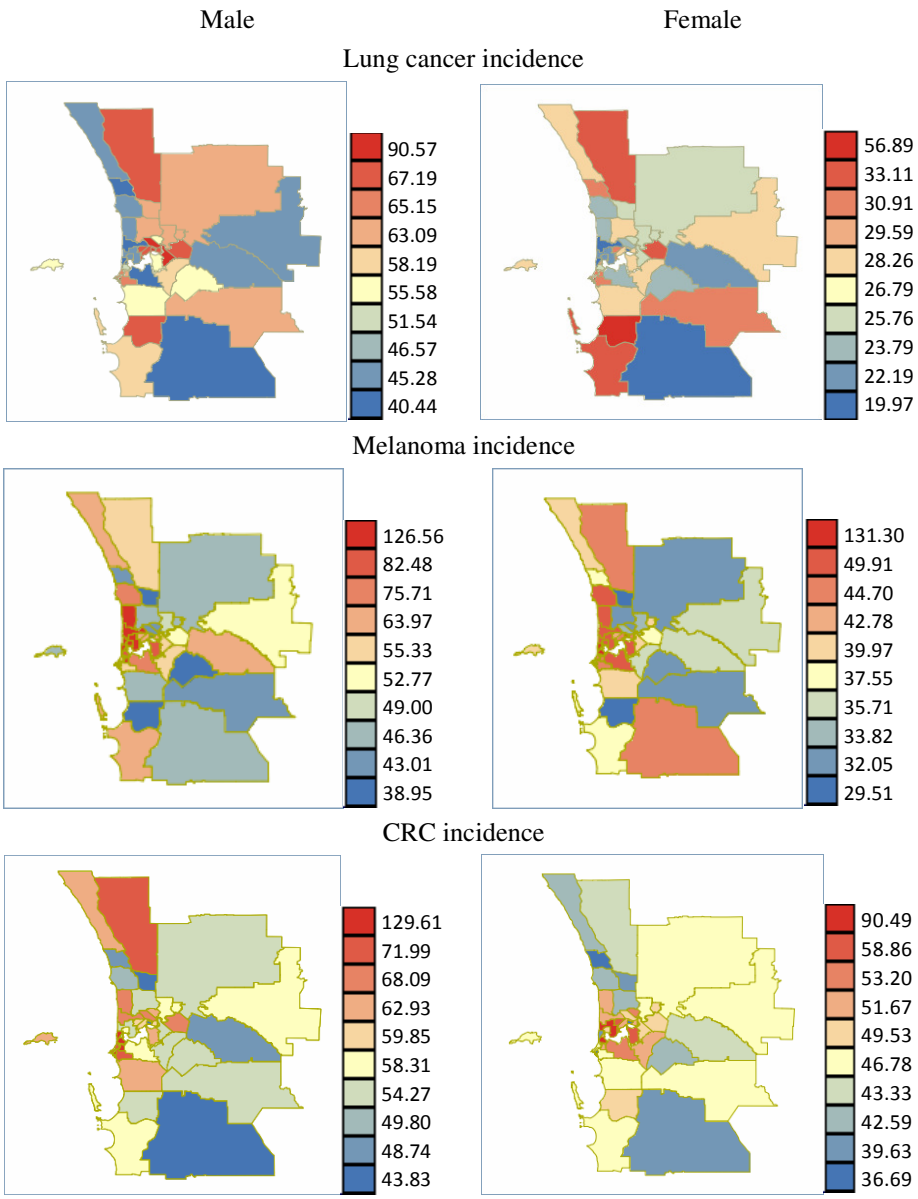


Figure 8 Cancer incidence rates by sex, 1990-2005.

Although lung cancer incidence rates are lower than other cancer rates, the mortality rates for lung cancer are relatively high according to Figure 9. Compared with incidence rates, the two mortality adjusted rates are quite similar. This indicates that age and sex probably have less influence on cancer mortality than on incidence. For melanoma cancer, the rates are much lower than for other cancer types. During the 16-year period, the highest mortality rate is 17 per 100,000 person-years across the Perth metropolitan area. However, the rate is 56 for lung cancer and 38 for colorectal cancer. For colorectal cancer, the mortality rates around Perth metropolitan are relatively high.

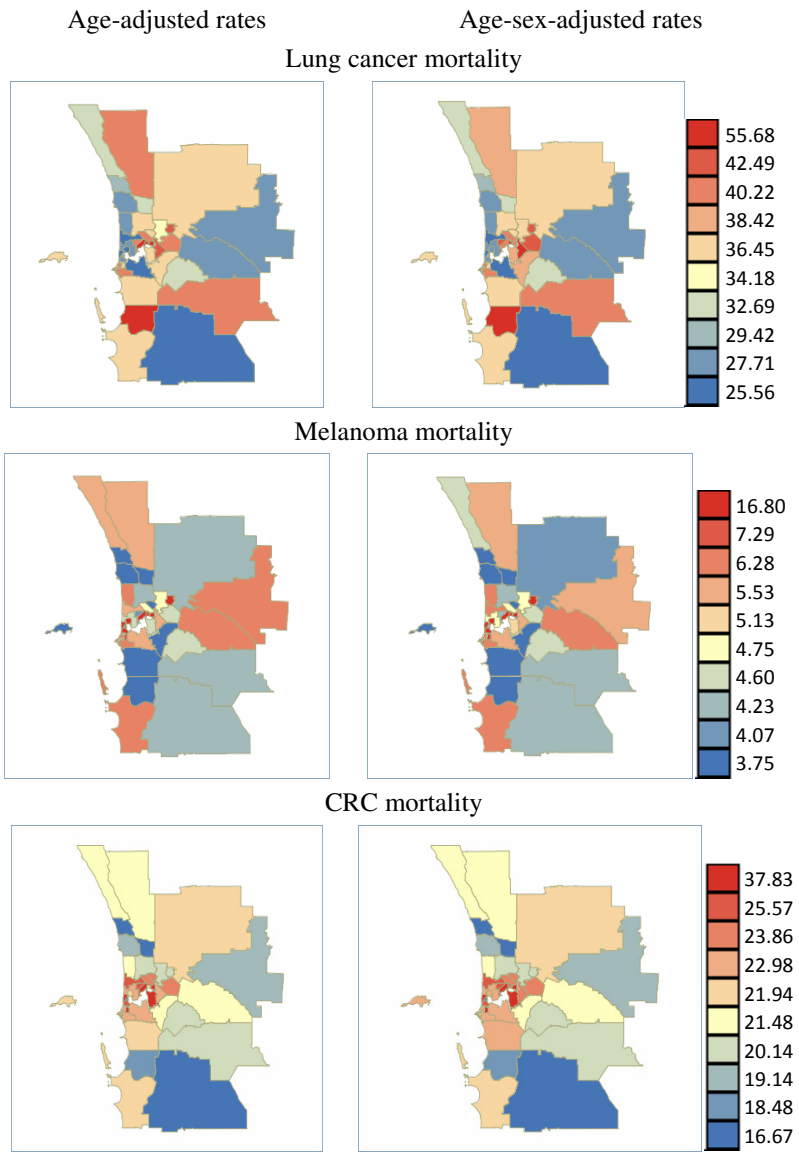


Figure 9 Adjusted rates for cancer mortality.

In Figure 10 the development with time of the mortality rates is shown. We see that there is an overall increase on the lung cancer rates in the Perth metropolitan area. For melanoma, it seems that there is a slight increase in mortality rates in most areas over three periods. For colorectal cancer, the change is slight. It seems that higher colorectal cancer rates are mainly near the inner city suburbs in each period.

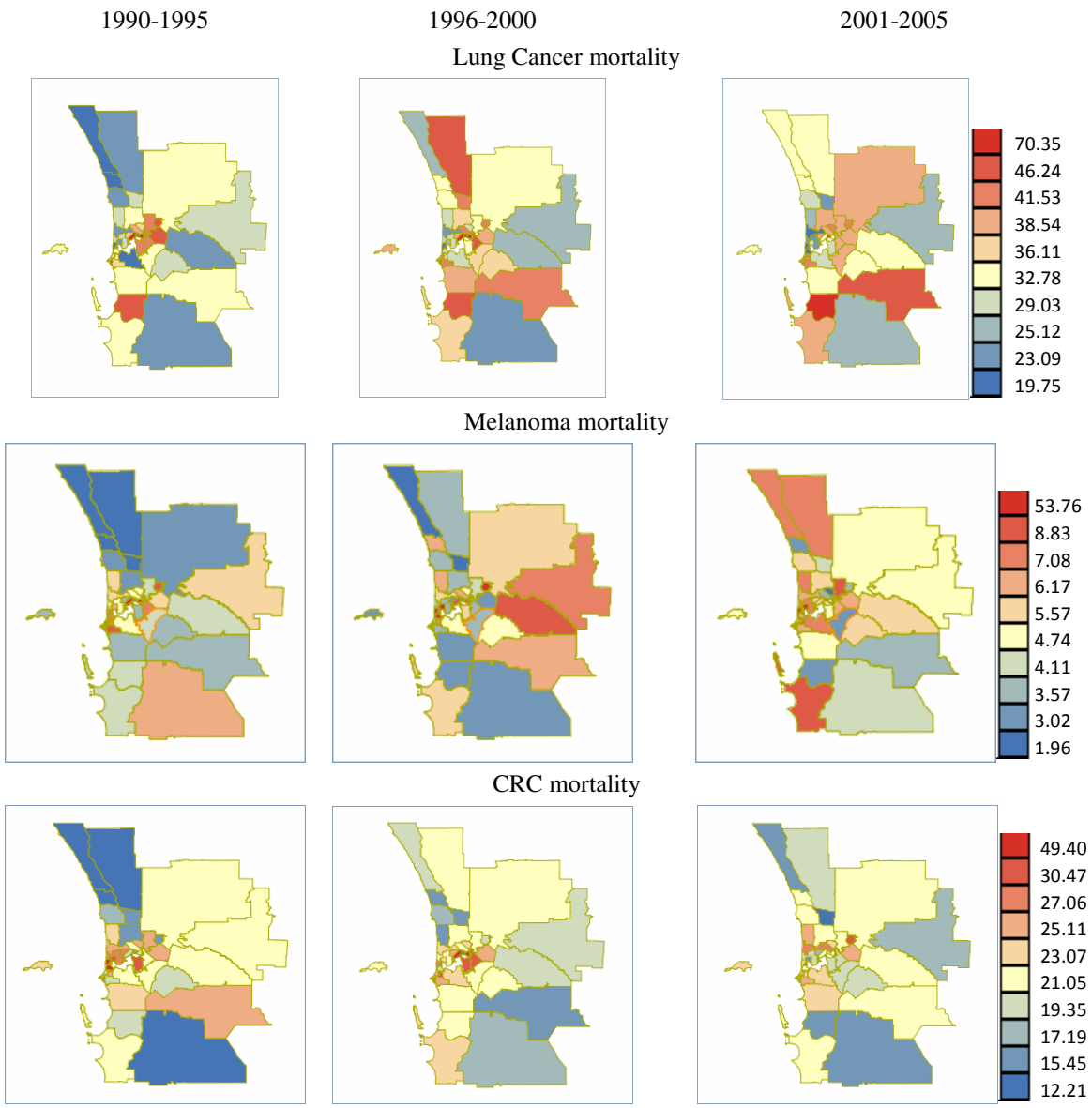


Figure 10 Age-adjusted rates for cancer mortality during three time periods.

Next we will consider the maps of the age-adjusted rates for sex specific cancers: breast and prostate cancers. For the incidence, overall prostate cancer incidence rates are relatively high compared with breast cancer in western part of the Perth metropolitan area (Figure 11). However, breast cancer rates are higher than prostate cancer rates in most of the eastern part of the study area. The overall higher rates arise in the inner suburbs of Perth. Based on Figure 7 and Figure 11, it can be seen clearly that breast cancer incidence is the most common cancer risk for females while the most common incidence risk for males is prostate cancer. For the mortality, it seems that breast cancer has higher overall mortality rates than others in western part of Perth metropolitan area. Generally speaking, the sex

specific cancer mortality rates are relatively low compared with lung cancer mortality (Figure 9).

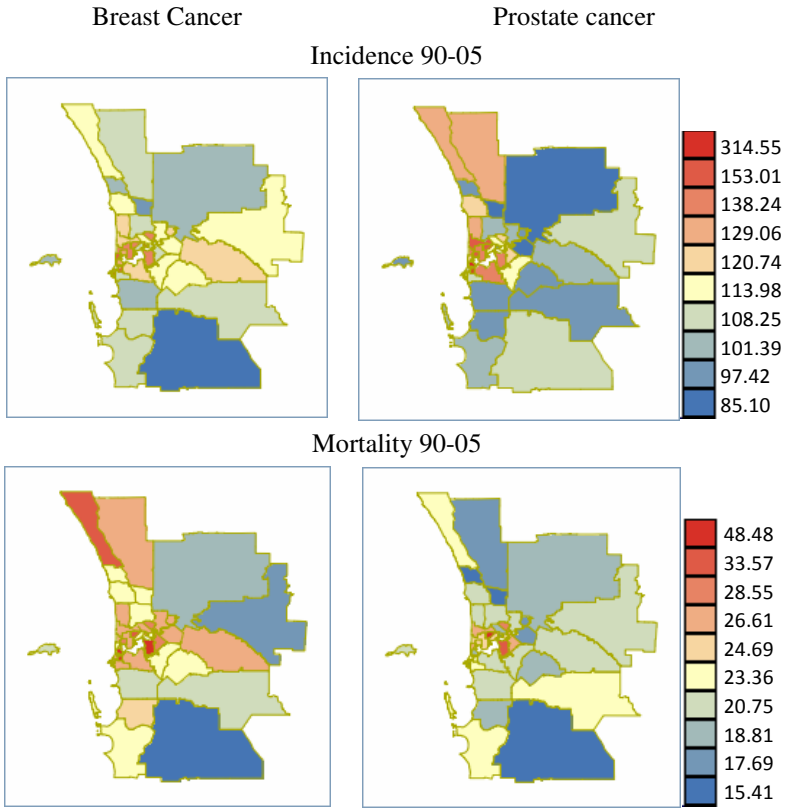


Figure 11 Age-adjusted rates for breast and prostate cancer in 1990-2005.

The cancer rates for the three time periods 1990-1995, 1996-2000 and 2001-2006 are shown in Figure 12. Breast cancer incidence increases steadily in these three periods in the Perth SLAs and the incidence is higher in the inner suburbs of metropolitan Perth. For prostate cancer incidence, it seems that there was a slight decrease in cancer rates around Perth metropolitan by the second period and a significant increase in cancer rates across Perth metropolitan area by the third period.

In comparison, higher mortality rates for breast cancer mortality are mainly in areas close to Perth city centre where the rates kept decreasing with time. However, mortality rates increased in other areas in three time periods. For prostate cancer, the overall mortality rates are relatively high in the second period and low in the third period.

Figure 12 Age-adjusted rates for breast and prostate cancer in three time periods.

From the discussions above, it can be seen that males have a much higher lung cancer incidence risk than females but they have quite similar incidence risk for melanoma and colorectal cancers. The lung cancer incidence risk is lower than melanoma and colorectal cancer risks for both males and females. However, lung cancer mortality risk is much higher than melanoma and colorectal cancer mortality risk. Although people have a very high melanoma cancer incidence risk, the mortality risk is very low compared with other cancers.

4.6.2 Population at risk for geostatistical analysis

In this section, the population at risk for the five types of cancer are provided for reference. The population at risk is based on the adjusted rates which are discussed in details in section 4.6.1. The population at risk for lung, melanoma and colorectal cancers are presented in Figure 13-Figure 16. The population at risk for sex specific cancers are given in Figure 17 and Figure 18. See section 4.5 for the calculation of the population at risk.

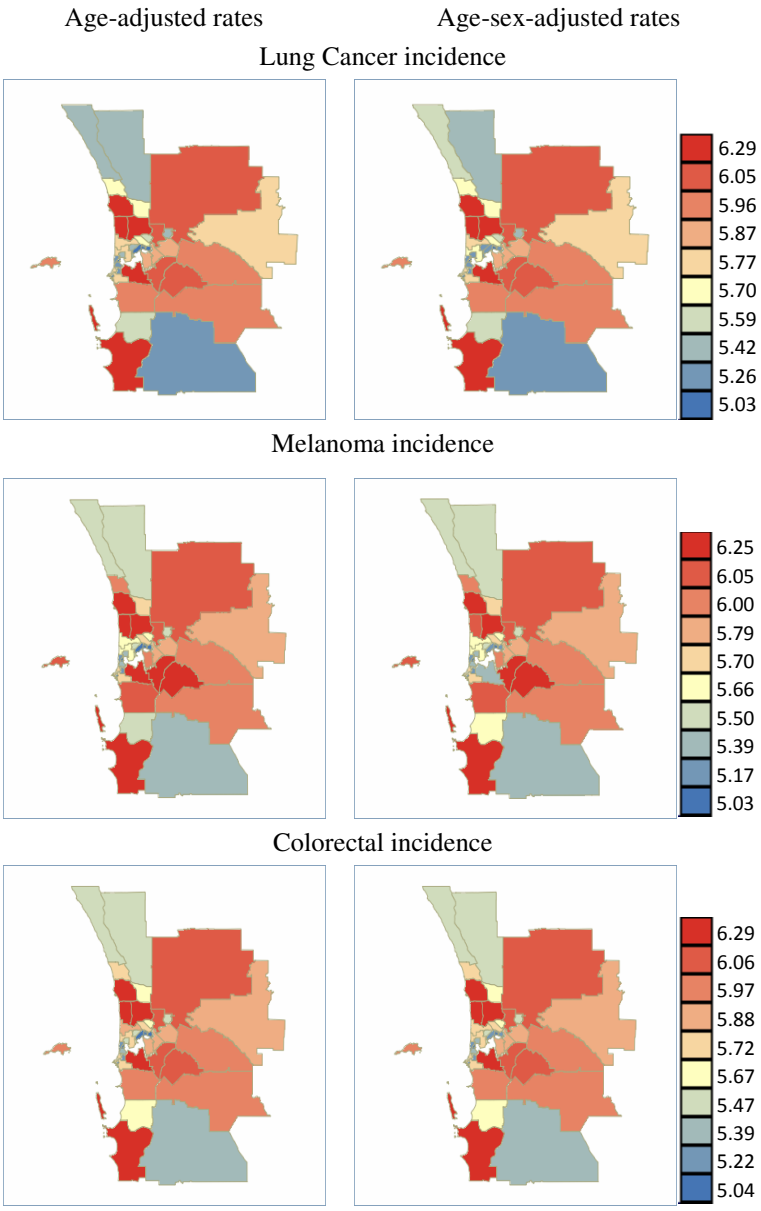


Figure 13 Population at risk (lognormal scale) for cancer incidence.

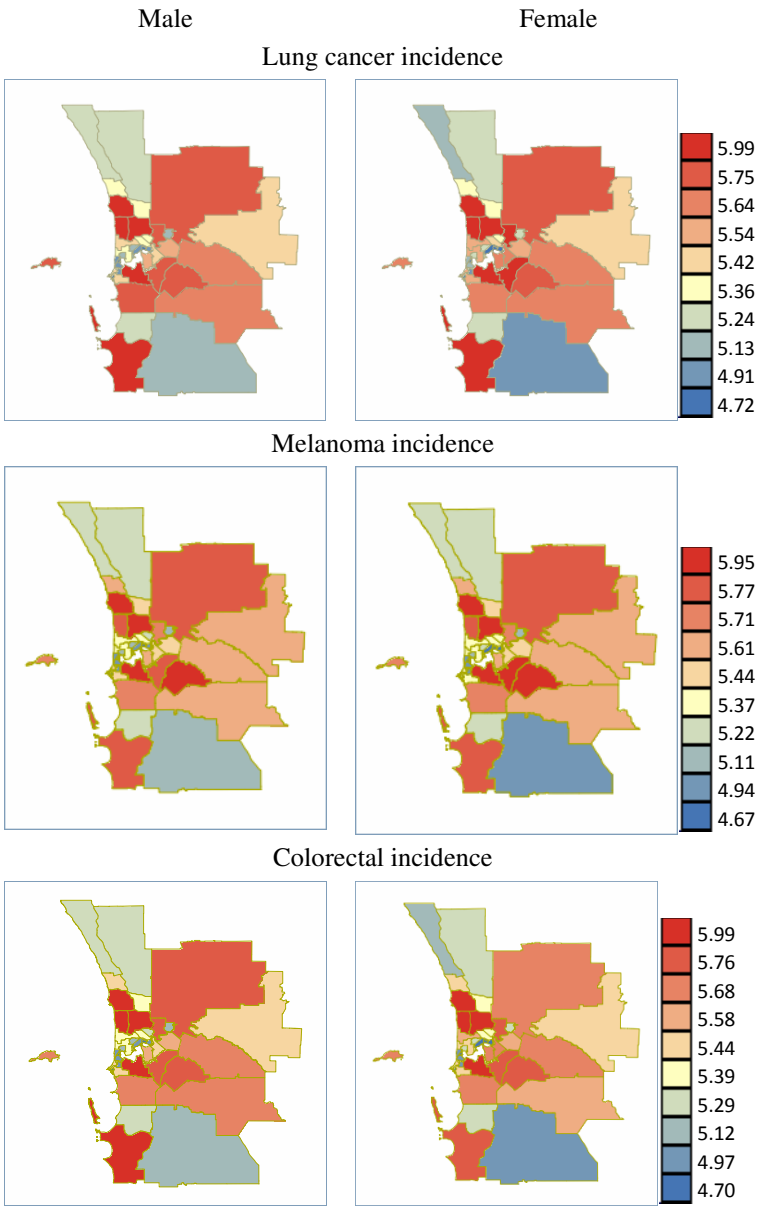


Figure 14 Population at risk by sex (lognormal scale) for cancer incidence.

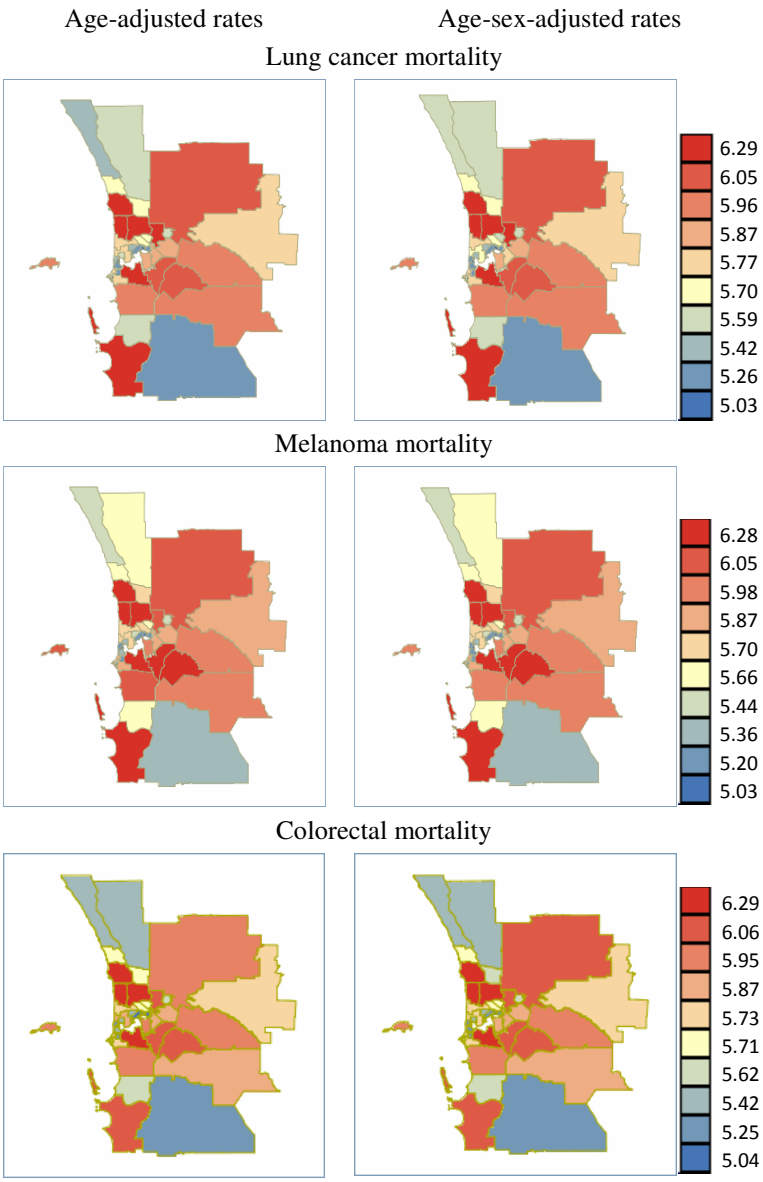


Figure 15 Population at risk (lognormal scale) for cancer mortality.

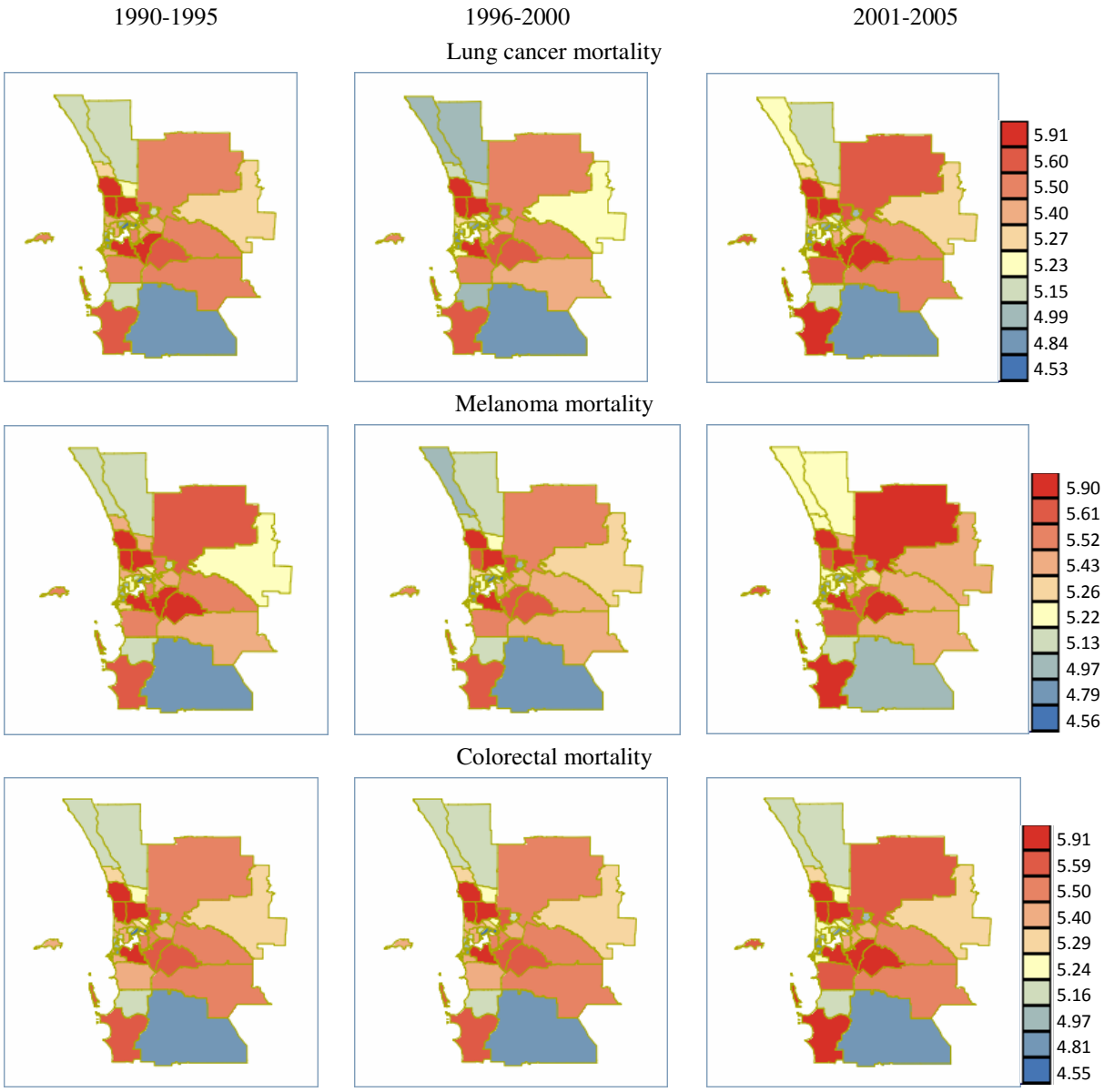


Figure 16 Population at risk (lognormal scale) based on age-adjusted rates for cancer mortality during three time periods.

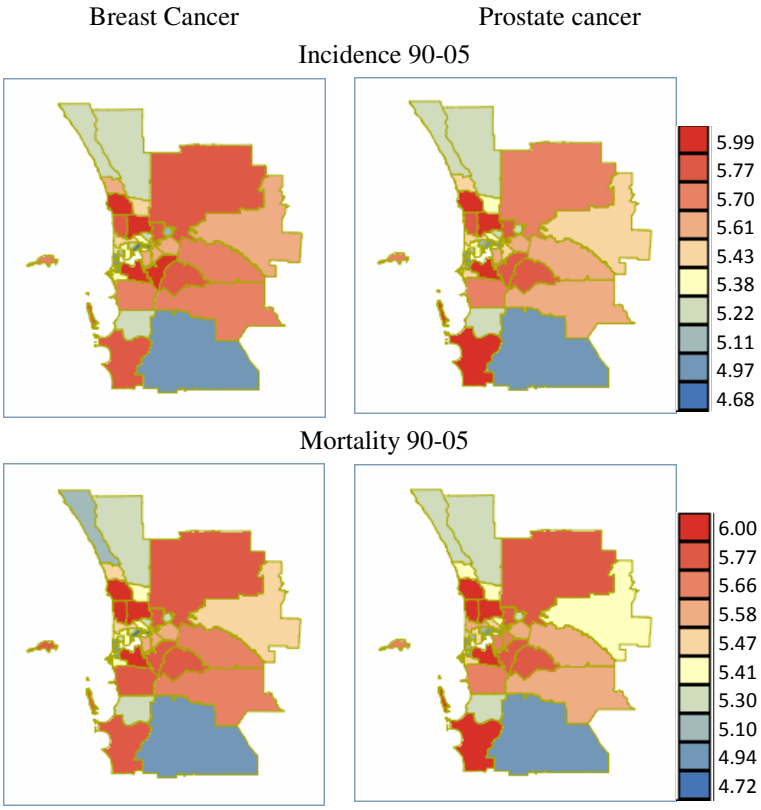


Figure 17 Population at risk (lognormal scale) based on age-adjusted rates for breast and prostate cancers.

Figure 18 Population at risk (lognormal scale) based on age-adjusted rates for breast and prostate cancers.

4.6.3 Covariates for spatial point pattern analysis

For spatial point pattern analysis of the 2005 incidence and mortality data, it is necessary to explore possible covariates which might be useful for the model fitting. The covariates to be used in spatial point pattern analysis for cancers are population density, the proportion of selected age groups and the spatially varying susceptible population intensity.

The population density in each suburb is computed using the total population in the suburb in 2006 divided by the area of the suburb in square kilometres. Based on the population density, the population density image is calculated using the equation (2.45). This pixel image is denoted by `popdens` (Figure 19). It will be used as covariate for lung cancer, melanoma and colorectal cancer model fitting. From Figure 4 it can be seen that people of age 50 or above (50+) have a much higher chance for lung cancer incidence than other age groups. The pixel image (see Figure 19) based on the proportion of this age group is denoted by `aged`. It will be used as covariate for lung cancer model fitting as well. According to Figure 4, pixel image `aged` is used as covariate for colorectal cancer model fitting as well. Similarly, pixel image `aged30+` (the proportion of people aged 30 or over) shown in Figure 19, will be used as covariate in modelling melanoma cancer. The pixel images `popdens`, `aged` and `aged30+` are calculated according to population data shown in Table 3.

For the sex specific cancers, the pixel image `agedf` (Figure 20) to be used as covariate for breast cancer model fitting is based on the proportion of females aged 40 above according to Figure 4. Another pixel image, `popdensf` (Figure 20) based on the overall population density of females, will be used for breast cancer model fitting as well. Similarly, pixel images `agedm` (the proportion of males aged 40 or over) and `popdensm` (male population density) shown in Figure 21, will be used as covariates for prostate cancer model fitting. All these images are calculated using the equation (2.45) and are obtained from the year 2006 population data using Gaussian Kernel smoothing. In the further discussion of the thesis, these covariates will be referred to by their pixel image name. The pixel images

popdensf, agedf, popdensm and agedm40+ are calculated according to population data shown in Table 4.

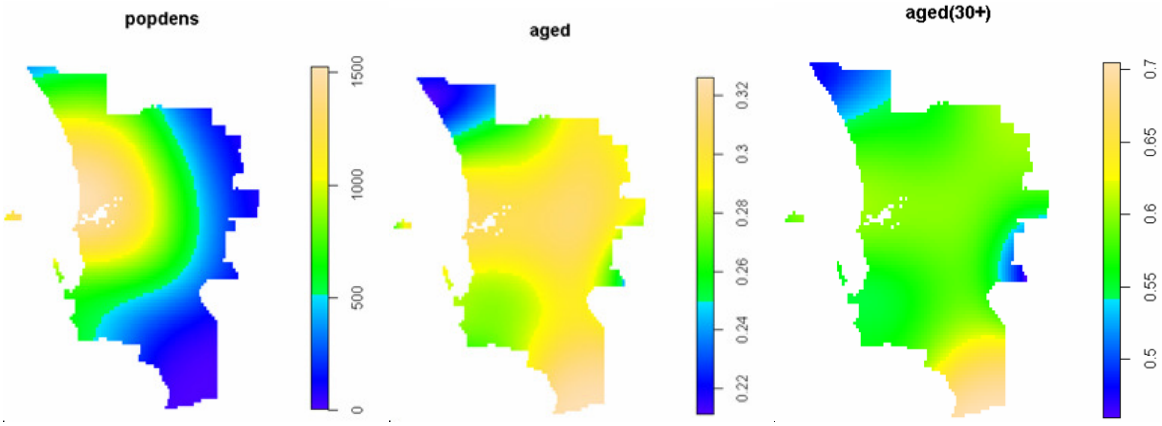


Figure 19 Kernel smoothing estimate of the population density (left, denoted by popdens), the percentage of people aged 50+ by suburb in 2006 (middle, denoted by aged) and the percentage of people aged 30 or above (right, denoted by aged30+).

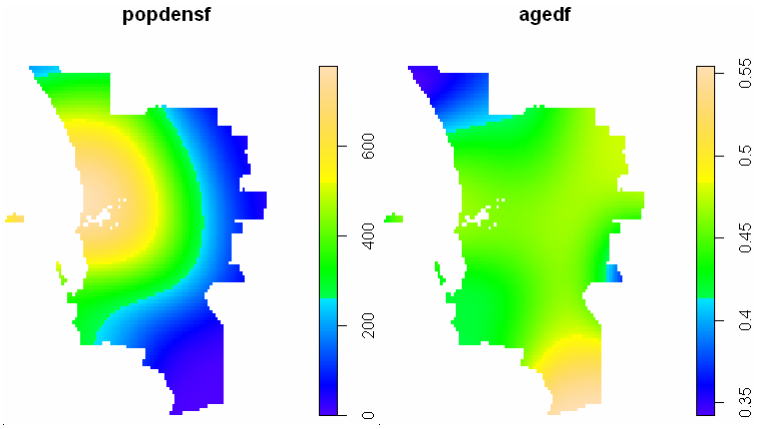


Figure 20 Kernel smoothing estimate of the population density for females (left, denoted by popdensf) and the percentage of females aged 40+ of all females by suburb in 2006 (right, denoted by agedf).

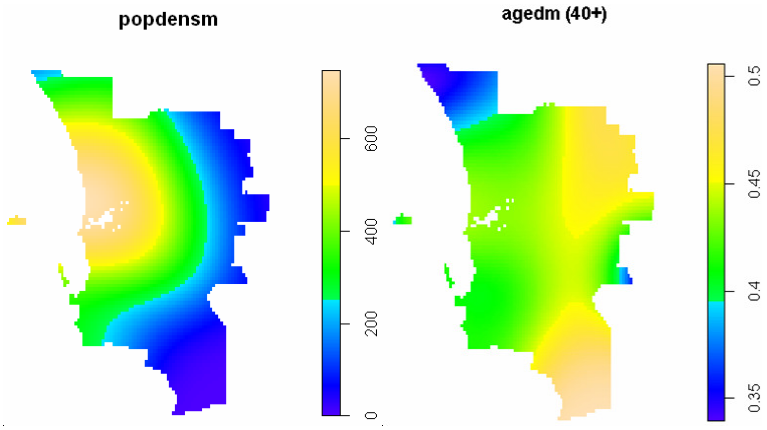


Figure 21 Kernel smoothing estimate of the population density for males (left, denoted by popdensm) and percentage of males aged 40+ of all males (right, denoted by agedm) by suburb in 2006.

Pixel images (`distmap`) based on cancer incidence case locations are calculated using empty space distances. The individual images from 1990-2005 will be used as covariates to reflect the spatially varying susceptible population density for each corresponding cancer model fitting. One example of this application can be found in Diggle (1990), in which the lung cancer point process is used as the spatial variation in risk population intensity for larynx cancer point process analysis. The `distmap`s, based on lung, melanoma, colorectal, breast and prostate cancer incidence cancers during the period 1990-2005, are shown in Figure 22 to Figure 26. The `distmap` images are calculated based on the equation (2.46) and will also be used as covariates for model fitting. At each pixel location the empty space distance is the distance from this location to the closest data point. And so they look quite similar when the colour scales with 10 kilometers apart are used to describe the empty space distance based on cancer incidence point patterns. The incidence case locations in densely populated areas are generally close to each other and so the value for the empty space distance is low. Inversely, the incidence case locations in sparsely populated areas are generally far away from each other and so the value for the empty space distance is high.

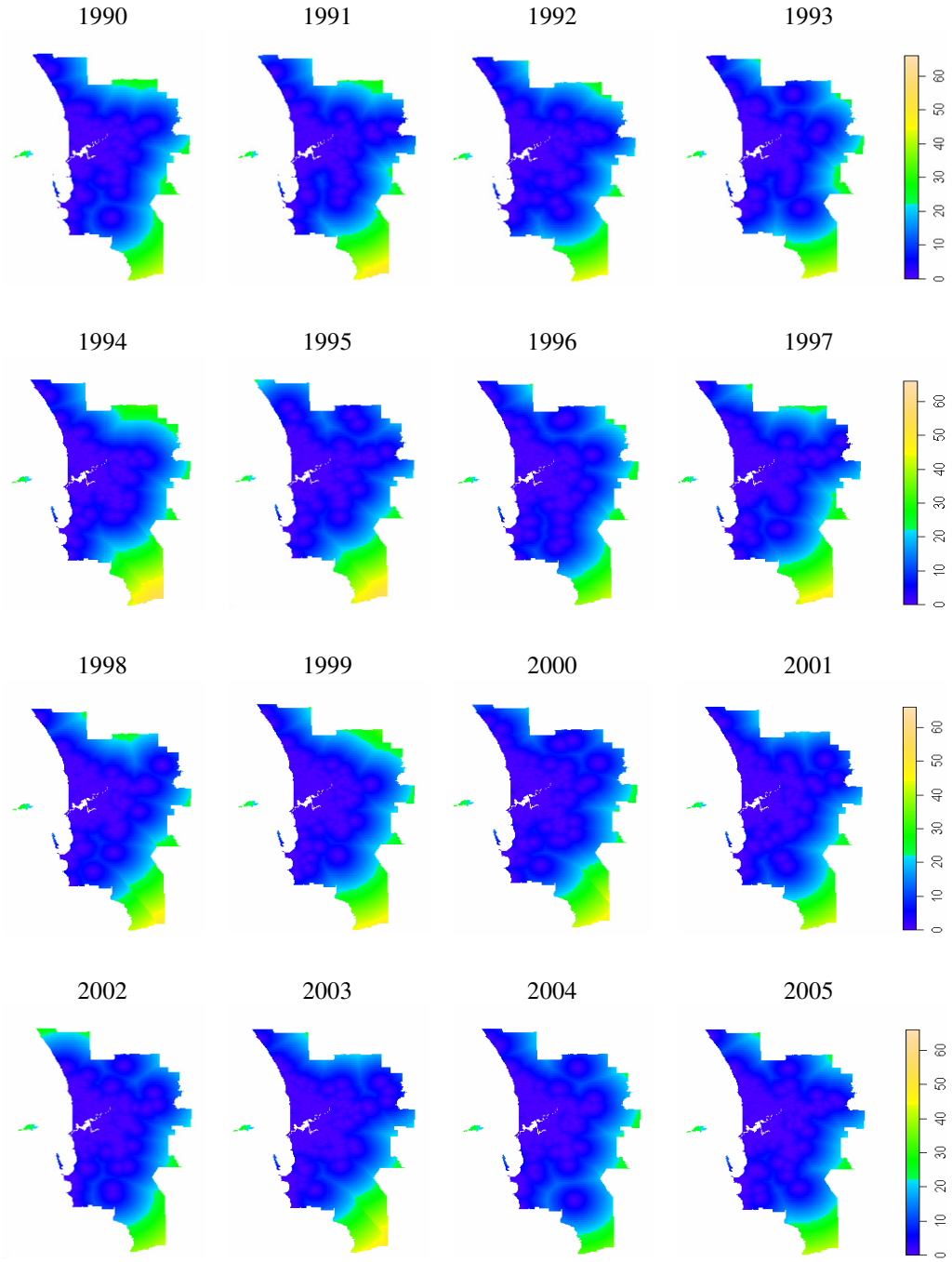


Figure 22 Empty space distances for lung cancer incidence case locations by year, 1990-2005.

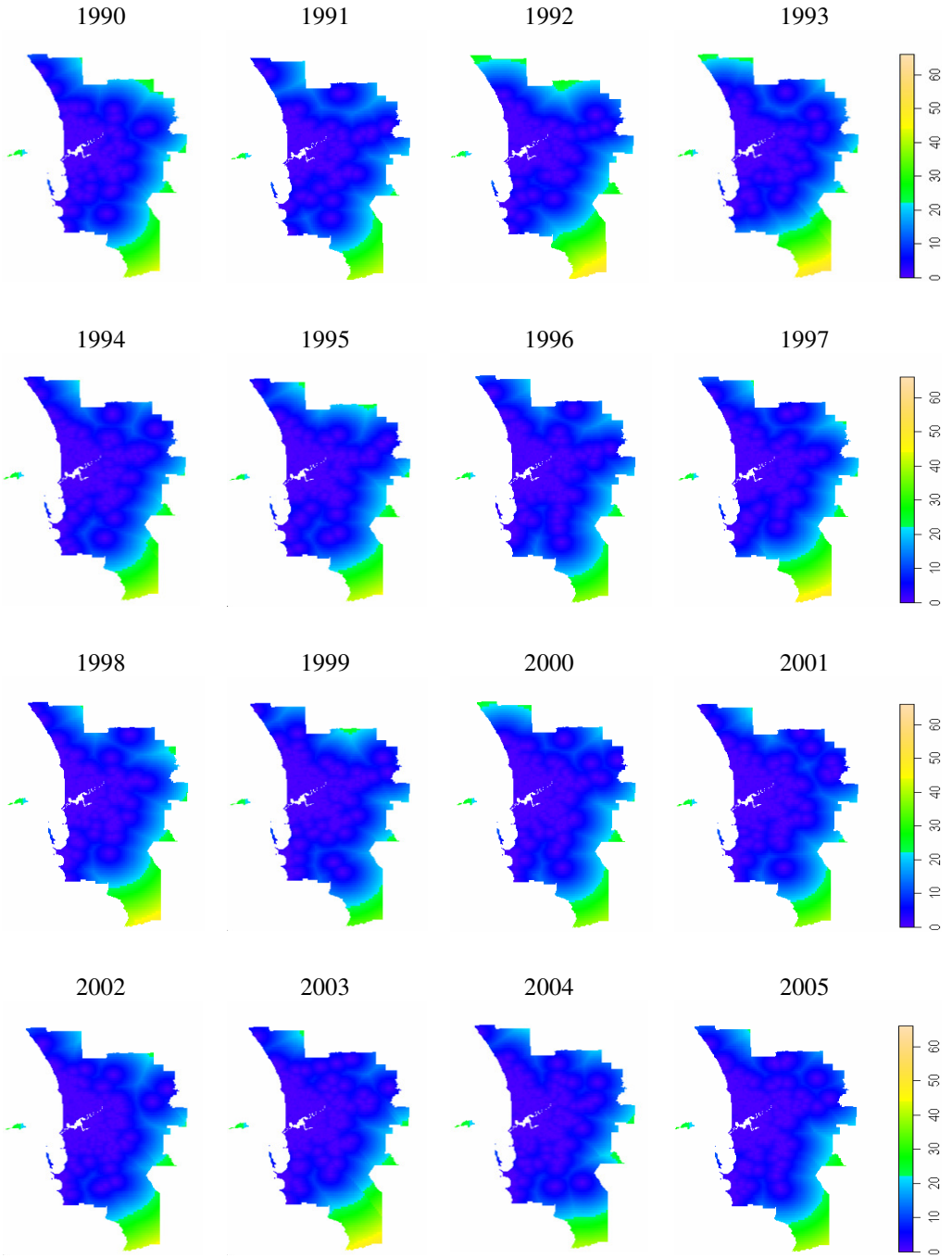


Figure 23 Empty space distances for melanoma cancer incidence case locations by year, 1990-2005.

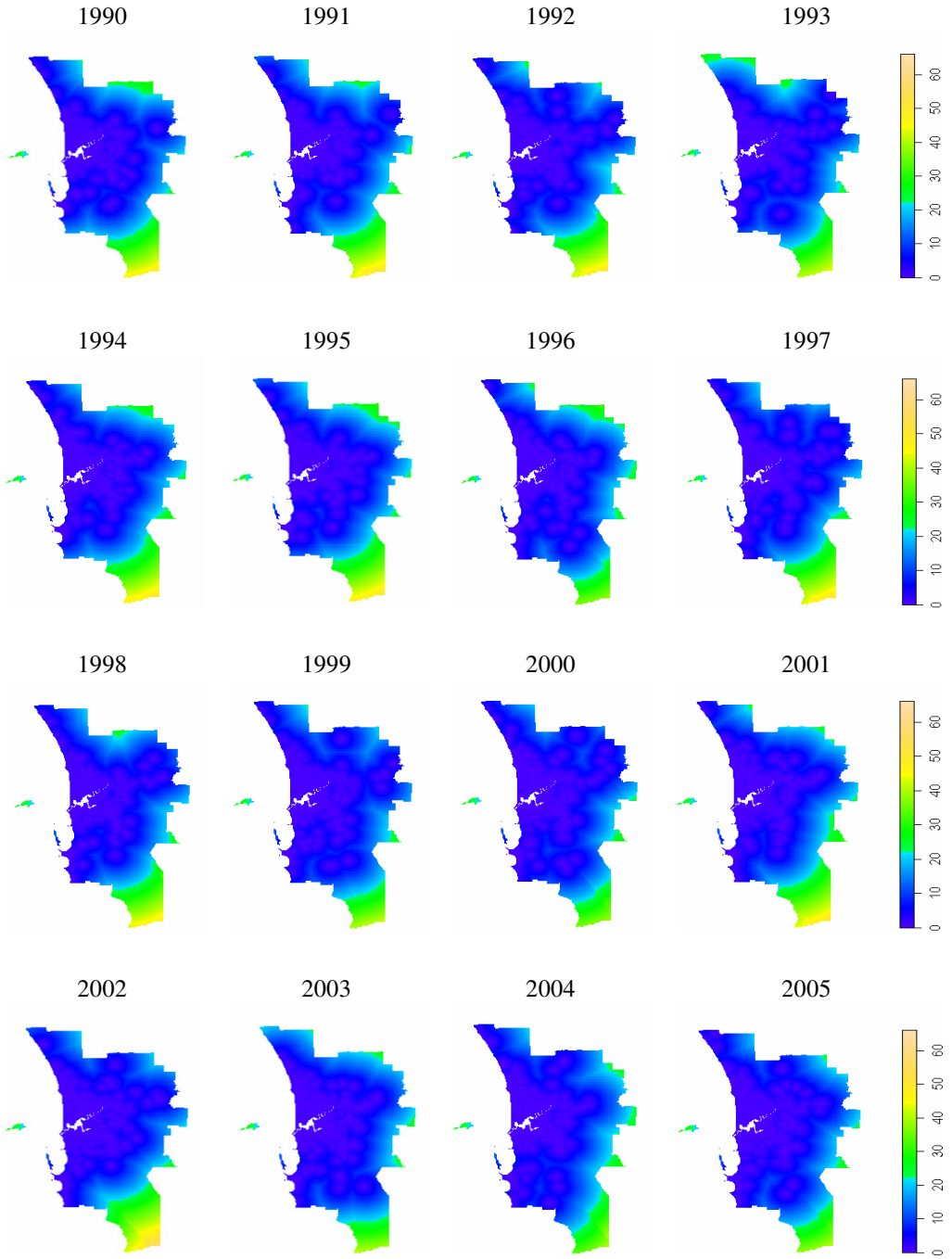


Figure 24 Empty space distances for colorectal cancer incidence case locations by year, 1990-2005.

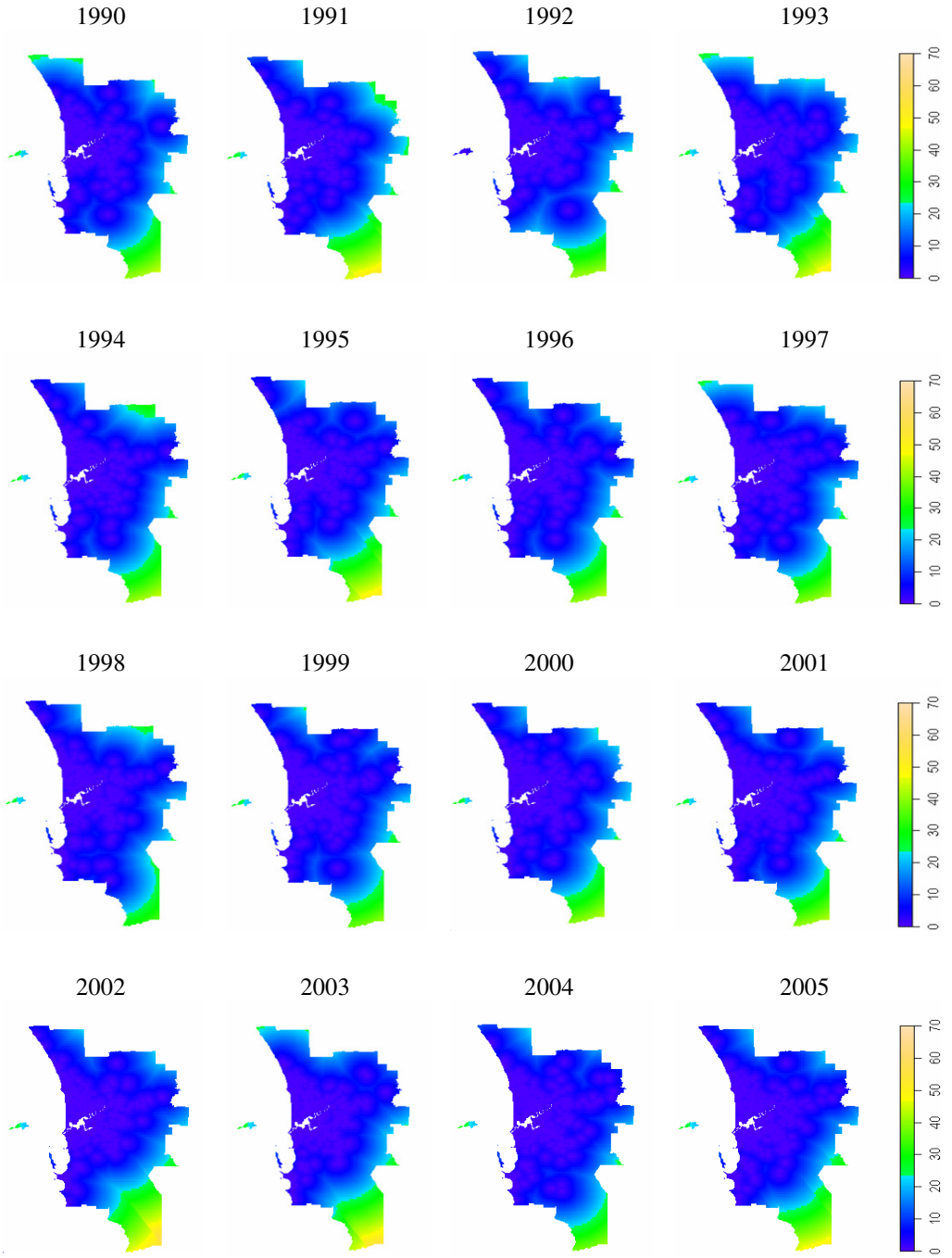


Figure 25 Empty space distances for breast cancer incidence case locations by year, 1990-2005.

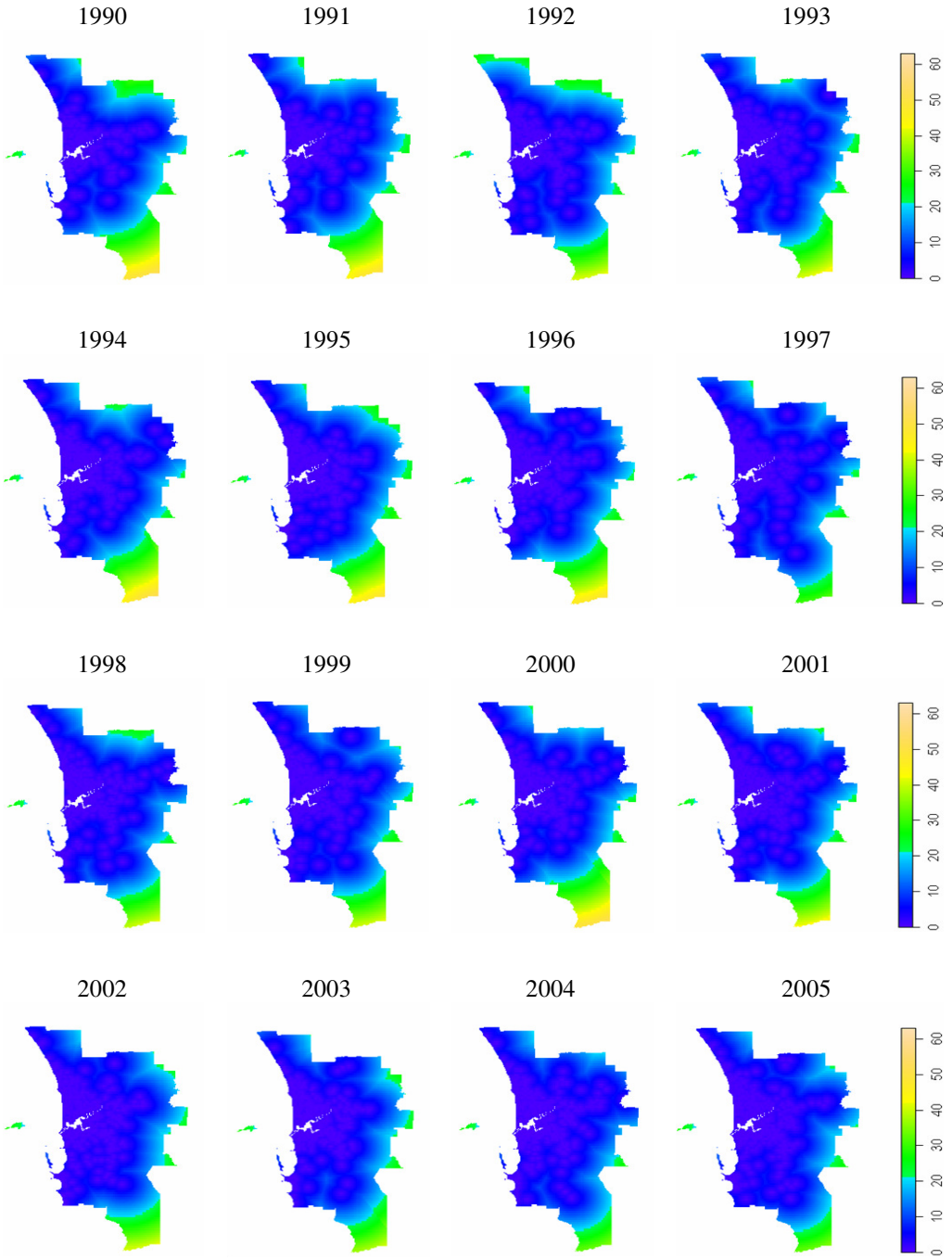


Figure 26 Empty space distances for prostate cancer incidence case locations by year, 1990-2005.

5 GEOSTATISTICAL ANALYSIS

Age-adjustment and age-sex-adjustment are statistical processes applied to rates of cancer in the geostatistical analysis. Age-adjusted rate and age-sex-adjusted rate are a weighted average of the age-specific and age-sex-specific rate compared with crude rate. Cancers occur more often among older people (WACR 1997) and this is also shown in Figure 4. A geographic unit with higher proportion of young people will have a lower cancer incidence and mortality rate than the one with more old people. This is true even if people from the two units have the same cancer risk. Similarly, males are more susceptible to some cancers such as lung cancer than females. The age-adjusted rate and age-sex-adjusted rate are used to remove confounding caused by age or sex. The calculation of the adjusted rates in this chapter is based on (4.1) and (4.4). Geostatistical estimations for cancers are based on the adjusted rates (4.6.1) and population at risk (4.6.2). Semivariograms will be used in the analysis to measure spatial continuity and variability of cancer data. Semivariogram models are fitted to enable kriging analysis and are plotted in meters. Unless otherwise stated, kilometres will be used to describe semivariograms for the sake of convenience.

The size and shape of geographical units in the applications can vary greatly, especially those large rural and sparsely populated areas, which make deconvolution very challenging (Goovaerts 2006a). Uniform discretizing grids across the study area used in deconvolution are not flexible enough because minor changes (coarser grids) in the deconvoluted model have little influence on the kriging estimates (Goovaerts 2008). The population density by SLA in Perth (Figure 27) shows that there is a great difference in population density between small SLAs in the middle of the study area and some large SLAs in the east. And so the deconvolution (2.35) is not used in this study due to these large and sparsely populated SLAs, which do not make distinct difference between computations based on point-support semivariogram model and those based on SLA-based semivariogram model.

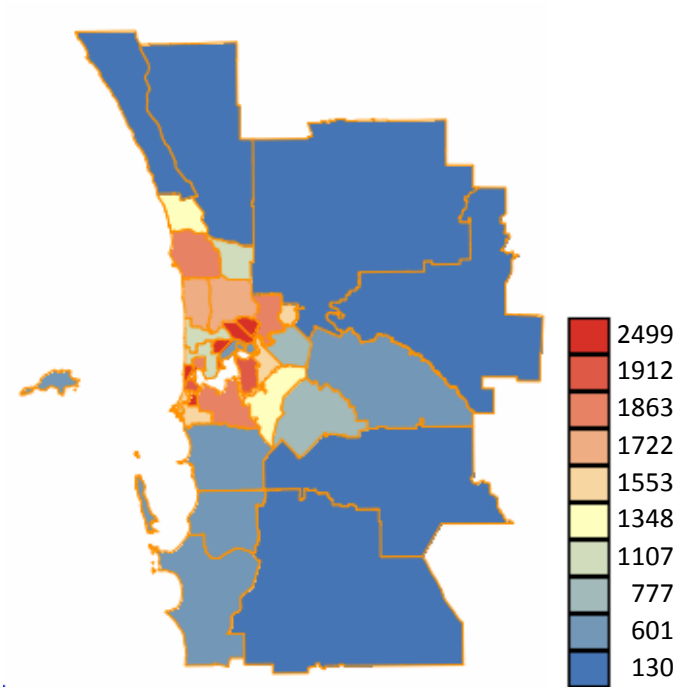


Figure 27 Population density per square kilometers in Perth 2000.

5.1 Lung cancer analysis

In this chapter, lung cancer data will be analysed using ATP Poisson kriging. Lung cancer incidence data during the period 1990-2005 are analysed in the first section. In addition, the incidence data divided by sex: male and female lung cancer incidence, are explored separately. Further discussion will show the differences in lung cancer risk between males and females. Mortality data are analysed in the second section.

5.1.1 Lung cancer incidence analysis

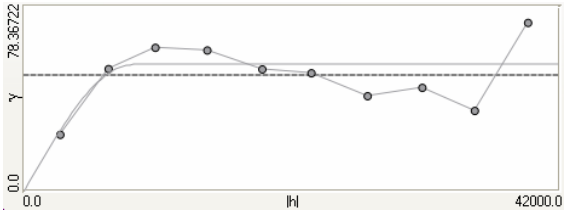
5.1.1.1 Lung cancer incidence analysis in 1990-2005

We have analysed lung cancer incidence data based on adjusted rates in which directional semivariograms are applied (Shao et al. 2009). In that paper, we have shown that the lung cancer incidence data can be analysed by taking into account sex and age. Slight differences can be found between the two adjusted rates and between the two kriging estimates, but the variance map of the age-sex adjusted rates show much less variability and relatively lower value than age-adjusted rates. The variation is spatially dependent and the

risk from age-sex (with zero nugget effect) is possibly more continuous than the risk only from age.

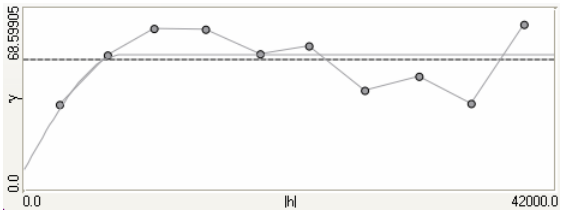
In this section, the estimates are based on adjusted rates in which omnidirectional semivariograms are applied (Figure 28). The nugget effect is zero for age-adjusted rates and 8 for age-sex-adjusted rates. The variogram for age-adjusted rates has greater range and higher sill. This indicates that the spatial continuity and spatial variability for age-adjusted rates and age-sex-adjusted rates are quite similar.

Age-adjusted incidence rates



Number of structures: 1
Nugget: 0
Spherical: Range 8.5, Sill: 53.5

Age-sex-adjusted incidence rates



Number of structures: 1
Nugget: 8
Spherical: Range 7.6, Sill: 43

Figure 28 Omnidirectional semivariograms of lung cancer incidence adjusted rates 1990-2005.

The estimates and variance based on the age-adjusted rates are calculated according to the Equations (2.32) and (2.34) respectively (Figure 29). According to the plots, the estimates based on the age-adjusted rates are relatively smooth and show less variability and the age-sex-adjusted rates map is patchier. However, the variance for the age-adjusted estimates is higher than the one based on age-sex-adjusted rates. The summary statistics for lung cancer incidence are shown in Table 5. Crude rates per 100,000 person-years have a higher mean value and variance value than other rates. ATP estimates are quite close to the adjusted rates and their variance is very low compared with crude rates and the adjusted rates.

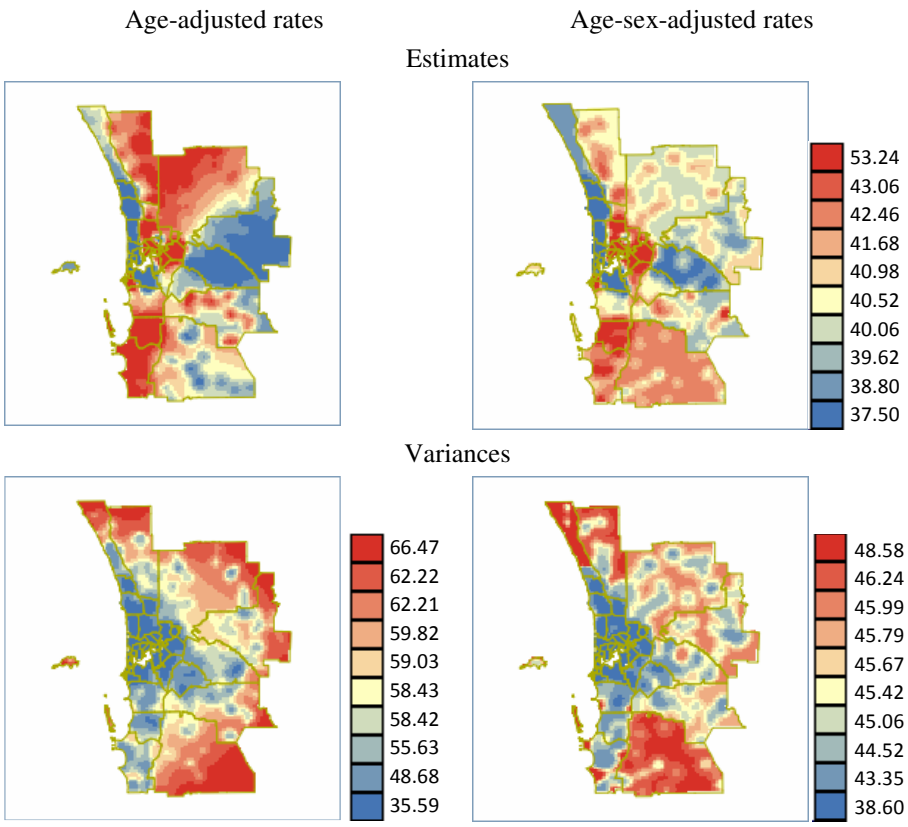


Figure 29 ATP Poisson kriging estimates and variances based on lung cancer incidence adjusted rates.

Table 5 Summary statistics for lung cancer incidence estimates during the period 1990-2005.

Estimator	Lung cancer incidence			
	Mean	Variance	Minimum	Maximum
Crude rates	42.37	148.56	19.66	66.23
Age-adjusted rates	39.78	82.89	21.64	60.74
Age-sex-adjusted rates	39.97	83.16	22.94	62.94
ATP (age) estimator	40.43	10.52	21.45	53.24
ATP (age-sex) estimator	40.4	6.11	23.1	51.24

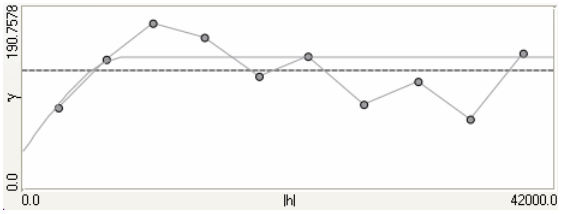
5.1.1.2 Lung cancer incidence analysis by sex

In this part, we will investigate the difference in incidence between males and females.

For experimental semivariograms, Figure 30 shows that the difference in spatial variation between males and females is strong. For the effective distance of spatial dependence, the value for females is more than 3 times the value for males. Semivariogram values of males are nearly 10 times those of females. The semivariogram at zero lag is 40 for males and 0

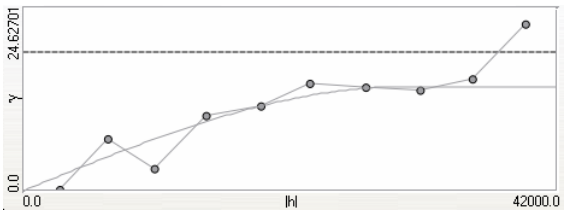
for females. As a result, female lung cancer incidence varies more continuously and has a much stronger spatial correlation than that of males. The risk from age (with zero nugget) is possibly more continuous for females than for males.

Male age-adjusted incidence rates



Number of structures: 1
Nugget: 40
Spherical: Range 8, Sill: 98

Female age-adjusted incidence rates



Number of structures: 1
Nugget: 0
Spherical: Range 30, Sill: 14

Figure 30 Omnidirectional semivariograms of lung cancer incidence age-adjusted rates 1990-2005.

The difference in estimation and the kriging variance (Figure 31) between males and females is strong. For both males and females the incidence is high in the south of the Perth metropolitan area and for males it is also high towards the northeast. The values of ATP Poisson kriging estimates vary between 23 and 80 per 100,000 person-years for males and between 21 and 34 for females. The estimation map of females is relatively smooth. For the kriging variance, the map for females is relatively smooth due to small semivariogram values. However, the male variance map shows great variability. For the maximum variance value of Poisson kriging estimates, it is 115 for males and 33 for females.

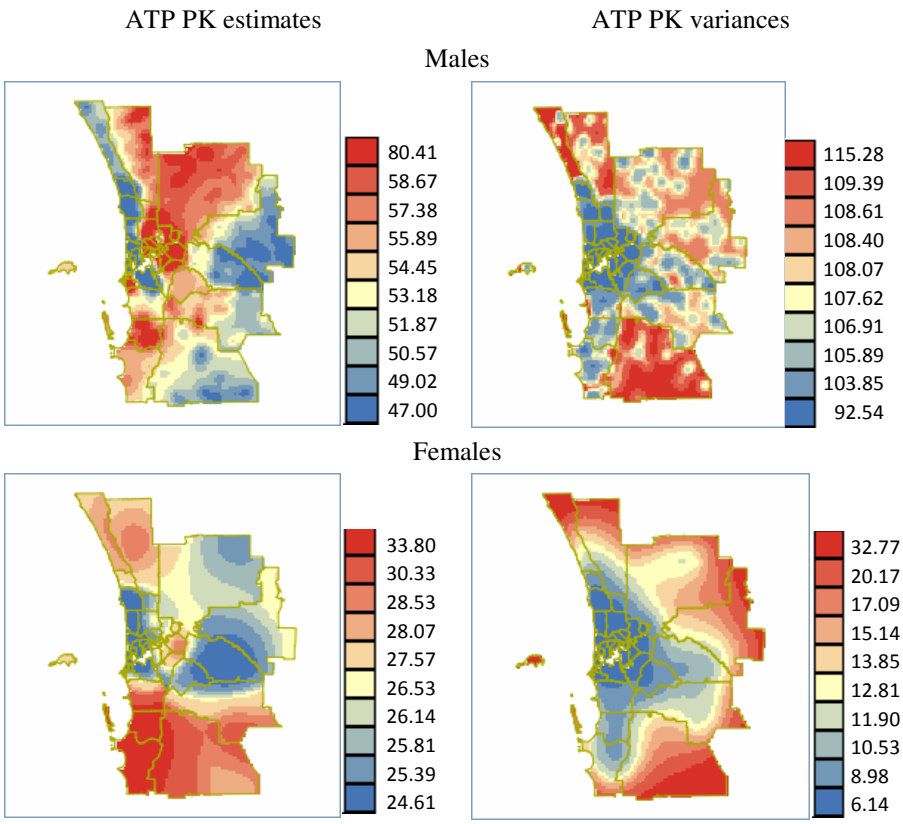


Figure 31 ATP Poisson kriging estimates and variances by sex based on lung cancer incidence age-adjusted rates.

The summary statistics for lung cancer incidence by sex are shown in Table 6. Males have much higher values than females in mean, variance, minimum and maximum for different rates in 1990-2005. All in all, male lung cancer incidence rates are relatively high and show greater variability compared with female lung cancer incidence rates during the period 1990-2005.

Table 6 Summary statistics for lung cancer incidence crude rates, age-adjusted rates and ATP Poisson kriging estimated rates for males and females during the period 1990-2005.

Lung cancer incidence estimator during 1990-2005		Mean	Variance	Min	Max
Crude rates	M	56.44	336.61	17.10	90.57
	F	27.75	79.28	0	44.04
Age-adj rates	M	53.81	195.96	18.50	83.76
	F	26.12	69.86	0	56.89
ATP estimator	M	53.00	27.51	23.33	80.41
	F	27.09	5.15	20.66	33.80

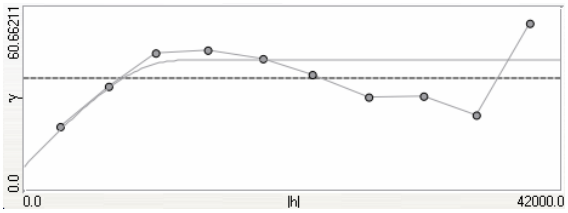
5.1.2 Lung cancer mortality analysis

5.1.2.1 Lung cancer mortality analysis in 1990-2005

In this section, the estimation for lung cancer mortality is based on age-adjusted rates and age-sex-adjusted rates for the period 1990-2005.

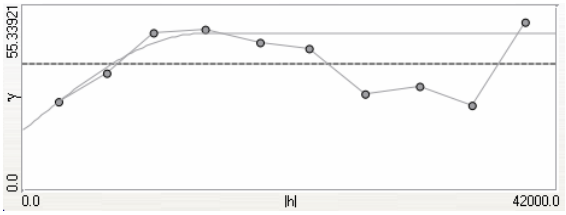
For the semivariograms, Figure 32 indicates that the difference in spatial variation between age-adjusted rates and age-sex-adjusted rates is slight. For age-adjusted rates, the nugget is 8 and the spatial dependence distance is 12km. While for age-sex-adjusted rates, the nugget is 18 and the spatial dependence distance is 14km. The semivariogram values for both rates are quite close.

Age-adjusted incidence rates



Number of structures: 1
Nugget: 8
Spherical: Range 12, Sill: 35

Age-sex adjusted incidence rates



Number of structures: 1
Nugget: 18
Spherical: Range 14, Sill: 29

Figure 32 Omnidirectional semivariograms of lung cancer mortality rates 1990-2005.

The difference in estimation and the kriging variance between adjusted rates is slight (Figure 33). The estimation maps show similar variability and the variance of age-sex-adjusted estimation is slightly higher than the other in the eastern exterior suburbs of Perth.

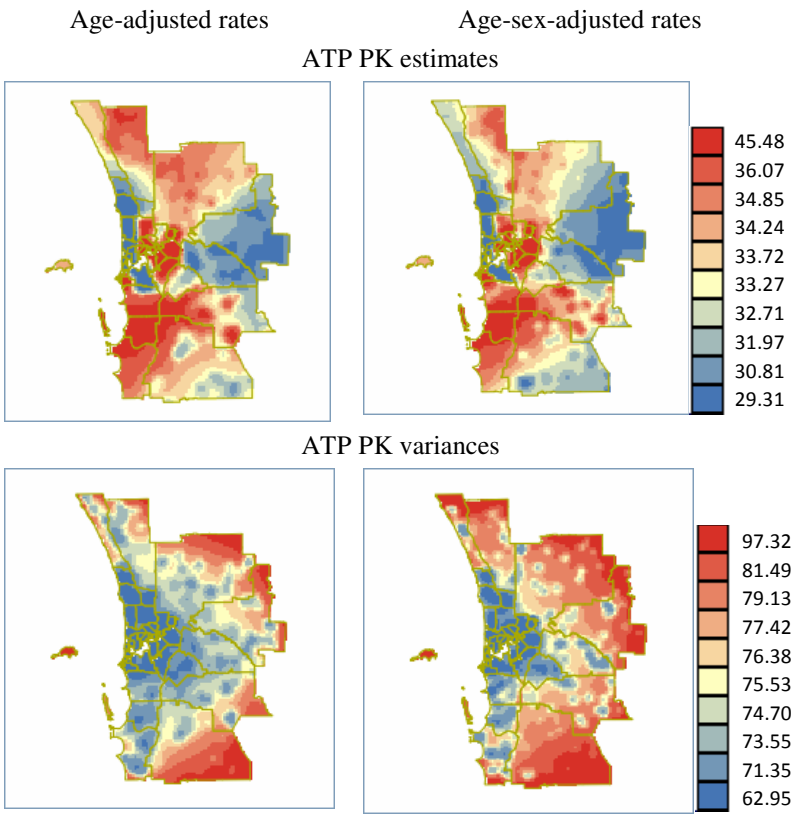


Figure 33 ATP Poisson kriging estimates and variance of lung cancer mortality based on age-adjusted (left) and age-sex-adjusted rates (right).

The details of summary statistics for lung cancer mortality in 1990-2005 are shown in Table 7. Crude rates have a higher mean value and variance value than other rates. ATP estimates are quite close to adjusted rates and their variance is relatively low.

Table 7 Summary statistics for lung cancer mortality estimates during the period 1990-2005.

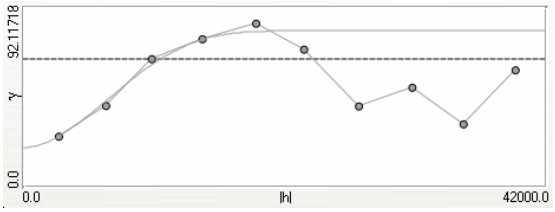
Estimator	Lung cancer mortality			
	Mean	Variance	Minimum	Maximum
Crude rates	35.93	125.37	16.85	55.68
Age-adjusted rates	33.5	64.74	13.15	51.16
Age-sex-adjusted rates	33.59	59.05	13.9	48.81
ATP (age) estimator	33.20	8.24	16.86	43.91
ATP (age-sex) estimator	32.76	8.35	16.64	45.48

5.1.2.2 Lung cancer mortality analysis for three periods: 1990-1995, 1996-2000 and 2001-2005

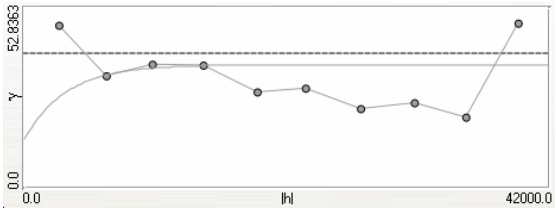
In this part, we will look at lung cancer mortality in three time periods.

For semivariograms, Figure 34 indicates that the difference between three periods is strong. The spatial correlation in the first period is very strong and the spatial dependence distance extends to 22km. Semivariogram values are the highest in the first period and the lowest in the second period. In addition, the risk semivariograms are not well structured in the second period and display a much shorter range of autocorrelation, and so the spatial correlation is very weak. The nugget effect is zero only in the third period and and 20 in the first period. This indicates that lung cancer mortality risk in the last period is probably more continuous and in the first period shows great variability.

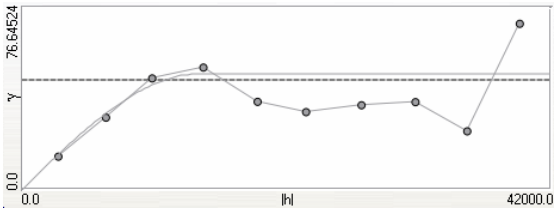
Age-adjusted mortality rates 1990-1995



Age-adjusted mortality rates 1996-2000



Age-adjusted mortality rates 2001-2005



Semivariogram model parameters
Number of structures for three models: 1
90-95 Nugget: 20
Cubic: Range 22, Sill: 60
96-00: Nugget: 14
Exponential: Range 10, Sill: 22
01-05: Nugget: 0
Spherical: Range 14, Sill: 48.5

Figure 34 Omnidirectional semivariograms of lung cancer mortality age-adjusted rates in 1990-1995 (top left), 1996-2000 (top right) and 2001-2005 (bottom).

The difference in estimation and the kriging variance for the three periods (Figure 35) is strong. For the estimated mortality rates, they had an increase from the first to the second period. However, the difference between the second period and the third period is not strong. The mortality rates in SLAs close to the Perth city centre are relatively high in each

period. The estimation map is relatively smooth in the third period and shows slightly more variability in other periods. The kriging variance for estimated rates, are relatively low in the second period and very high in most of Perth SLAs in the first period. The variance map shows greatest variability in the first period and the highest value is 161. In contrast, the values are very low in the second period and the highest variance is 82 with slight variability. The variance map is relatively smooth for the third period.

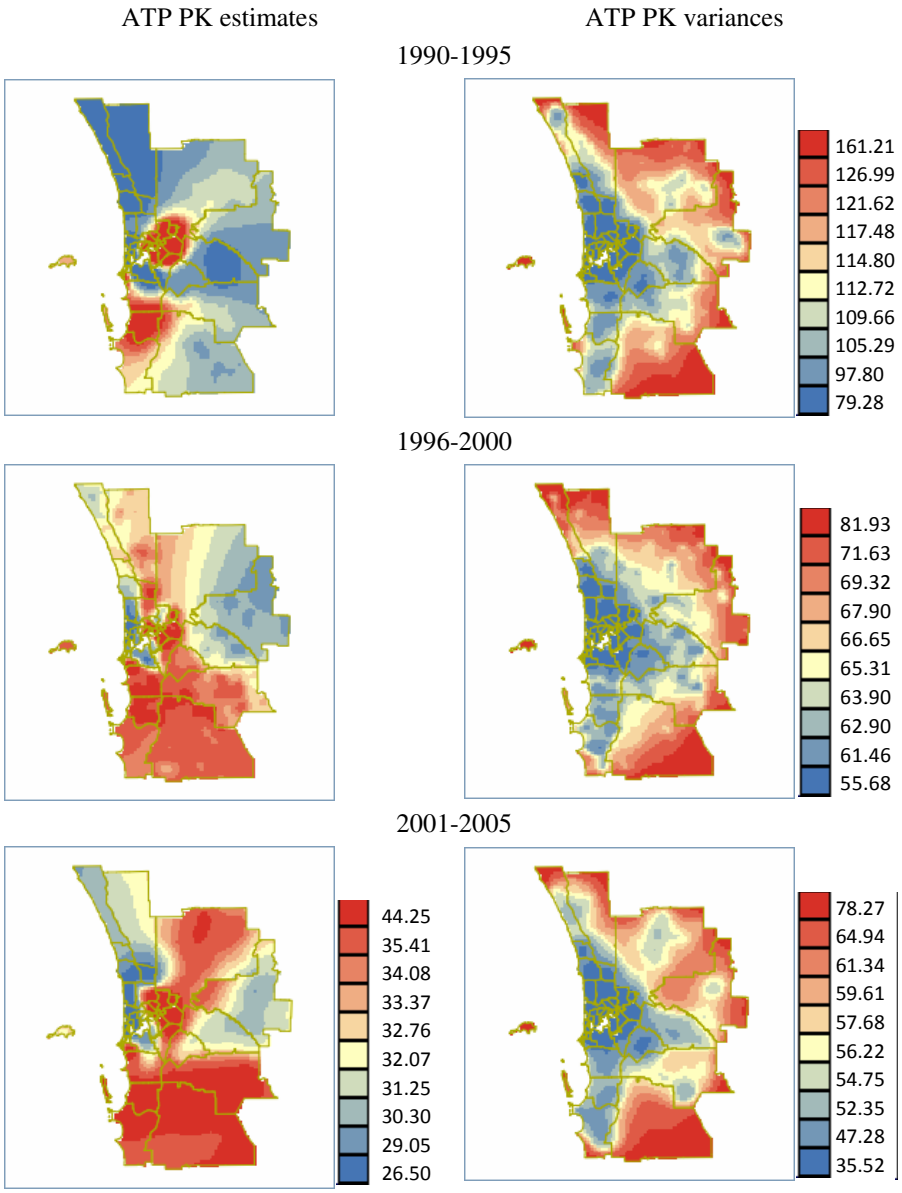


Figure 35 ATP estimated rates (the scale applies to all estimates) and variances based on lung cancer mortality age-adjusted rates.

The details of summary statistics for lung cancer mortality are shown in Table 8. The mean of rates is relatively high in the second period and the rates show great variability in the first period. Overall, lung cancer mortality rates increase with time in some SLAs.

Table 8 Summary statistics for lung cancer mortality crude rates, age-adjusted rates and ATP Poisson kriging estimated rates during 1990-1995, 1996-2000 and 2001-2005.

Lung cancer mortality estimator during: 1990-1995, 1996-2000 and 2001-2005		Age-adjusted rates	Crude rates	ATP estimator
Mean	1990-1995	30.60	35.14	28.73
	1996-2000	33.08	36.61	32.48
	2001-2005	30.82	32.62	33.44
Variance	1990-1995	135.99	203.76	26.01
	1996-2000	138.50	202.52	5.71
	2001-2005	106.17	137.08	10.63
Minimum	1990-1995	8.19	10.59	12.9
	1996-2000	0	0	22.86
	2001-2005	0	0	18.27
Maximum	1990-1995	69.76	68.93	44.25
	1996-2000	70.35	69.08	42.49
	2001-2005	54.69	54.86	40.73

5.2 Melanoma analysis

Melanoma cancer incidence data based on adjusted rates from 1990 to 2005 are analysed in the first section. In the second section, the differences in melanoma cancer risk by sex are investigated. The mortality data for 1990-2005 are analysed in the third section. To get a temporal view, mortality data in three contiguous periods: 1990-1995, 1996-2000 and 2001-2005 are also investigated.

5.2.1 Melanoma incidence analysis

5.2.1.1 Melanoma cancer incidence analysis in 1990-2005

For melanoma cancer incidence risk in the period 1990-2005, Figure 4 in Chapter 4 indicates that the difference between males and females is becoming obvious for people aged 50 above. The analysis in this section is based on age-adjusted rates and age-sex-adjusted rates.

For semivariograms, Figure 36 displays that there is slight difference between age-adjusted and age-sex-adjusted rates. The nugget effect is 0 for both rates. The age-adjusted rates have relatively lower semivariogram values and a slightly shorter range than the age-sex-adjusted rates.

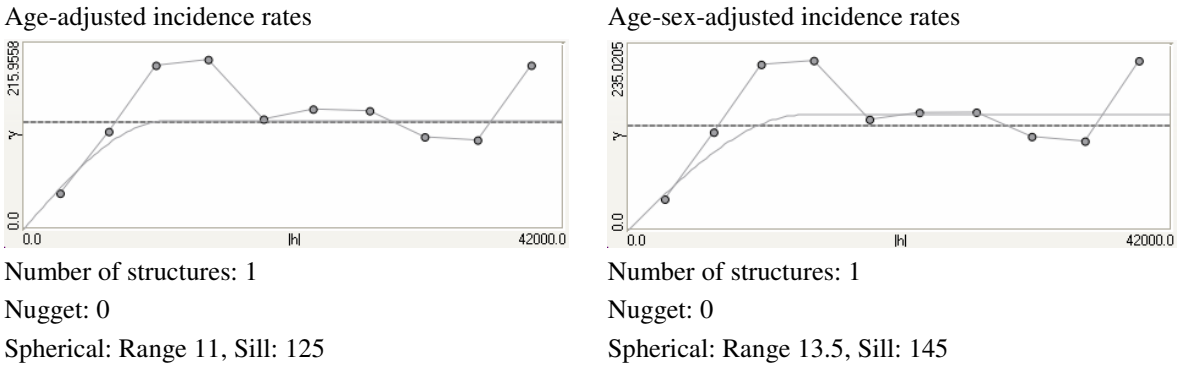


Figure 36 Omnidirectional semivariograms of melanoma cancer incidence adjusted rates in 1990-2005.

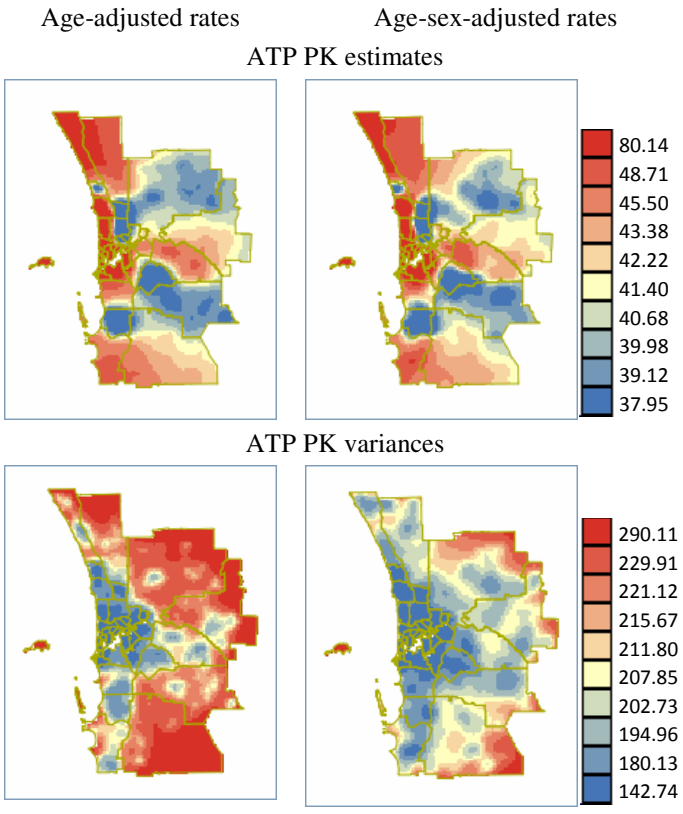


Figure 37 ATP Poisson kriging estimates and variances based on melanoma incidence adjusted rates.

The estimates and variance based on the age-adjusted rates are shown in Figure 37. For both ATP Poisson kriging estimates, higher estimates are mainly along west coast of the

study area. The estimation map for age-sex-adjusted rates shows slightly more variability than the age-adjusted rates. For the variance, age-adjusted rates have higher values than age-sex-adjusted rates in sparsely populated areas (eastern outer suburbs of Perth).

The summary statistics for melanoma cancer incidence during the period 1990-2005 are shown in Table 9. Crude rates show more variability and there is little difference between adjusted rates and ATP estimates.

Table 9 Summary statistics for melanoma cancer incidence estimates during the period 1990-2005.

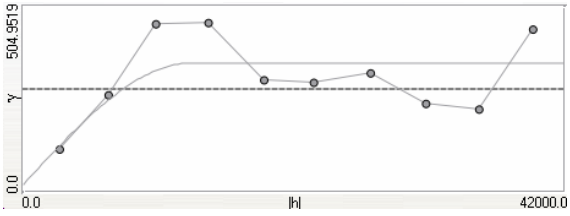
Estimator	Melanoma cancer incidence			
	Mean	Variance	Minimum	Maximum
Crude rates	51.55	366.02	8.03	92.55
Age-adjusted rates	50.07	168.6	28.46	78.33
Age-sex-adjusted rates	50.48	188.38	28.55	80.08
ATP (age) estimator	42.6	31.92	26.61	76.34
ATP (age-sex) estimator	42.67	36	25.01	80.14

5.2.1.2 Melanoma incidence analysis by sex

In this part, we will analyse male and female melanoma cancer incidence in 1990-2005.

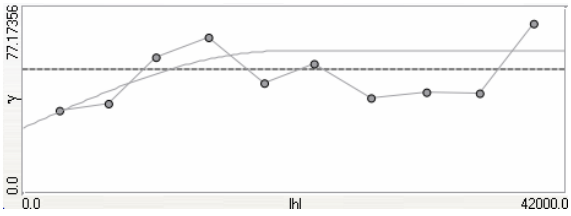
Semivariogram plots for males and females are shown in Figure 38. Semivariogram values of males are nearly six times the values of females. The maximum distance of spatial dependence is 13km for males and 20km for females. The semivariogram at zero lag is 20 for males and 27 for females. Thus female melanoma cancer incidence varies less continuously and spatial correlation is relatively strong compared with males.

Male age-adjusted incidence rates



Number of structures: 1
Nugget: 20
Spherical: Range 13, Sill: 330

Female age-adjusted incidence rates



Number of structures: 1
Nugget: 27
Spherical: Range 20, Sill: 32

Figure 38 Omnidirectional semivariograms of melanoma cancer incidence age-adjusted rates 1990-2005.

The estimation and the kriging variance for males and females are shown in Figure 39. For females the incidence is high in the west and south of the study area. For males it is high in most areas along the coast and also in the middle of the study area. ATP Poisson kriging estimates for males are relatively high, and also the estimation map shows great variability compared with females. The highest estimated rate for males is 108 per 100,000 person-years during the 16-year period but only 57 for females. The kriging variance is much higher for males than for females although the spatial trends are quite similar. The variance map for females is relatively smooth compared with that for males.

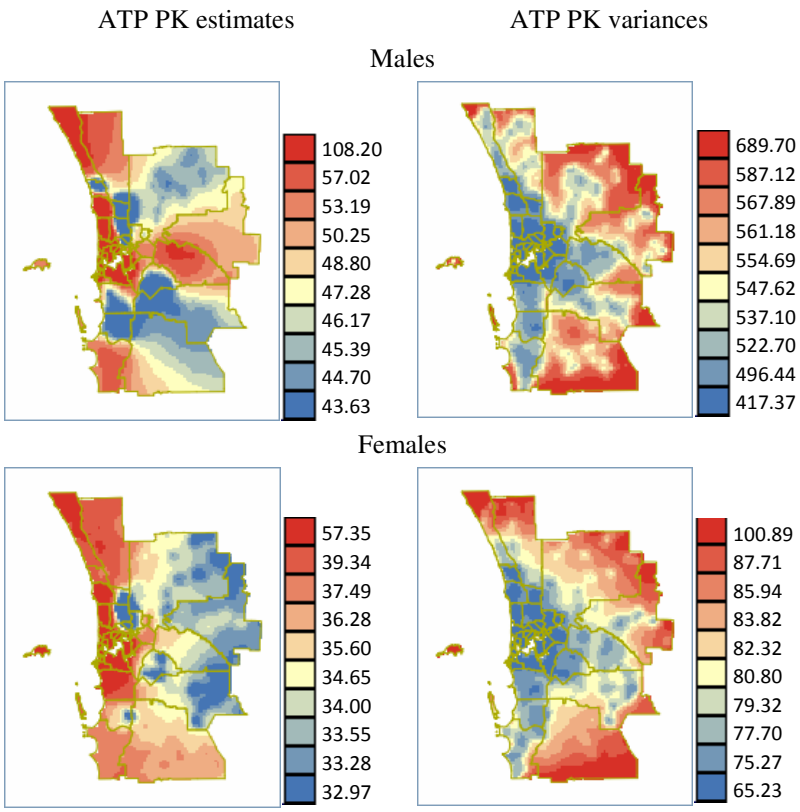


Figure 39 ATP Poisson kriging estimates and variances by sex based on melanoma cancer incidence age-adjusted rates.

The summary statistics for melanoma cancer incidence by sex are shown in Table 10. According to the table, the mean value of male rates is higher than that of females, and the difference in variance between males and females is strong. Overall, male rates show more variability than female rates.

Table 10 Summary statistics for melanoma cancer incidence crude rates, age-adjusted rates and ATP Poisson kriging estimated rates for males and females during the period 1990-2005.

Melanoma cancer incidence estimator during 1990-2005		Mean	Variance	Min	Max
Crude rates	M	63.81	757.07	18.33	126.56
	F	43.06	406.19	0	131.3
Age-adj rates	M	60.28	485.45	19.12	107.75
	F	40.69	213.58	0	99.89
ATP estimator	M	49.36	69.89	24.68	108.2
	F	35.6	9.73	28.73	57.35

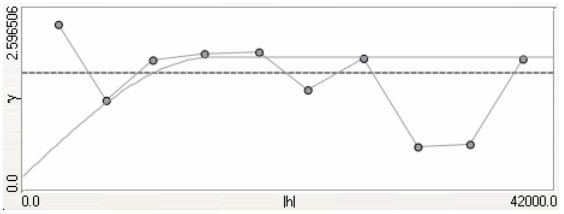
5.2.2 Melanoma mortality analysis

5.2.2.1 Melanoma mortality analysis in 1990-2005

In this section, we analyse melanoma cancer mortality for the period 1990-2005 based on age-adjusted and age-sex-adjusted rates. We then investigate age-adjusted rates for the periods: 1990-1995, 1996-2000 and 2001-2005.

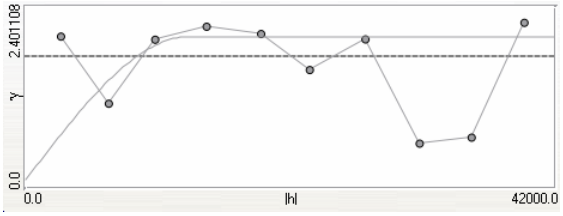
There is slight difference between age-adjusted rates and age-sex-adjusted rates for semivariograms indicated by Figure 40. During the period 1990-2005, the semivariogram values for melanoma cancer mortality are less than 2 and they are very low compared with other cancers. As a result, the variability for melanoma cancer mortality is not as strong as for other cancer types. The range and nugget effect are 15km and 0.2 for age-adjusted rate, and 13km and 0.1 for the age-sex-adjusted rates. The semivariogram values are quite close for the adjusted rates. This indicates that the risk from age-sex is possibly less variable than the risk only from age but spatial correlation is relatively strong for age-adjusted rates.

Age-adjusted mortality rates



Number of structures: 1
Nugget: 0.2
Spherical: Range 15, Sill: 1.7

Age-sex-adjusted mortality rates



Number of structures: 1
Nugget: 0.1
Spherical: Range 13, Sill: 1. 9

Figure 40 Omnidirectional semivariograms of melanoma cancer mortality adjusted rates 1990-2005.

The estimation and the kriging variance for adjusted rates are shown in Figure 41. For the mortality estimation, the rates along west coast are relatively high. For the kriging variance, age-sex-adjusted rates have relatively high values and the map shows slightly more variability than the map based on the age-adjusted rates.

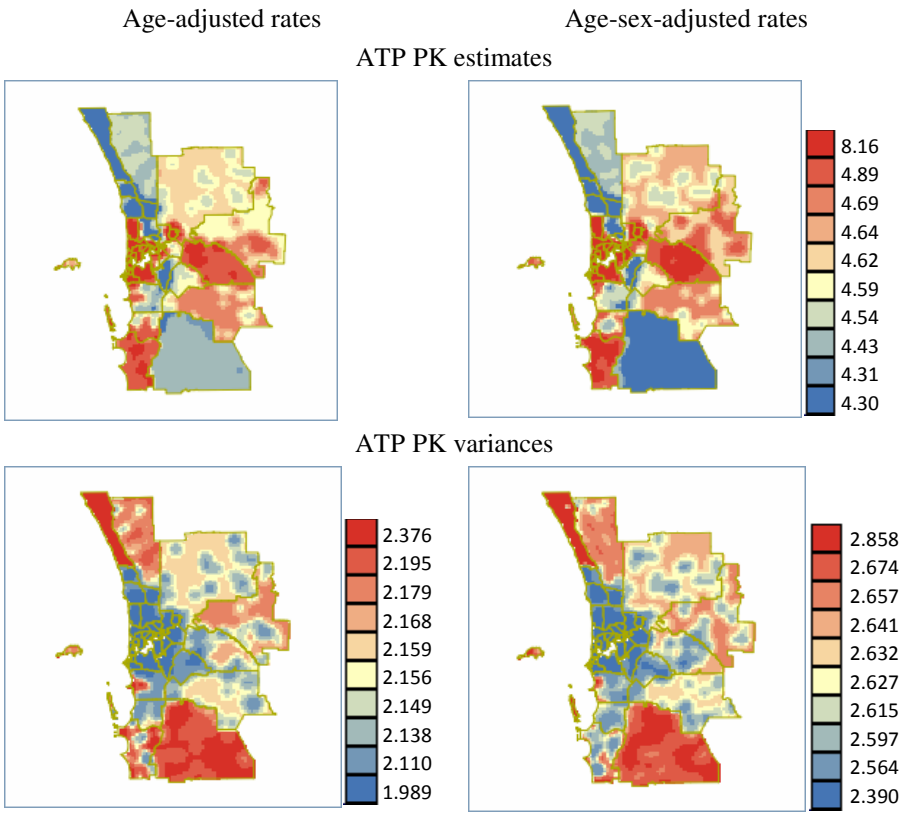


Figure 41 ATP Poisson kriging estimates and variance based on melanoma cancer mortality adjusted rates.

The summary statistics for melanoma cancer mortality in 1990-2005 are shown in Table 11. The mean and variance of different rates based on melanoma mortality is very low compared with other cancers.

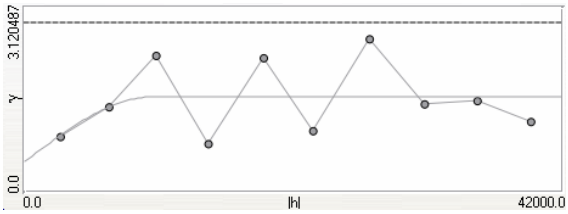
Table 11 Summary statistics for melanoma cancer mortality estimates during the period 1990-2005.

Estimator	Melanoma cancer mortality			
	Mean	Variance	Minimum	Maximum
Crude rates	6.12	11.4	1.87	16.8
Age-adjusted rates	5.68	7.93	2.38	16.59
Age-sex-adjusted rates	5.52	5.35	2.39	12.57
ATP (age) estimator	4.56	0.1	3.33	7.7
ATP (age-sex) estimator	4.56	0.13	3.2	8.37

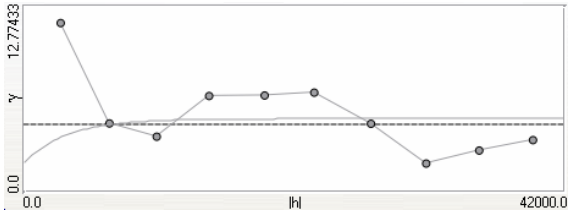
5.2.2.2 Melanoma mortality analysis during three periods: 1990-1995, 1996-2000 and 2001-2005

For experimental semivariograms Figure 42 shows that there is difference between age-adjusted rates in three periods. The variogram in the third period does not show structure. The semivariogram values in the second period are much higher than other periods. Nugget effects are 0.5 in the first period while 2 in the second period. This indicates that cancer mortality shows more variability in the second period.

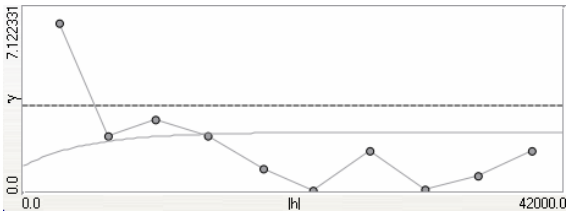
Age-adjusted mortality rates 1990-1995



Age-adjusted mortality rates 1996-2000



Age-adjusted mortality rates 2001-2005



Semivariogram model parameters for three models

Number of structures for three models: 1

90-95 Nugget: 0.5
Spherical: Range 10, Sill: 1.1

96-00: Nugget: 2
Exponential: Range 10, Sill: 3

01-05: Nugget: 1
Exponential: Range 15, Sill: 1.3

Figure 42 Omnidirectional semivariograms of melanoma cancer mortality age-adjusted rates.

The difference of spatial trends is strong as for estimation and the kriging variance in three periods (Figure 43). The estimates increase with time in most SLAs of Perth. In each period the map of estimates shows great variability, and the estimated rates are very low compared with other cancers. For the kriging variance, the maps show great variability in each period and the variance values are relatively high in the second period.

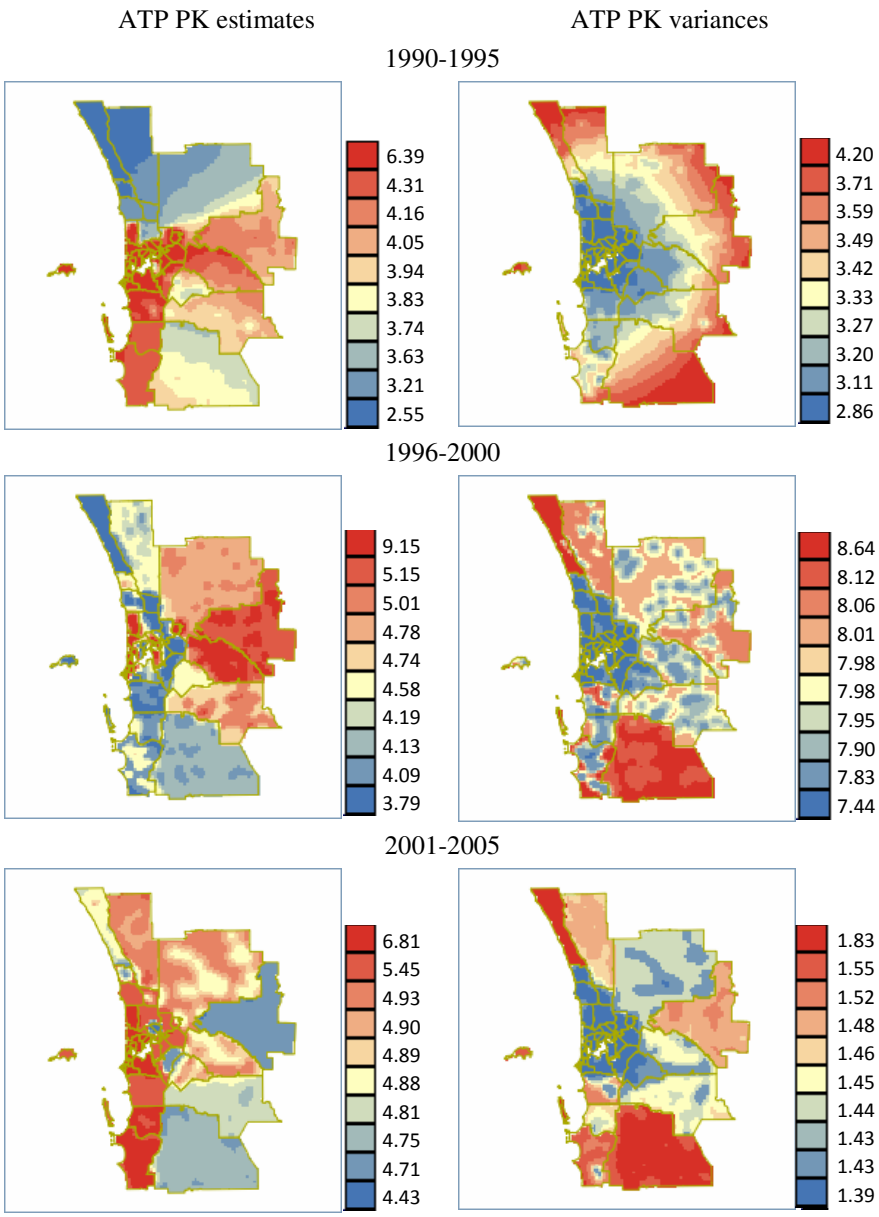


Figure 43 ATP estimated rates and variances based on melanoma cancer mortality age-adjusted rates.

The summary statistics for melanoma cancer mortality during three contiguous periods are shown in Table 12. Mortality rates in the second period are relatively high and so is the variance.

Table 12 Summary statistics for melanoma cancer mortality crude rates, age-adjusted rates and ATP Poisson kriging estimated rates during 1990-1995, 1996-2000 and 2001-2005.

Melanoma cancer mortality estimator during: 1990-1995, 1996-2000 and 2001-2005		Crude rates	Age-adjusted rates	ATP estimator
	1990-1995	5.18	4.47	3.69
Mean	1996-2000	6.7	6.03	4.47
	2001-2005	6.04	5.89	4.89
	1990-1995	13.45	7.81	0.5
Variance	1996-2000	76.35	69.5	0.39
	2001-2005	16.9	34.18	0.1
	1990-1995	0	0	1.67
Minimum	1996-2000	0	0	2.51
	2001-2005	0	0	3.99
	1990-1995	18.69	15.53	6.39
Maximum	1996-2000	53.76	52.16	9.15
	2001-2005	17.9	37.46	6.81

5.3 Colorectal cancer analysis

Colorectal cancer incidence data during the period 1990-2005 is analysed in the first section. Also, in this section we will explore the incidence data divided by sex: male and female. Further discussion will show the difference in colorectal cancer risk between males and females. Mortality data are analysed in the second section.

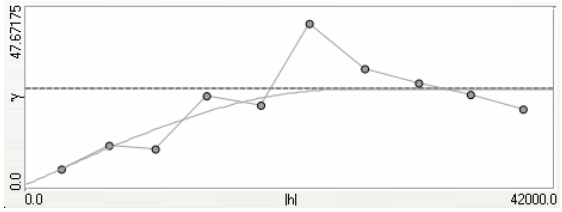
5.3.1 Colorectal cancer incidence analysis

5.3.1.1 Colorectal cancer incidence analysis in 1990-2005

First, the data to be analysed in this section are colorectal cancer data based on age-adjusted rates and age-sex-adjusted rates in 1990-2005.

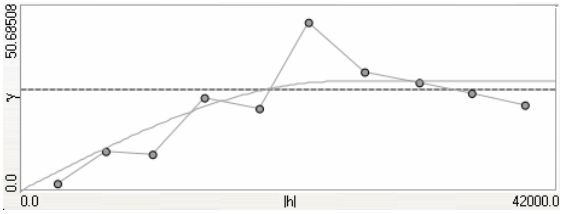
There is only a slight difference between age-adjusted and age-sex-adjusted rates as for semivariograms according to Figure 44. The spatial correlation is the same for both rates and extends to 25km. For age-sex-adjusted rates, the semivariogram values are slightly higher than age-adjusted rates. The nugget effect is strictly zero for age-sex-adjusted rates and 1 for the other rate.

Age-adjusted incidence rates



Number of structures: 1
Nugget: 1
Spherical: Range 25, Sill: 25

Age-sex-adjusted incidence rates



Number of structures: 1
Nugget: 0
Spherical: Range 25, Sill: 30

Figure 44 Omnidirectional semivariograms of colorectal cancer incidence adjusted rates 1990-2005.

For the estimation and the kriging variance, the maps are shown in Figure 45. These two adjusted rates are quite similar. The estimation map and the krigin variance map for age-sex-adjusted rates show a relatively smooth trend compared with the age-adjusted rates.

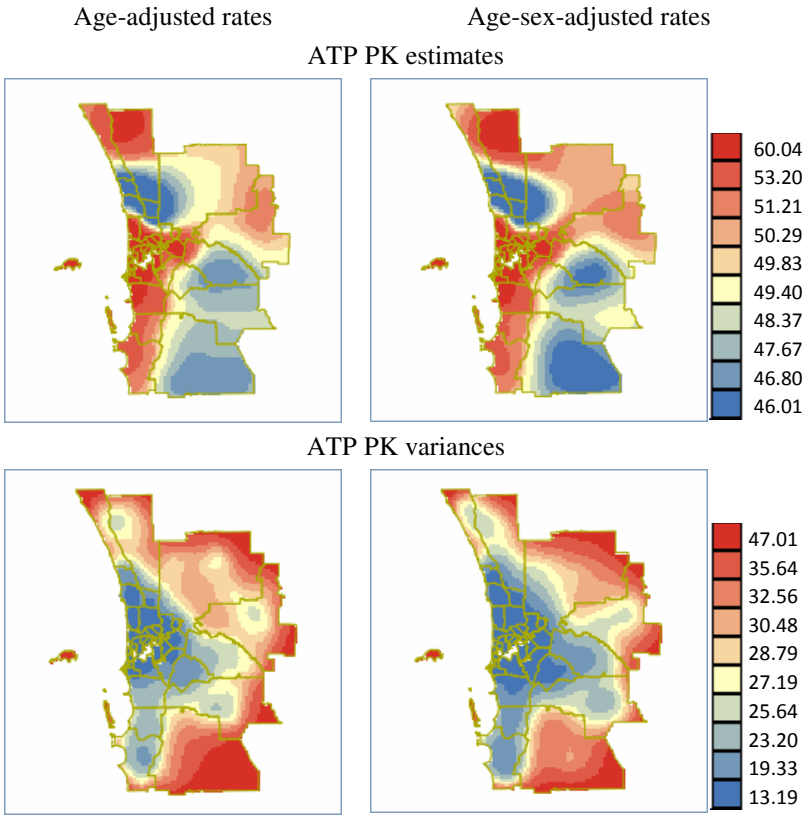


Figure 45 ATP Poisson kriging estimates and variances for colorectal incidence adjusted rates.

The summary statistics for colorectal cancer incidence during the period 1990-2005 are shown in Table 13. For mean value, there is little difference between adjusted rates and

ATP estimates but they are all lower than crude rates. The difference between age-adjusted rates and age-sex-adjusted rates is not significant.

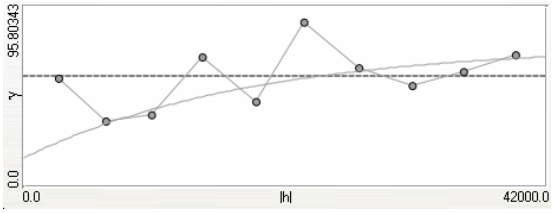
Table 13 Summary statistics for colorectal cancer incidence estimates during the period 1990-2005.

Estimator	Colorectal cancer incidence			
	Mean	Variance	Minimum	Maximum
Crude rates	57.73	350.18	21.59	92.41
Age-adjusted rates	52.48	96.43	21	86.33
Age-sex-adjusted rates	52.36	74.19	23.78	72.27
ATP (age) estimator	49.18	7.24	42.01	59.92
ATP (age-sex) estimator	49.28	8.76	41.1	60.04

5.3.1.2 Colorectal cancer incidence analysis by sex

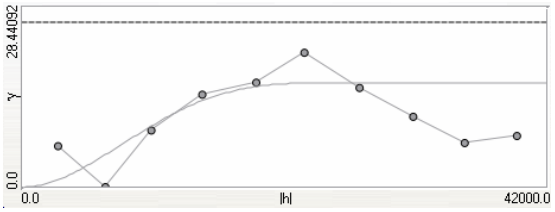
For semivariograms of colorectal cancer incidence, the difference between males and females is distinct in terms of nugget, range and sill according to Figure 46. The semivariogram values for males are nearly five times those of females, and so female colorectal cancer incidence varies relatively continuously. The distance of spatial dependence extends to 55km for males and 26km for females, so spatial correlation for males is relatively strong. The semivariogram at zero lag is zero for females and 15 for males and so the risk from age is possibly more continuous for females than males.

Male age-adjusted incidence rates



Number of structures: 1
Nugget: 15
Exponential: Range 55, Sill: 60

Female age-adjusted incidence rates



Number of structures: 1
Nugget: 0
Cubic: Range 26, Sill: 16.5

Figure 46 Omnidirectional semivariograms of colorectal cancer incidence age-adjusted rates 1990-2005.

The estimates and the kriging variance for males and females are displayed in Figure 47. The estimates for both males and females are relatively high in the inner suburbs and southwest of Perth. The kriging variance is high in eastern exterior suburbs of Perth for

males and females. Males have much higher estimated rates and variances than females, and the maps for females are relatively smooth.

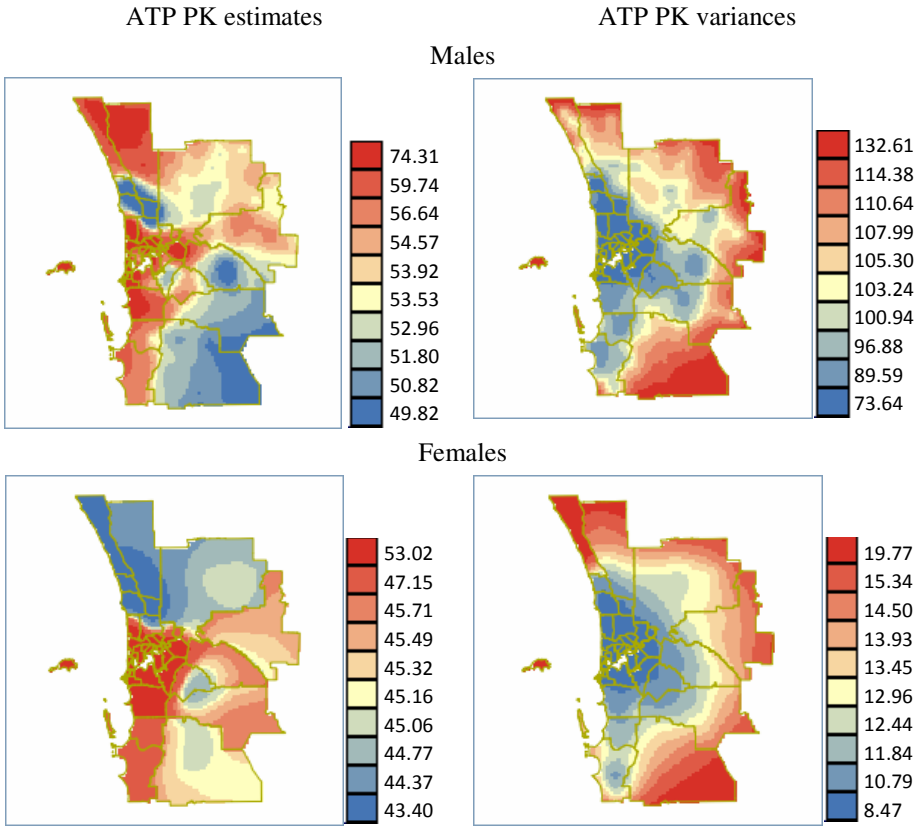


Figure 47 ATP Poisson kriging estimates and variances by sex based on colorectal cancer incidence age-adjusted rates.

The summary statistics for colorectal cancer incidence by sex are shown in Table 14. The mean value of male rates is slightly higher than that of females, but the difference in variance between males and females is significant. Generally speaking, male rates vary within a greater range than females and so have a higher variance.

Table 14 Summary statistics for colorectal cancer incidence crude rates, age-adjusted rates and ATP Poisson kriging estimated rates for males and females during the period 1990-2005.

Colorectal cancer incidence estimator during 1990-2005		Mean	Variance	Min	Max
Crude rates	M	63.59	516.58	18.33	129.61
	F	51.6	329.39	23.94	90.49
Age-adj rates	M	58.75	230.1	19.12	114.24
	F	45.97	47.39	28.44	60.38
ATP estimator	M	53.96	14.36	44.84	74.8
	F	45.18	3.5	39.34	53.05

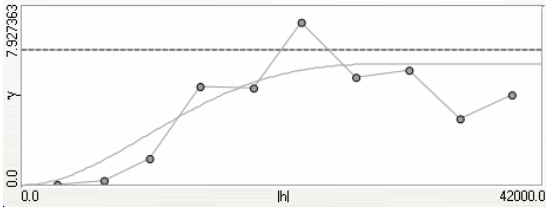
5.3.2 Colorectal cancer mortality analysis

5.3.2.1 Colorectal cancer mortality analysis in 1990-2005

The data to be analysed in this section are colorectal cancer mortality based on age-adjusted and age-sex-adjusted rates for the period 1990-2005.

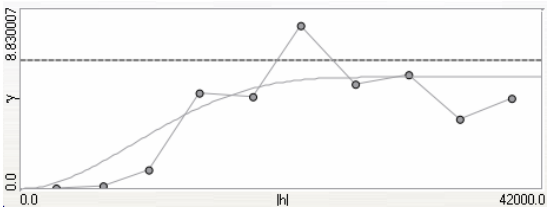
For semivariograms, the difference between age-adjusted rates and age-sex-adjusted rates is slight (Figure 48). The nugget effect is strictly zero for both adjusted rates. In addition, the effective spatial dependency distance for age-adjusted rates is 31km and 31.6km for the other. The sill for both rates is the same. Overall the spatial correlation is slightly stronger for age-sex-adjusted rates.

Age-adjusted mortality rates



Number of structures: 1
Nugget: 0
Cubic: Range 31, Sill: 5

Age-sex-adjusted mortality rates



Number of structures: 1
Nugget: 0
Cubic: Range 31.6, Sill: 5

Figure 48 Omnidirectional semivariograms of colorectal cancer mortality adjusted rates 1990-2005.

The estimates and the kriging variance for colorectal cancer mortality are indicated in Figure 49. The difference between two adjusted rates is slight. The map of estimates based on age-sex-adjusted rates is slightly smooth compared with age-adjusted rates. The map for the kriging variance is smooth for both adjusted rates.

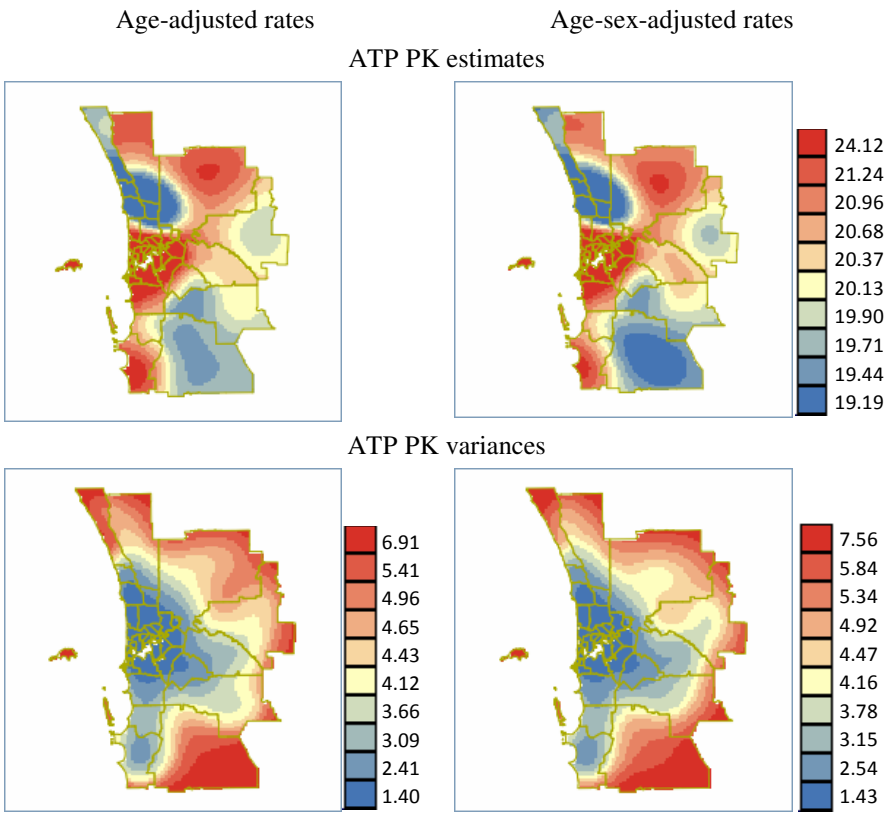


Figure 49 ATP Poisson kriging estimates in the same colour scale and variance based on colorectal cancer mortality adjusted rates.

The summary statistics for colorectal cancer mortality in 1990-2005 are shown in Table 15. The difference between adjusted rates is not large and so is the difference between ATP estimates based on both adjusted rates.

Table 15 Summary statistics for colorectal cancer mortality estimates during the period 1990-2005.

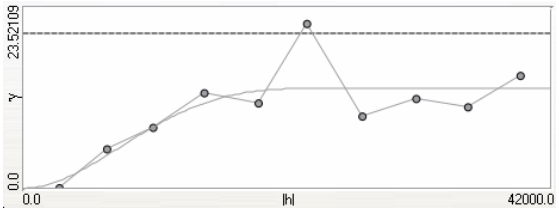
Estimator	Colorectal cancer mortality			
	Mean	Variance	Minimum	Maximum
Crude rates	23.77	71.24	10.36	37.83
Age-adjusted rates	21.38	12.61	13.3	29.03
Age-sex-adjusted rates	21.44	12.52	13.08	29.74
ATP (age) estimator	20.25	0.96	17.06	23.91
ATP (age-sex) estimator	20.18	1.1	17.1	24.12

5.3.2.2 Colorectal cancer mortality analysis for three periods: 1990-1995, 1996-2000 and 2001-2005

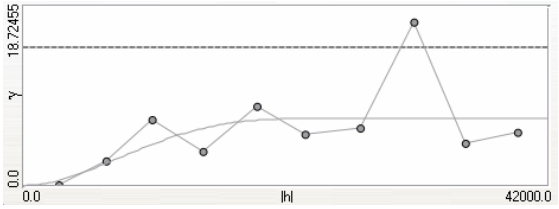
Here colorectal cancer mortality analysis during the period 1990-2005 is divided into three periods: 1990-1995, 1996-2000 and 2001-2005.

For semivariograms of colorectal cancer mortality age-adjusted rates, Figure 50 indicates that the difference between three periods is large. Nugget effects are zero in each period. Spatial correlation is relatively strong in the first two periods, and extends to 25km. In comparison, the effective distance for spatial dependence is about 17km in the third period. As for semivariogram values, they are nearly 13 in the first period but only 1 in the third period. Overall cancer mortality varies relatively continuously in the third period and shows more variability in the first period.

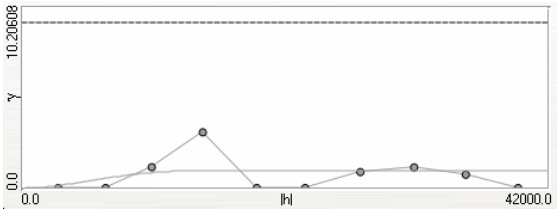
Age-adjusted mortality rates 1990-1995



Age-adjusted mortality rates 1996-2000



Age-adjusted mortality rates 2001-2005



Semivariogram model parameters for three models

Number of structures for three models: 1

90-95 Nugget: 0
Spherical: Range 25, Sill: 13

96-00: Nugget: 0
Exponential: Range 25, Sill: 7

01-05: Nugget: 0
Exponential: Range 17.4, Sill: 1

Figure 50 Omnidirectional semivariograms of colorectal cancer mortality age-adjusted rates.

The estimation and the kriging variance in three periods are indicated by Figure 51. For the estimation, the map is relatively smooth in the third period and shows great variability in other periods. The estimates are high in the inner suburbs of Perth in three periods. For the kriging variance, the values of variance are very high in the first period and low in the third

period, but the maps in three periods do not indicate great variability. In addition, the spatial trends for the variance maps are quite similar in three periods.

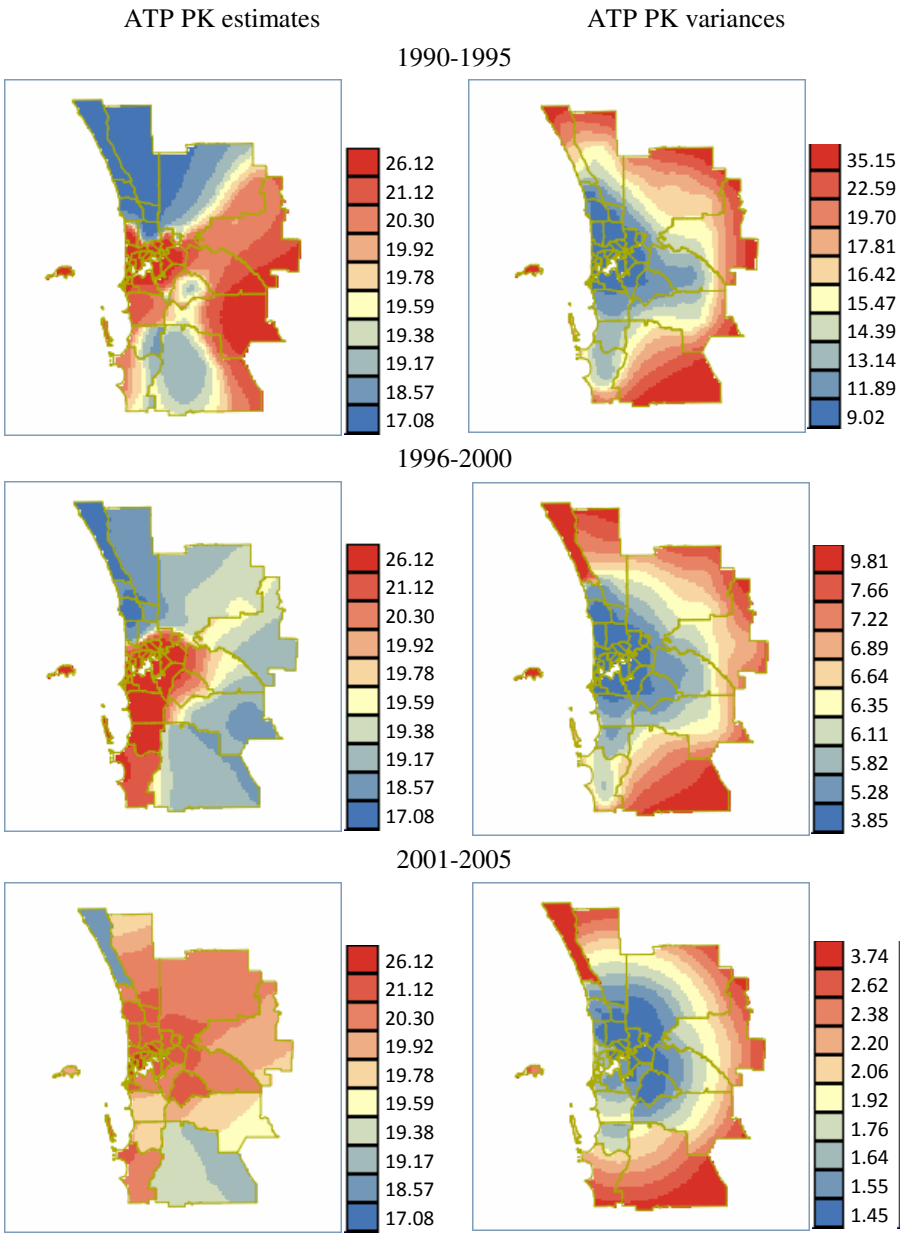


Figure 51 ATP estimated and variances based on colorectal cancer mortality age-adjusted rates.

The summary statistics for colorectal cancer mortality during three contiguous periods are shown in Table 16. The mean values of mortality rates decrease over time. The variance is lowest in the third period. Overall, colorectal cancer mortality rates do not increase with time in most areas during the recorded time. The rates in eastern part of the study area vary greatly with time. The rates have decreased sharply by the second period and then increased gradually by the third period in most areas except the southwest of the study area.

Table 16 Summary statistics for colorectal cancer mortality crude rates, age-adjusted rates and ATP Poisson kriging estimated rates during 1990-1995, 1996-2000 and 2001-2005.

Colorectal cancer mortality estimator during: 1990-1995, 1996-2000 and 2001-2005		Crude rates	Age-adjusted rates	ATP estimator
	1990-1995	24.91	20.62	19.03
Mean	1996-2000	22.52	19.42	19.32
	2001-2005	21.85	19.35	19.83
	1990-1995	128.09	33.95	7.34
Variance	1996-2000	120.81	46.2	1.93
	2001-2005	51.53	17.92	0.4
	1990-1995	7.92	9.2	11.87
Minimum	1996-2000	0	0	15.95
	2001-2005	8.76	8.03	17.5
	1990-1995	49.4	33.46	26.23
Maximum	1996-2000	45.82	37.23	24.92
	2001-2005	36.57	27.9	21.9

5.4 Breast cancer analysis

Breast cancer incidence data based on adjusted rates in different periods are analysed in the first section. In the second section, mortality data are considered. Breast cancer data in three contiguous periods: 1990-1995, 1996-2000 and 2001-2005 are also investigated in each section to get a temporal view.

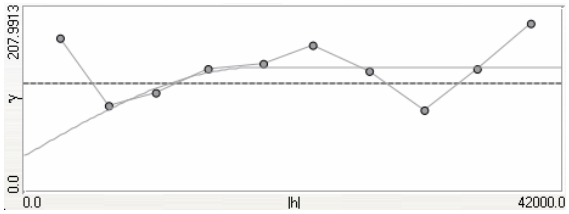
5.4.1 Breast cancer incidence analysis

Breast cancer incidence analysis during three contiguous periods and the entire study period will be carried out in this section.

A distinct difference exists between semivariograms for the three periods (see Figure 52). For the effective distance of spatial dependence, it is 35km in the first period and 10km in other periods. Semivariogram values are the highest in the second period with the value more than 400 and lowest in the third period with the value 110. The semivariogram at zero lag is 30 and 100 in the first and second period respectively, and 1 in the third period. As a result, there is great variability (higher semivariogram values) in breast cancer incidence during the second period. The best spatial continuity is shown in the variogram for the first period and spatial correlation is relatively strong.The second period shows almost pure

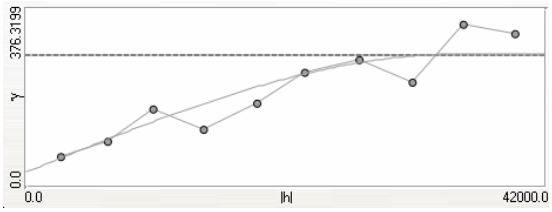
nugget. For the entire study period, the semivariogram values are relatively low and spatial variation ceases to be autocorrelated once case locations are farther apart than 18km. As a result, the spatial correlation for breast cancer incidence is relatively strong. The semivariograms at zero lag are 40.

Age-adjusted incidence rates 1990-2005



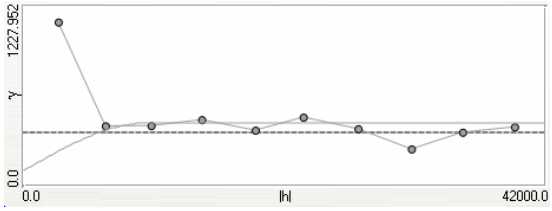
Number of structures: 1
Nugget: 40
Spherical: Range 18, Sill: 100

Age-adjusted incidence rates 1990-1995



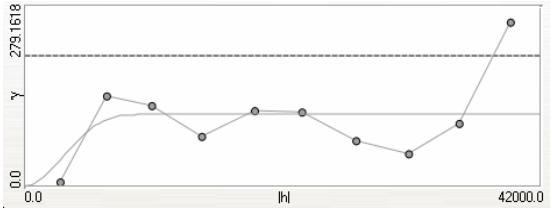
Number of structures: 1
Nugget: 30
Spherical: Range 35, Sill: 250

Age-adjusted incidence rates 1996-2000



Number of structures: 1
Nugget: 100
Spherical: Range 10, Sill: 328

Age-adjusted incidence rates 2001-2005



Number of structures: 1
Nugget: 1
Cubic: Range 10, Sill: 110

Figure 52 Omnidirectional semivariograms of breast cancer incidence age-adjusted rates.

ATP Poisson kriging estimates and the kriging variance for breast cancer incidence are shown in Figure 53. It can be seen clearly that the incidence estimates in 1990-2005 are relatively low across Perth SLAs. Higher incidence rates occur mainly in the middle of the study area. The estimation map of breast cancer incidence shows slight variability. The kriging variance in 1990-2005 is very low compared with other individual periods. For the three individual periods, breast cancer incidence goes up with time in the Perth metropolitan area. Estimation maps are very smooth during the first and third period compared with the map in the second period (higher variogram values and nugget effects). This is confirmed by the semivariogram plots in Figure 52. Low incidence rates are in the north and south, while high rates are in the older suburbs of Perth. When turning to the kriging variance, it is the greatest in the second period and lowest in the third period. The

variability in the map for kriging variance in the second period is highest while the maps for other periods are relatively smooth.

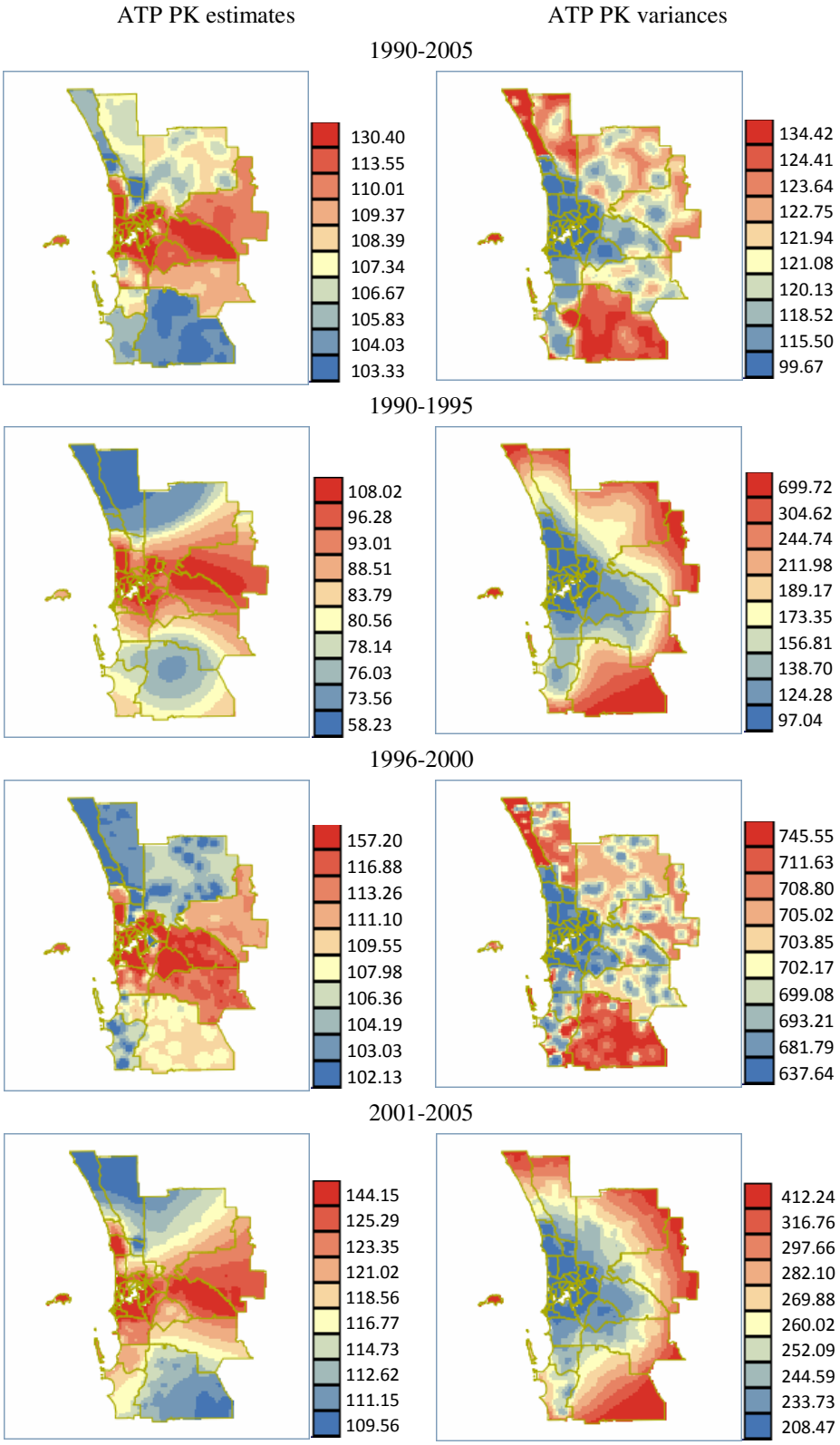


Figure 53 ATP estimated rates and variances based on breast cancer incidence age-adjusted rates.

The summary statistics for breast cancer incidence are shown in Table 17. Crude rates have lower minimum value than other rates in each period. Breast cancer incidence increased over time and varied greatly in the second period.

Table 17 Summary statistics for breast cancer incidence crude rates, age-adjusted rates and ATP Poisson kriging estimated rates during different periods.

Breast cancer incidence estimator during: 1990-1995, 1996-2000 and 2001-2005		Crude rates	Age-adjusted rates	ATP estimator
Mean	1990-2005	123.77	120.3	107.9
	1990-1995	102.58	91.20	79.90
	1996-2000	131.26	125.17	108.62
	2001-2005	130.63	123.03	117.11
Variance	1990-2005	1327.68	1343.89	17.98
	1990-1995	1026.04	460.52	207.94
	1996-2000	4035.77	4685.03	47.47
	2001-2005	1039.47	513.49	38.69
Minimum	1990-2005	72.66	77.43	96.06
	1990-1995	30.41	36.35	38.51
	1996-2000	66.53	82.59	85.55
	2001-2005	56.52	63.78	105.11
Maximum	1990-2005	262.1	314.55	130.4
	1990-1995	164.79	133.47	108.02
	1996-2000	451.61	508.38	157.20
	2001-2005	193.14	174.84	144.15

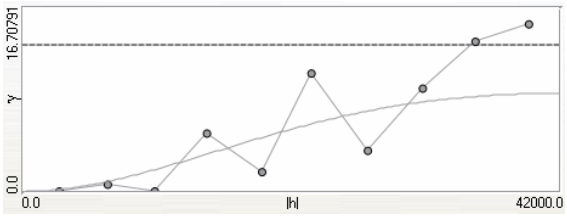
5.4.2 Breast cancer mortality analysis

Breast cancer mortality analysis during three contiguous periods and the entire study period will be carried out in this section.

For semivariogram plots of mortality, there is good structure in the experimental semivariogram for 1990-1995 and 1990-2005, and rather little structure in the other periods. The semivariogram values are relatively high in the third period (Figure 54). The effective distance for spatial dependence is 9km. The spatial correlation in the first period and the entire study period is relatively strong and extends to longer distance: 40km and 50km respectively. Although breast cancer mortality varies continuously (lower semivariogram values) in the second period, the range is relatively small and all of the variation ceases to

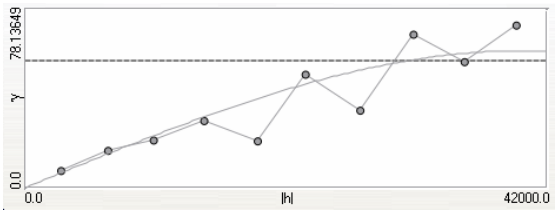
be autocorrelated after the distance 10km, and so the data exhibit a pure nugget structure. The semivariogram at zero lag is 0 in 1990-1995 and 1990-2005, and 100 in 2001-2005. As a result, the mortality risk from the age is possibly more continuous in the first period and the entire study period than other periods.

Age-adjusted mortality rates 1990-2005



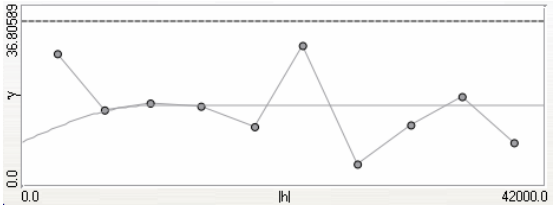
Number of structures: 1
Nugget: 0
Cubic: Range 50, Sill: 9

Age-adjusted mortality rates 1990-1995



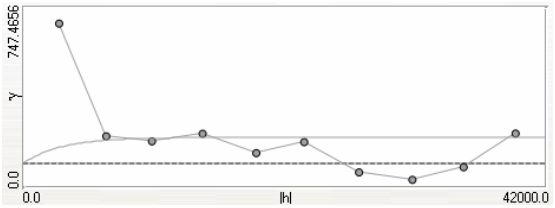
Number of structures: 1
Nugget: 0
Spherical: Range 40, Sill: 60

Age-adjusted mortality rates 1996-2000



Number of structures: 1
Nugget: 9
Spherical: Range 10, Sill: 7.4

Age-adjusted mortality rates 2001-2005



Number of structures: 1
Nugget: 100
Exponential: Range 9, Sill: 105

Figure 54 Omnidirectional semivariograms of breast cancer mortality age-adjusted rates.

Breast cancer mortality ATP Poisson kriging estimates and the kriging variance are shown in Figure 55. The mortality rate estimates in 1990-2005 are very high except northeast and southeast of the study area, and the estimation map is relatively smooth. For the kriging variance, the value is very low in 1990-2005 and the map is relatively smooth. For the three individual periods, the estimated mortality rates in the first period in areas around the Perth city centre are quite a contrast to those in other areas where lower rates are found. During the second period, there was little variation in mortality rates across Perth SLAs. However, mortality rates vary considerably in the third period. There is marked difference in the kriging variance between maps of three periods. The kriging variance in the third period is extremely high and in the second period is extremely low. ATP Poisson kriging variance based on age-adjusted rates illustrates more variability in the last two periods than in the first period.

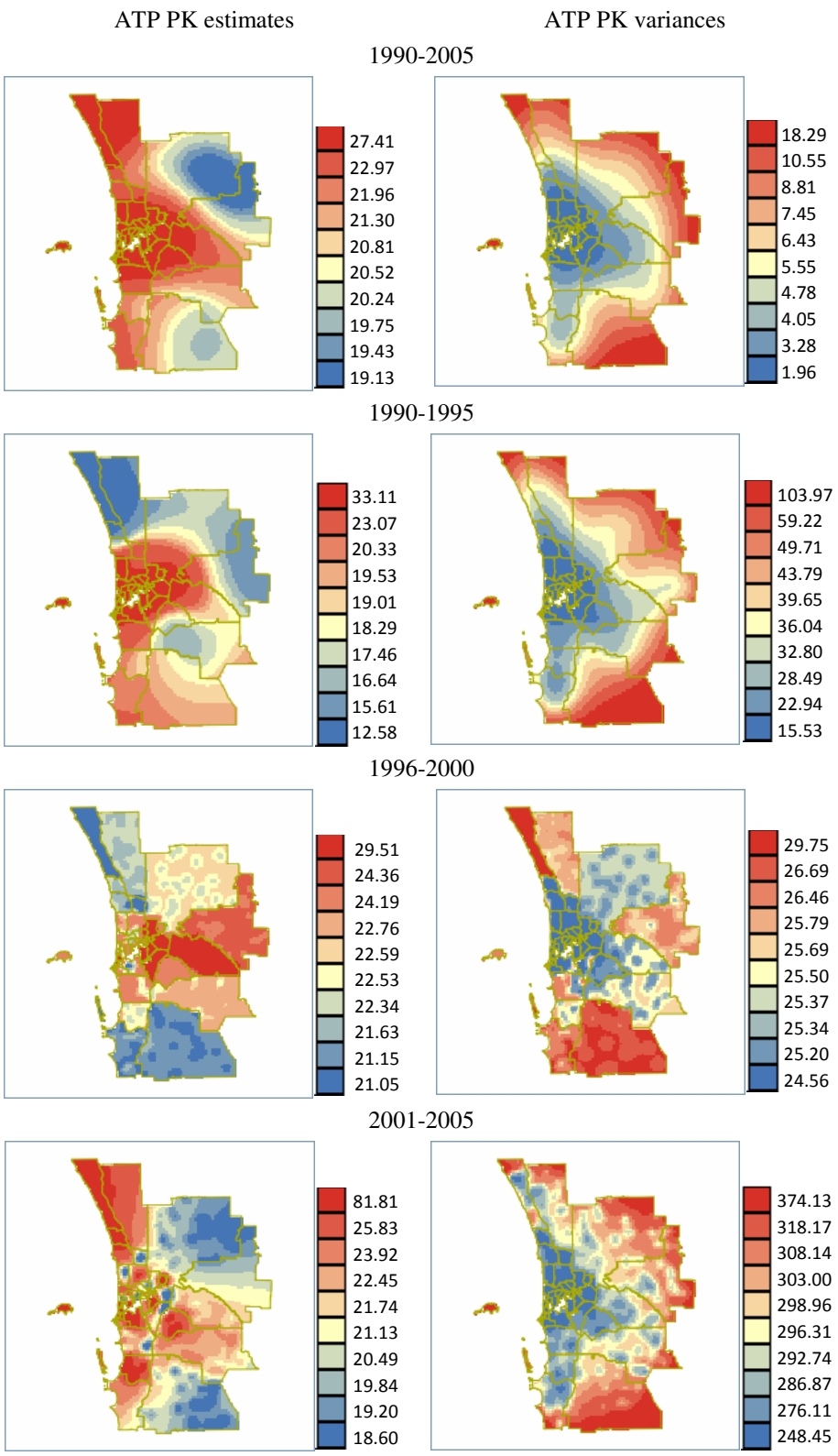


Figure 55 ATP estimated rates and variances based on breast cancer mortality age-adjusted rates.

The summary statistics for breast cancer mortality are shown in Table 18. The mean value of crude rates is higher than other rates in each period. Age-adjusted rates have lower

variance values than crude rates. In addition, the mean value is the lowest in the second period and highest in the third period for crude and adjusted rates. Breast cancer mortality varied over time and increased by the third period when the mortality rates showed the greatest variability. The variance is very low in 1990-2005 compared with other periods.

Table 18 Summary statistics for breast cancer mortality crude rates, age-adjusted rates and ATP Poisson kriging estimated rates during different periods.

Breast cancer mortality estimator during: 1990-1995, 1996-2000 and 2001-2005		Crude rates	Age-adjusted rates	ATP estimator
Mean	1990-2005	27.16	24.62	21.68
	1990-1995	30.59	26.64	18.15
	1996-2000	23.99	20.71	22.52
	2001-2005	32.71	30.85	21.69
Variance	1990-2005	88.98	26.71	4.37
	1990-1995	223.26	206.29	18.92
	1996-2000	187.17	92.81	2.04
	2001-2005	1035.62	1160.18	11.7
Minimum	1990-2005	11.32	11.83	18.27
	1990-1995	3.97	4.03	6.78
	1996-2000	0	0	19.38
	2001-2005	0	0	15.29
Maximum	1990-2005	48.48	36.76	27.41
	1990-1995	65.88	93.6	33.11
	1996-2000	64.52	50.75	29.51
	2001-2005	190.27	206.72	81.81

5.5 Prostate cancer analysis

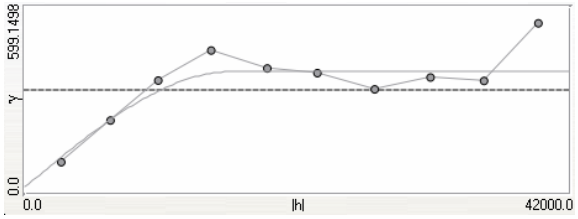
Prostate cancer incidence data based on adjusted rates in different periods are analysed in the first section. In the second section, mortality data are considered. Prostate cancer data in three contiguous periods: 1990-1995, 1996-2000 and 2001-2005 are also investigated in each section to get a temporal view.

5.5.1 Prostate cancer incidence analysis

Prostate cancer incidence analysis during three contiguous periods and the entire study period will be carried out in this section. In addition, we will see how the incidence varies over time.

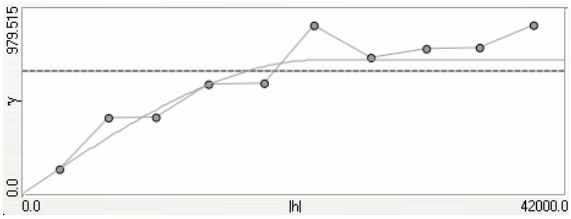
There is quite good continuity except the second period for semivariogram plots of prostate incidence (Figure 56). In fact, the time series graph shown in Figure 2 in Chapter 4 also indicates that there are large fluctuations in the number of prostate cancer incidence cases in the second period. The semivariogram values are very high in the first period but the spatial correlation is quite strong. The spatial variation is autocorrelated when the distance between locations is within 23km. In comparison, semivariogram values are very low and the effective distance for spatial dependence is about 15km in the second period. In the third period, the semivariogram values are relatively high. Locations farther apart than 20km are not spatially autocorrelated. The semivariogram at zero lag is not 0 in any period: it is lowest in the first period with value 10 and highest in the second period with value 40. Thus prostate cancer incidence rates vary more continuously but spatial correlation is weak in the second period. The short range variability is probably due to the fewer incidence cases available in the second period. Prostate cancer incidence risk from age is possibly more continuous in the first period than the risk in other periods.

Age-adjusted incidence rates 1990-2005



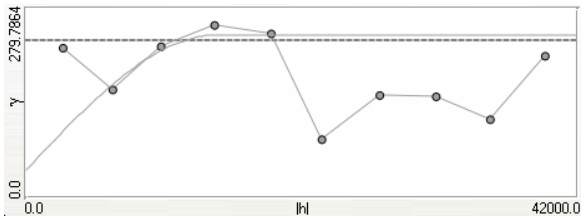
Number of structures: 1
Nugget: 20
Spherical: Range 16, Sill: 370

Age-adjusted incidence rates 1990-1995



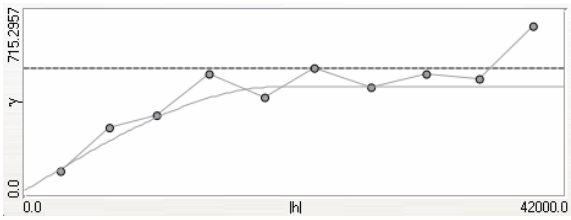
Number of structures: 1
Nugget: 10
Spherical: Range 23, Sill: 700

Age-adjusted incidence rates 1996-2000



Number of structures: 1
Nugget: 40
Spherical: Range 14.5, Sill: 200

Age-adjusted incidence rates 2001-2005



Number of structures: 1
Nugget: 20
Spherical : Range 20, Sill: 400

Figure 56 Omnidirectional semivariograms of prostate cancer incidence age-adjusted rates.

The estimates and variance of prostate cancer incidence are indicated in Figure 57. Higher prostate cancer rates are mainly in northwest of the study area in 1990-2005 and the estimation map is relatively smooth. For the estimates in three individual periods, incidence rates vary more continuously in the first and third periods. During the first period, incidence rates in SLAs near the Perth city centre are much higher than those in other areas. The incidence rates decreased by the second period but there was little variation in incidence rates map across Perth. However, incidence rates increased considerably by the third period. There is a distinct difference in the kriging variance between the maps for the different periods. The kriging variance in the entire period is relatively low. During the first period, the kriging variance is very high and shows slight variability. The kriging variance in the second period is very low and the map indicates more variability.

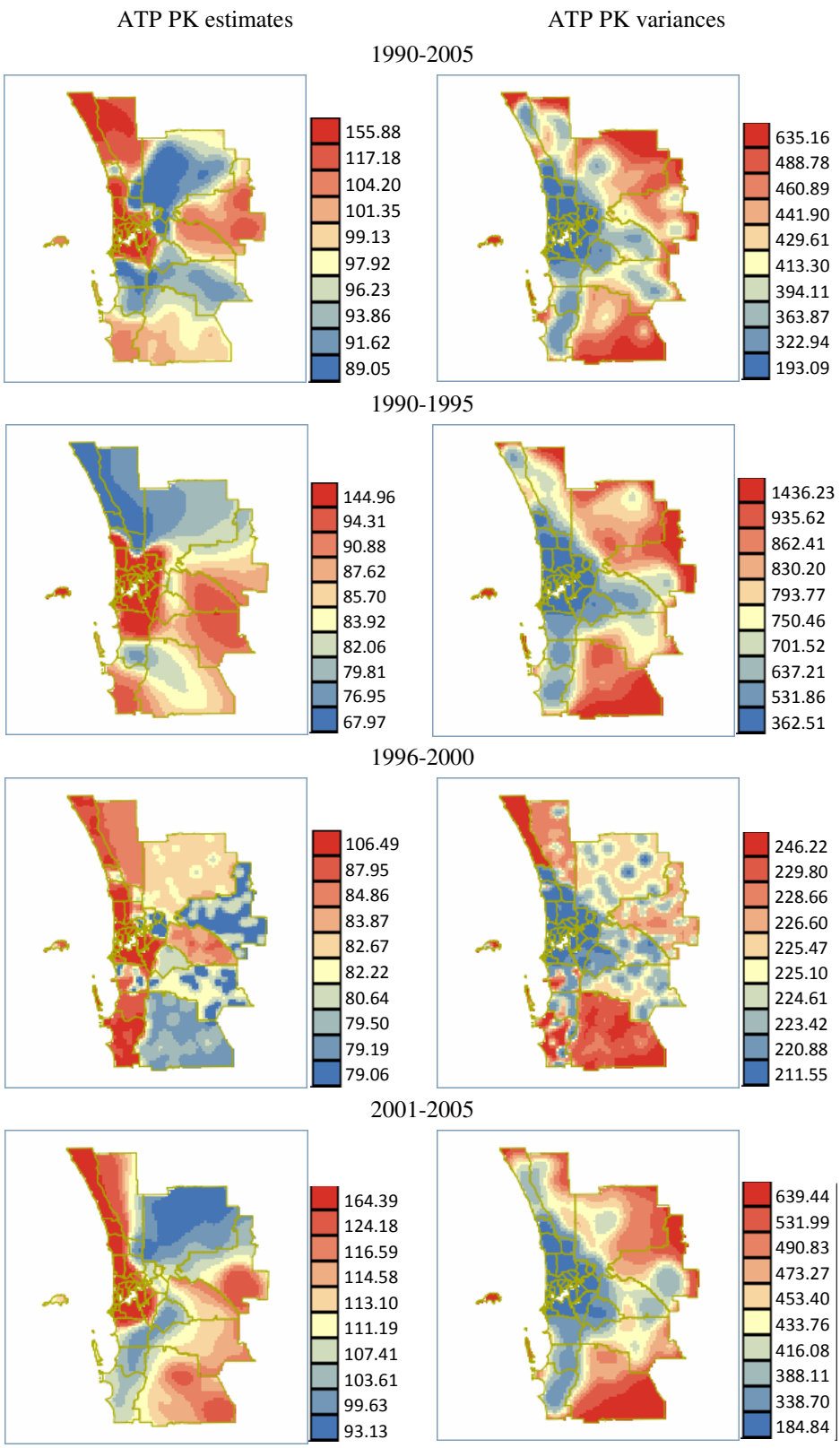


Figure 57 ATP estimated rates and variances based on prostate cancer incidence age-adjusted rates.

The summary statistics for prostate cancer incidence are shown in Table 19. The mean value, maximum value and variance of crude rates are relatively high compared with other rates in different periods. In addition, the mean value and maximum value are lowest in the second period and highest in the third period. The variance is the highest in the first period and relatively low in the second period.

Table 19 Summary statistics for prostate cancer incidence crude rates, age-adjusted rates and ATP Poisson kriging estimated rates during different periods.

Prostate cancer incidence estimator during: 1990-1995, 1996-2000 and 2001-2005		Crude rates	Age-adjusted rates	ATP estimator
Mean	1990-2005	121.68	113.08	99.68
	1990-1995	124.76	102.06	84.5
	1996-2000	92.46	83.66	82.67
	2001-2005	137.72	125.41	109.74
Variance	1990-2005	1513.68	495.12	125.66
	1990-1995	3380.52	1066.66	213.45
	1996-2000	1177.91	647.85	17.14
	2001-2005	2200.16	717.54	156
Minimum	1990-2005	59.13	81.52	77.18
	1990-1995	27.82	27.22	47.58
	1996-2000	0	0	72.43
	2001-2005	57.8	60.66	84.88
Maximum	1990-2005	198.47	163.42	155.88
	1990-1995	235.38	158.19	145.12
	1996-2000	157.29	141.12	106.49
	2001-2005	244.23	188.06	164.39

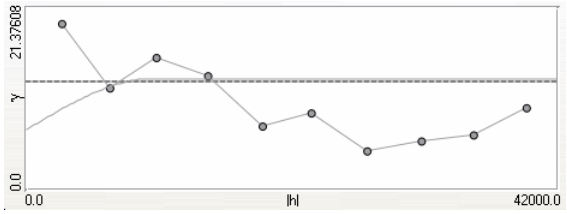
5.5.2 Prostate cancer mortality analysis

This section analyses prostate cancer morality and we also study how the prostate cancer mortality varies over time.

The experimental semivariogram of prostate cancer mortality show differences among different periods (Figure 58). The semivariograms in the second period and the third period are not reaching the sample variance (dashed line). The semivariogram values are very low compared with those of prostate cancer incidence displayed in Figure 56. The effective distance for spatial dependence is relatively long in the second period. The semivariogram values are the lowest and nugget effect is strictly zero in the third period. In comparison,

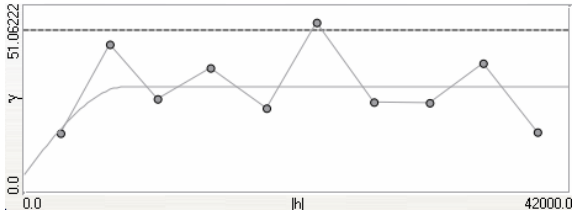
semivariogram values are the highest in the first period. Extremely poor structures exhibited by the semivariograms are partly caused by the fluctuations of low mortality cases during the study period (Figure 2).

Age-adjusted mortality rates 1990-2005



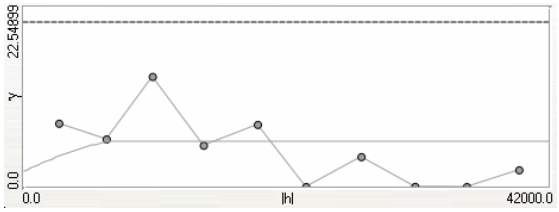
Number of structures: 1
Nugget: 7
Spherical: Range 10, Sill: 6

Age-adjusted mortality rates 1990-1995



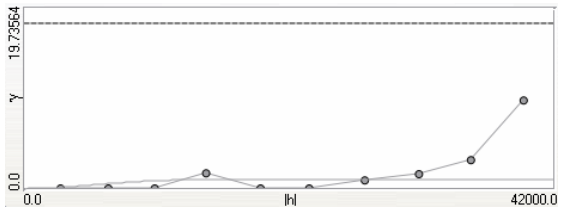
Number of structures: 1
Nugget: 5
Spherical: Range 7.8, Sill: 24

Age-adjusted mortality rates 1996-2000



Number of structures: 1
Nugget: 2
Spherical: Range 8, Sill: 3.8

Age-adjusted mortality rates 2001-2005



Number of structures: 1
Nugget: 0
Cubic: Range 18, Sill: 1

Figure 58 Omnidirectional semivariograms of prostate cancer mortality age-adjusted rates.

ATP Poisson kriging estimates and the kriging variance for prostate cancer mortality are shown in Figure 59. It can be seen that the rates in 1990-2005 are relatively low in most areas except the middle part of the study region. The mortality rates show the most variability in the first period: the rates are very high in south and relatively low in north. There is slight variability in the estimation map in the second period and the estimates are low in most areas. The mortality rates vary more continuously and the rates are high in most areas by the third period. When referring back to the time series plots for prostate mortality (Figure 2), this feature is visible across Perth: the number of mortality cases is relatively high during the first six years and goes down greatly in the next five years and then increases slowly in the last period. It needs to be noted that the spatial trends for the prostate mortality estimated rates in the third period are slightly different from those of the crude and age-adjusted rates shown in Figure 104. One reason is that the fluctuations of low

number of mortality cases during this period (Figure 2). The other reason is that kriging technique always overestimates low values and underestimates high values.

For the kriging variance, there is great difference between maps in different periods. Spatial trends in prediction maps have been confirmed by the variogram plots (Figure 58). The kriging variance is extremely high in the first period, and low in other periods. Prediction maps indicate more variability in the first two periods and the entire period. The map varies more continuously in the last period. In contrast, the kriging variance of prostate cancer mortality rates is much lower than that of incidence rates (see Figure 57) in the corresponding period.

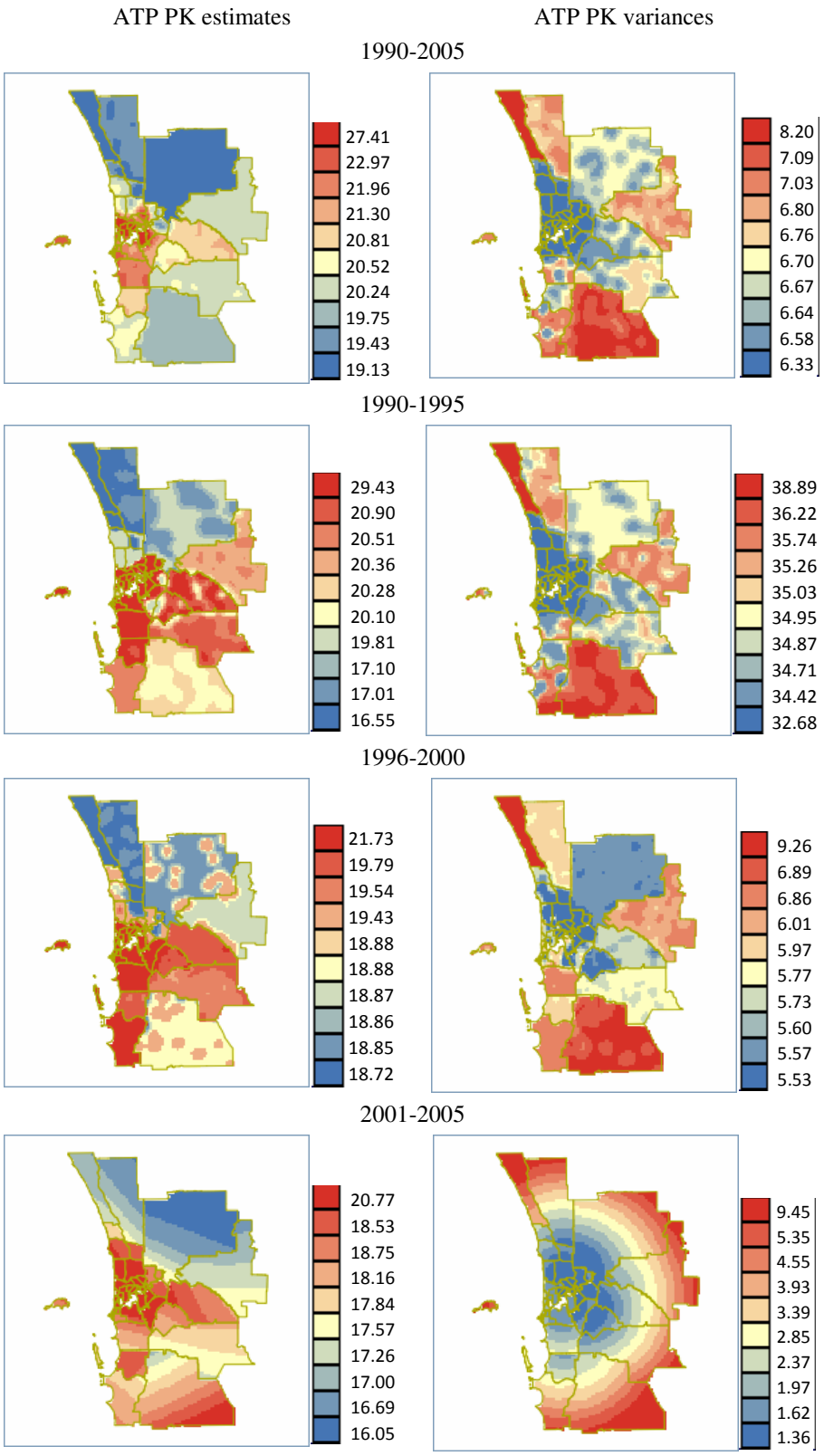


Figure 59 ATP estimated rates and variances based on prostate cancer mortality age-adjusted rates.

The summary statistics for prostate cancer mortality are shown in Table 20. Crude rates have relatively large range (low minimum and high maximum) and high variance value. The mean value of crude rates is higher than other rates in each period. In addition, the mean value and maximum value decreased with time except adjusted rates. In short, prostate cancer mortality rates decreased over time and the variance decreased accordingly.

Table 20 Summary statistics for prostate cancer mortality crude rates, age-adjusted rates and ATP Poisson kriging estimated rates during different periods.

Prostate cancer mortality estimator during: 1990-1995, 1996-2000 and 2001-2005		Crude rates	Age-adjusted rates	ATP estimator
Mean	1990-2005	21.98	19.61	19.87
	1990-1995	23.18	18.22	19.19
	1996-2000	21.37	18.62	19.18
	2001-2005	20.35	17.5	17.49
Variance	1990-2005	127.51	42.42	0.98
	1990-1995	265.72	87.52	7.02
	1996-2000	137.03	71.61	0.52
	2001-2005	112.02	45.18	1.02
Minimum	1990-2005	0	0	17.65
	1990-1995	0	0	10.04
	1996-2000	0	0	17.42
	2001-2005	0	0	15.00
Maximum	1990-2005	45.35	35.63	25.76
	1990-1995	61.11	40.17	29.43
	1996-2000	51.05	40.52	21.73
	2001-2005	47.76	27.54	20.77

5.6 Summay

This chapter has provided the geostatistical analysis for cancers recorded in 1990-2005 in metropolitan Perth.

For lung cancer incidence in 1990-2005, the difference between age-adjusted rates and age-sex-adjusted rates is not significant. The incidence risk from age is possibly more continuous than the risk from age-sex. The estimated incidence rates for males in this period are much higher than for females in most SLAs of Perth, and the kriging variance is nearly ten times that of females. Also, female incidence varies more continuously and the spatial correlation is much stronger than males across Perth SLAs. For melanoma cancer

incidence in 1990-2005, the rates are relatively high in SLAs near the west coast of Perth. The spatial correlation is relatively strong for age-sex-adjusted rates compared with age-adjusted rates across Perth SLAs. Males have higher incidence rates than females in Perth and female incidence rates vary more continuously than males. The highest incidence estimated rate is 108 per 100,000 person-years for males while it is only 57 for females. For colorectal incidence in 1990-2005, the cancer risk from age-sex is possibly more continuous than the risk only from age and so the variance map is relatively smooth. Males have higher incidence rates than females in Perth, and the incidence shows great variability compared with females.

For lung cancer mortality rates in 1990-2005, the mortality risk from age is possibly more continuous than the risk from age-sex. For mortality rates in the three contiguous time periods, there is an overall increase in most SLAs of Perth over time. However, the mortality rates are very high in the middle of the study area in the first period. For melanoma cancer mortality in 1990-2005, the difference between age-adjusted rates and age-sex-adjusted rates is slight. The kriging variance for melanoma mortality based on adjusted rates is the lowest compared with other estimates. The mortality risk from age-sex is possibly more continuous than the risk only from age. Melanoma mortality rates are relatively low in the first period and the rates vary more continuously than other periods. The highest melanoma mortality rates occur in the second period and the rates show more variability in this period. For colorectal cancer mortality, the difference between age-adjusted rates and age-sex-adjusted rates is slight. The mortality rates in eastern part of the study area vary greatly with time. The rates have decreased sharply by the second period and then increased gradually by the third period.

As for sex specific cancers, breast cancer incidence rates are relatively low in 1990-1995 and mortality rates are relatively low in 1996-2000. Cancer incidence rates increased gradually with time in most SLAs of Perth. Breast cancer mortality rates have decreased by the second period and then have increased a little by the third period. Spatial dependency for breast cancer is relatively strong in 1990-1995. Breast cancer incidence risk from age is relatively continuous in the third period, while breast cancer mortality risk from age is possibly more continuous in the first period.

For prostate cancer analysis during three periods, prostate cancer incidence rates are relatively low in 1996-2000 but have increased by the third period. Prostate mortality rates decrease with time in most Perth SLAs. The population at risk has had a slow increase by the third period for both incidence and mortality rates. Spatial correlation is relatively strong in the third period and very weak in the second period for prostate cancer across Perth SLAs. The mortality risk from the age is possibly more continuous than the incidence risk from the age.

For the analysis of breast and prostate cancer in the period 1990-2005, breast cancer rates are slightly higher than prostate cancer rates. The same trends are found for mortality rates in three contiguous periods: 1990-1995, 1996-2000 and 2001-2005. However, for incidence rates, breast cancer rates are only higher than prostate cancer rates in the second period.

In brief, the difference in lung cancer incidence risk between males and females is significant. Estimation and variance maps show more variability for males than for females. Age-sex-adjusted rates have stronger spatial correlation than age-adjusted rates across Perth SLAs. For melanoma cancer, males have higher cancer incidence risk than females and the kriging variance for male estimates is relatively high compared with females. The difference in colorectal cancer incidence risk between males and females is not significant. For cancer mortality kriging estimates in three periods, they are generally high in the inner suburbs of Perth and increase gradually with time in the north of the study area. Lung cancer mortality rates have increased gradually across Perth SLAs with time. In the west of the study area which is close to the coast, melanoma cancer mortality rates have decreased by the second period but increased greatly by the third period. It seems that there are fluctuations in the east of the study area during three periods and the mortality rates have decreased by the third period. Colorectal cancer mortality rates have increased with time in the north of the study area. In comparison, melanoma mortality rates are the lowest and lung cancer mortality rates are relatively high. The difference between breast and prostate cancer in 1990-2005 is slight. The spatial correlation for breast cancer is relatively strong and cancer risk varies continuously compared with prostate cancer across Perth SLAs.

This chapter has provided an analysis of cancer incidence and mortality at the (relatively) large SLA scale. In the next chapter a finer scale will be used. Spatial point pattern analysis

will be used to explore the intensity of the point patterns for different cancers. Going back to the literature review, Turner (2009) illustrated the techniques used for spatial point pattern analysis but no actual conclusions based on the modelling were drawn. To the best of our knowledge, our study provides the first detailed application of point process modeling to cancer data and the results are interpreted to draw conclusions on the spatial intensity of the point patterns.

6 SPATIAL POINT PATTERN ANALYSIS

Spatial point pattern theory focuses on the cancer analysis at finer scales. To investigate model fitting, we explore the point pattern of 2005. There are two reasons for doing this: firstly, we have more choice for the spatially varying susceptible population density compared with point patterns in other years. Secondly, the population data by suburb available in this study are those for the year 2006.

The intensity of the point process for five cancer types will be investigated in this chapter. Perth Central Business District (CBD) is chosen as the reference point in spatial statistical modelling. In addition, for lung cancer incidence the Kwinana industrial area is also used as reference point and for melanoma cancer incidence the coastline is used as reference line as well. To find an appropriate inhomogeneous Poisson process model, potential covariates will be incorporated to determine how they account for cancer point process. We will use lung cancer mortality data to explore whether different point patterns have an influence on spatial point pattern analysis and to decide what method is appropriate for treating missing values. We also use prostate mortality data as an example to illustrate the process to select an appropriate model based on the available covariates and that process illustrated will be applied to the other cancers.

The quadrat counting test (Table 21) indicates that the incidence and mortality point processes for the five cancers in 2005 are not completely random. In the further discussion, we will use this conclusion for model fitting. See Appendix 9.2 for the R codes used in spatial point pattern analysis.

Table 21 χ^2 test for testing the hypothesis of CSR for cancers in 2005.

Cancer types		χ^2	Degree of freedom	p-value
Lung cancer	Incidence	1827.78	31	Less than 2.2e-16
	Mortality	1279.91	31	Less than 2.2e-16
Melanoma cancer	Incidence	1737.94	31	Less than 2.2e-16
	Mortality	223.85	31	Less than 2.2e-16
Breast cancer	Incidence	2902.62	31	Less than 2.2e-16
	Mortality	424.99	31	Less than 2.2e-16
Prostate cancer	Incidence	3007.12	31	Less than 2.2e-16
	Mortality	315.27	31	Less than 2.2e-16
Colorectal cancer	Incidence	2248.93	31	Less than 2.2e-16
	Mortality	593.43	31	Less than 2.2e-16

6.1 Lung cancer analysis

Lung cancer data will be studied in this section using spatial point pattern analysis. The available covariates used in lung cancer model fitting are the population density (`Popdens`), the percentage of people aged 50 or over (`aged`) and the susceptible population intensity (Figure 22).

6.1.1 Exploratory data analysis

In this section, two methods will be used to treat case locations without recorded coordinates using lung cancer mortality as an example: method 1 is to remove the case locations without coordinates, and the original point pattern treated using this method will be called point pattern in the discussion of lung cancer analysis. Method 2 is to assign the mean location for the 16-year period in the SLA where they occur to them, and the point pattern treated using this method will be called modified point pattern in the discussion.

The lung cancer incidence point process in 2005 (see Figure 60) is inhomogeneous. The rejection based on quadrat counting test can be explained partly by the population density in the Perth metropolitan area (see Figure 19). Under the null hypothesis of inhomogeneous Poisson process, the seventeen duplicate case locations (about 2.6%) were removed. There are two reasons for doing this. One is the requirement of the characteristics of the

homogeneous Poisson process (Senn and Scott 2008 Ch2), in which two points never coincide. The other is to avoid the "nugget" effect at zero distance when the inhomogeneous *K*-function is estimated in the further study (Turner 2009).

In order to take account of the inhomogeneity of the year 2005 data, the inhomogeneous *K*-function is applied to test data independence. The selection of `sigma` value is generally based on the data point patterns and the window size of the study region. In Figure 61, the "smoothing parameter" `sigma` is set to 7 for incidence. For mortality, the kernel smoothing bandwidth is set to `sigma=4.9`. The point process is clustered when the inhomogeneous *K*-function estimate stays above the high bound of the critical envelope and inhibited when the estimate stays below the low bound of the critical envelope. According to Figure 61, there is no indication of attraction or inhibition for the year 2005 incidence data point process (left plot). Although the estimate stays very close to the lower bound of the critical envelope when the r values are in the range 0 to 5, it can be seen clearly that the point process is independent. As for the mortality point process, the estimated *K* values are slightly lower than the lower bound of the critical envelope when the r values are in the range 7 to 14. However, it is too early to conclude that there is repulsion between mortality points within this range at this stage. Although there is evidence against the process being an inhomogeneous Poisson process, the mortality cases are more dispersed than the incidence cases from Figure 60. There are almost no mortality cases in the eastern study region. This could account for the "inhibition" indicated by the *K*-function.

The point process model to be fitted might include an inhomogeneous spatial trend, dependence on covariates, and interpoint interactions (Baddeley and Turner 2005). Based on the characteristics of an inhomogeneous Poisson process, the year 2005 data will be modelled to see whether the covariates influence the intensity of the point process and how.

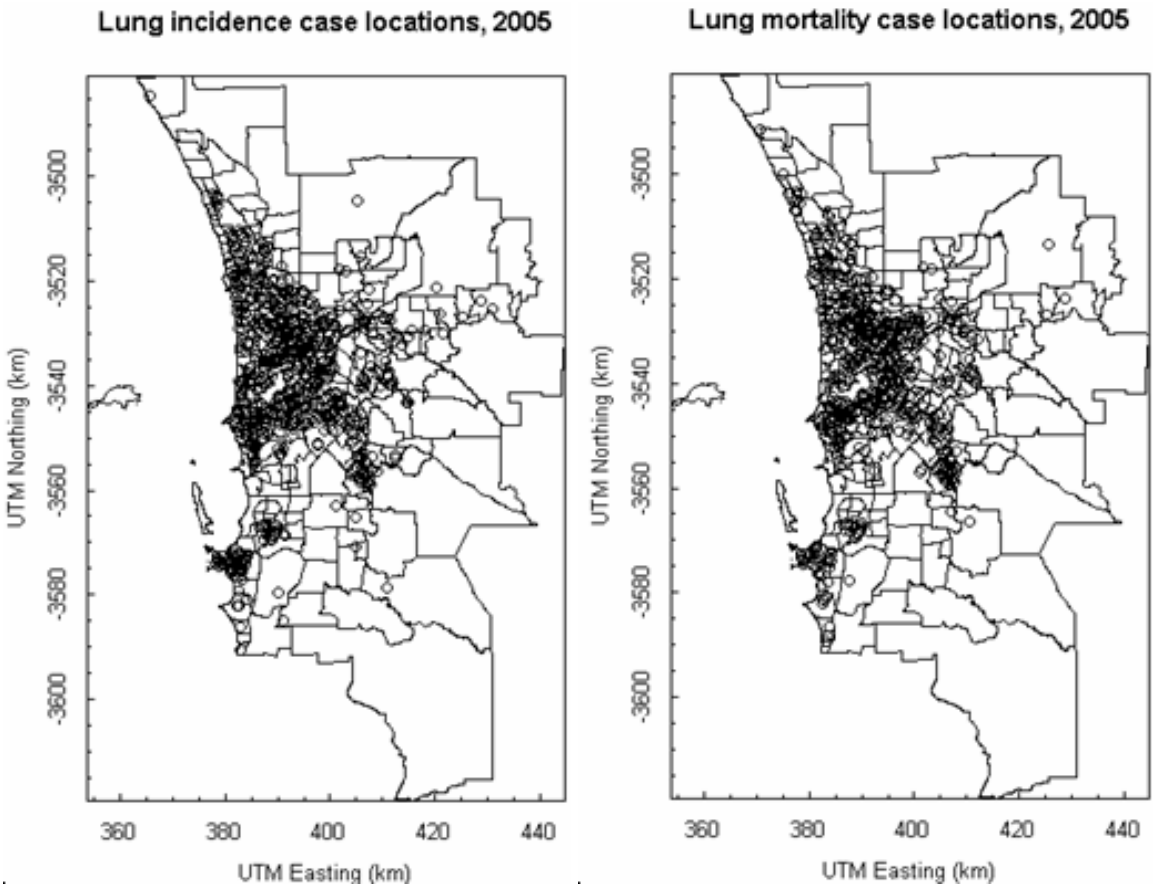
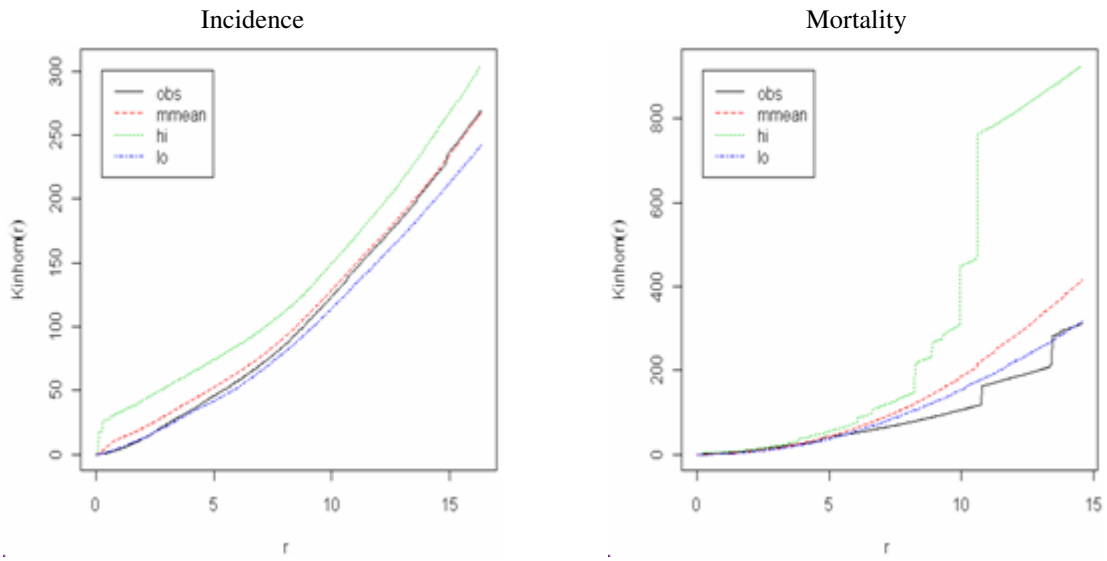


Figure 60 The modified lung cancer incidence (left) and mortality (right) case locations in year 2005 in Perth.



Bandwith for smoothing intensity : 7 Bandwith for smoothing intensity: 4.9

Bandwith for simulation from data: 7 Bandwith for simulation from data: 4.9

Figure 61 Estimates of the inhomogeneous K-function for the year 2005 lung cancer incidence and mortality data.

6.1.2 Analysis of modified lung cancer incidence point patterns

There are 665 lung cancer incidence cases with 17 cases without coordinates in the original 2005 lung cancer data. In this section, the data to be analysed are treated using method 2 and the spatial intensity of the susceptible population is based on Figure 22. The Kwinana Industrial Area is located in the southwest of the Perth metropolitan area (Figure 6). According to kriging analysis (Figure 29 and Figure 31), lung cancer incidence risk is relatively high in Kwinana compared to other SLAs. It has been found that air pollution increases the risk of lung cancer according to Cancer Research UK (CRUK). Diggle (1990) developed a methodology for fitting a class of inhomogeneous Poisson point process models to investigate whether there is evidence of raised cancer incidence near the industrial incinerator. Similarly, the Kwinana Industrial Area is used as reference point for modelling incidence data to see whether there is underlying relationship between the existence of industrial area and lung cancer incidence.

Inspecting the data (see Figure 60), it is apparent that there are more cancer cases in the inner suburbs of Perth due to higher population density. Since the Perth CBD (391.063, -3536.509) is roughly in the centre of cancer case locations, it is chosen as the reference point in spatial statistical modelling. The spatial trend model only includes a transformation of the coordinates of the case locations. A quantile-quantile plot for this model is shown in left plot of Figure 62. This plot indicates that this model is not suitable for describing the data. The plot compares empirical quantiles of the smoothed residual field of data with those of the fitted model (estimated by Monte Carlo sampling).The mean quantile of the simulated values is not zero, and the range of simulated values is much smaller than the range of data values. Inclusion of additional covariates, such as `popdens` or `aged`, does not improve the fit. One example shown here is the quantile-quantile plot for the model including the spatial trend and dependence on covariate `popdens` (see right plot of Figure 62).

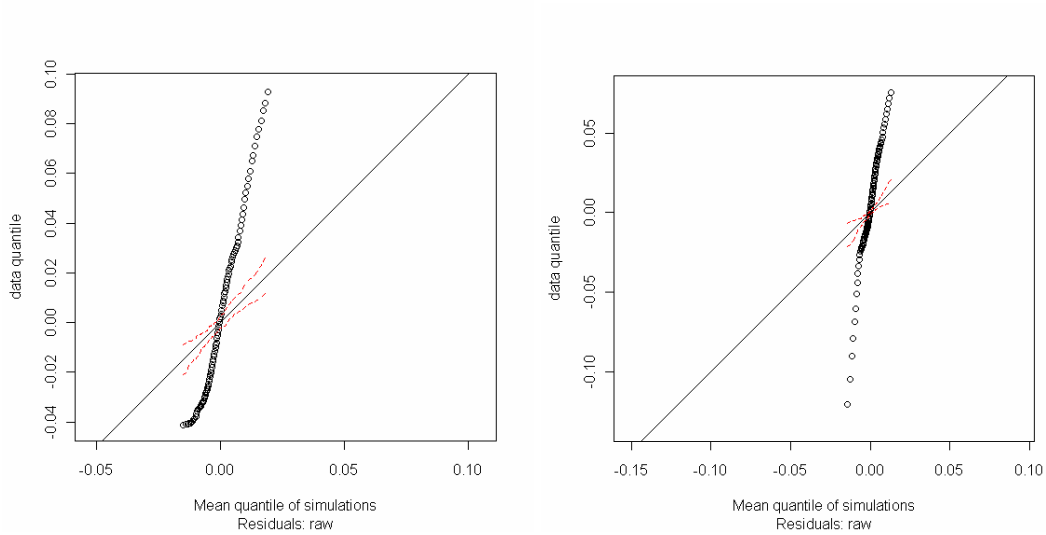


Figure 62 The quantile-quantile plots for the inhomogeneous Poisson model fitted to the year 2005 incidence data: trend only model (left) and the model including spatial trend and covariate `popdens` (right).

According to Figure 60, there is a very high intensity of cases in the middle of the western part of the study area where the Central Business district is, so it is important to take into account the spatial variation in vulnerable population density.

The empty space distances of cancer incidence case locations reflect how dense the susceptible population is in Perth. Among the empty space distances of lung cancer incidence during the period 1990-2004, the fit is adequate when only those of year 1993 are used for the model fitting data 2005. This probably due to the similarity of the point patterns in these two years. The quantile-quantile plot for the model is shown in left plot of Figure 63. The fit has improved greatly. The lower end of the plot stays close to the theoretical ideal line. However, the middle of the plot touches the lower edge of the critical envelope and the upper end of the plot protrudes the upper edge of the critical envelope. The smoothed residual field for the data has lighter right-hand tail and less variability than the smoothed residual field for simulation from the fitted model. That is, the pattern is more inhibited than the model. To improve the fit, `popdens` (population density) is added as covariate. Converting back to the form (2.39), the revised fitted model is

$$\lambda_\theta(\mathbf{u}) = ((x-391.063)^2 + (y+3536.509)^2)^{0.03} \times \exp(-0.708 - 0.987\, Z_1(\mathbf{u}) + 0.001 Z_2(\mathbf{u})) \quad (6.1)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 1993 incidence point patterns and $Z_2(\mathbf{u})$ represents population density. The quantile plot based on model (6.1) indicates that the fit is adequate (see middle plot of Figure 63). The middle of this plot just sticks to the theoretical ideal line. This model is acceptable since the plot stays well within the critical band. This plot confirms that the process is inhomogeneous and that the model (6.1) has a correct trend and covariate effects.

When the proportion of people aged 50 or above (`aged`) is included as an additional covariate instead of the overall population density, the fit (see right plot of Figure 63) is worse: the residual field for the data has a lighter right-hand tail and higher variability than the residual field for simulation from the fitted model. It indicates that this covariate is not a suitable explanatory variable for lung cancer incidence.

Distmap + spatial trends

Distmap+ spatial trends **+ popdens**

Distmap+spatial trends **+ aged**

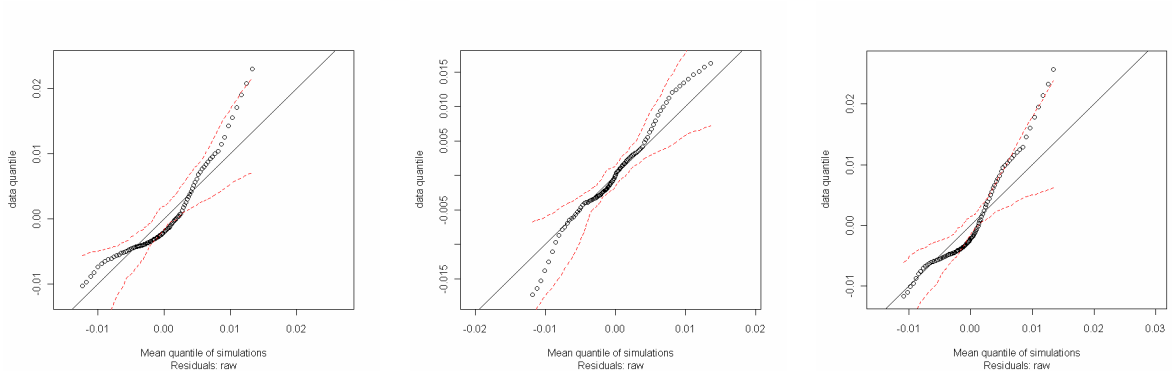


Figure 63 The quantile-quantile plots for the inhomogeneous Poisson model fitted to the year 2005 incidence process as function of coordinates and the `distmap` of year 1993 incidence data. The middle plot and the right plot incorporate `popdens` and `aged` respectively.

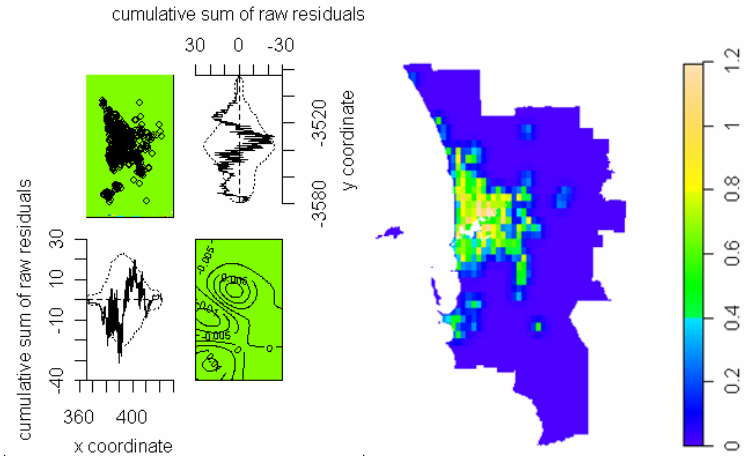


Figure 64 Diagnostic plots and the estimated rates map for model (6.1).

The diagnostic plots are developed to detect spatial trend and covariate effects. The diagnostic plots for model (6.1) are shown in left plot of Figure 64. The top left of the diagnostic plots is a mark plot, the top right and the bottom left are the lurking variable plots for the y-coordinate and x-coordinate respectively, and in the bottom right is the contour plot of the smoothed mark field. These plots can be used to assess goodness-of-fit, to identify outliers in the data, and to disclose departures from the fitted model (Baddeley et al. 2005). The dotted lines in the lurking variable plots are the 2σ-limits which have the usual interpretation of significance under the model (6.1). The mark plot does not show extreme outliers. For the x coordinate, the lurking variable plot indicates a slight overestimation between 400 and 410, and there is underestimation at $y=-3530$ and overestimation at $y=-3570$ and around $y=-3510$. Despite this, nearly all values of the coordinates stay within the 2σ-limits, and the plot in the lower right corner does not appear to have a visible trend, confirming that the model provides a good fit. The estimation map (right plot of Figure 64) indicates that the highest estimated incidence rate per 100 square kilometres in 2005 is 120 people.

For the standard deviations and coefficients of parameters in the fitted models (especially for `popdens` and `aged`) shown in Table 22, the difference in the coefficient of variation between parameters θ_2 is distinct. The coefficient of variation for the parameter `aged` is $\frac{3.2086}{1.3184} \approx 2.5$ and for the parameter `popdens` is $\frac{0.0002}{0.0005} \approx 0.4$. This indicates that

`popdens` is a better covariate than `aged` in terms of the measurement of dispersion of a probability distribution.

Table 22 Standard deviation of the parameters of the fitted model when different covariates are used to account for the intensity of lung cancer incidence point patterns.

	Parameters for model (6.1)		Parameters for model with covariate `aged`	
Parameter θ_i	Coefficient of θ_i	Std. Dev. $(\hat{\theta}_i)$	Coefficient f θ_i	Std. Dev. $(\hat{\theta}_i)$
intercept	-0.7083	0.4726	0.0273	1.0823
mapc93	-0.9874	0.0539	-1.0058	0.0535
popdens	0.0005	0.0002	----	----
aged	----	----	1.3184	3.2086
Transformation of coordinates	0.0601	0.0979	-0.1198	0.0724

Lung cancer incidence rates are high in Kwinana industrial area according to the kriging analysis (5.1.1). Does the distance from the Kwinana industrial area have an influence on lung cancer? Finally, Kwinana industrial area (384.161, 3536.393) is used as reference point for modelling lung cancer incidence data. The quantile-quantile plots shown in Figure 65 involve the spatial trends (distance from Kwinana industrial area) and the susceptible population intensity. The middle and the right plot incorporate population density and the percentage of people aged 50 or above (see Figure 19) respectively. The middle plot indicates that the distance from the Kwinana industrial area accounts for lung cancer incidence in 2005 when the population density is considered. The fitted model is:

$$\lambda_\theta(\mathbf{u}) = ((x-384.161)^2 + (y+3536.393)^2)^{0.149} \times \exp(-1.886 - 0.977\, Z_1(\mathbf{u}) + 0.001\, Z_2(\mathbf{u})) \quad (6.2)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 1993 incidence point patterns and $Z_2(\mathbf{u})$ represents population density. The diagnostic plots for the model (6.2) shown in left of Figure 66 confirm that the fitted model is adequate. The estimation map (right plot of Figure 66) demonstrates that the highest estimated incidence rate per 100 square kilometres in 2005 is 120 people.

The intensity of the point process in 2005 is given by model (6.1) and model (6.2) from different perspectives. The first model is based on the distance from the Perth Business district and the second is based on the distance from the Kwinana industrial area. They all

indicate that the point process in 2005 is similar to that in 1993 and the fitted models are adequate to describe the incidence data in 2005. In fact, the point patterns in other individual years have been tested and only the point pattern for year 1993 is a good covariate.

Distmap + spatial trends

distmap+ spatial trends + **popdens**

distmap+spatial trends + **aged**

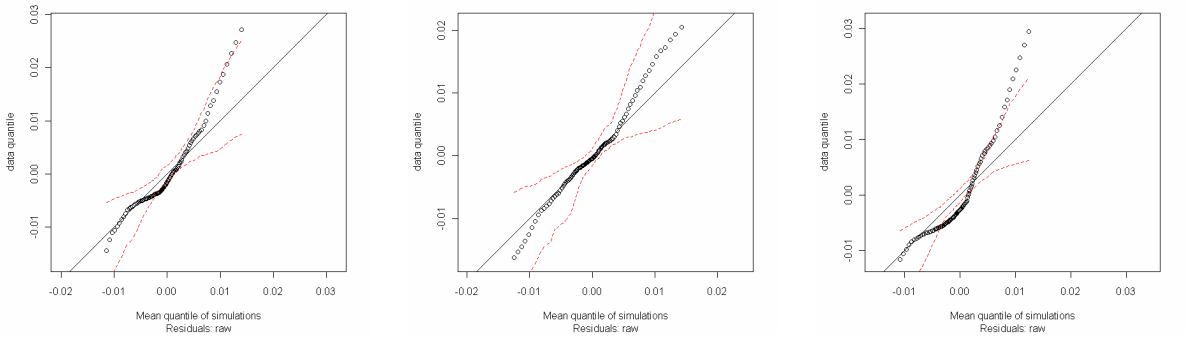


Figure 65 The quantile-quantile plots for the inhomogeneous Poisson model fitted to the year 2005 incidence process as function of coordinates in which Kwinana industrial area is used as reference, and the `distmap` of year 1993 incidence data. The middle and the right plot incorporate `popdens` and `aged` respectively.

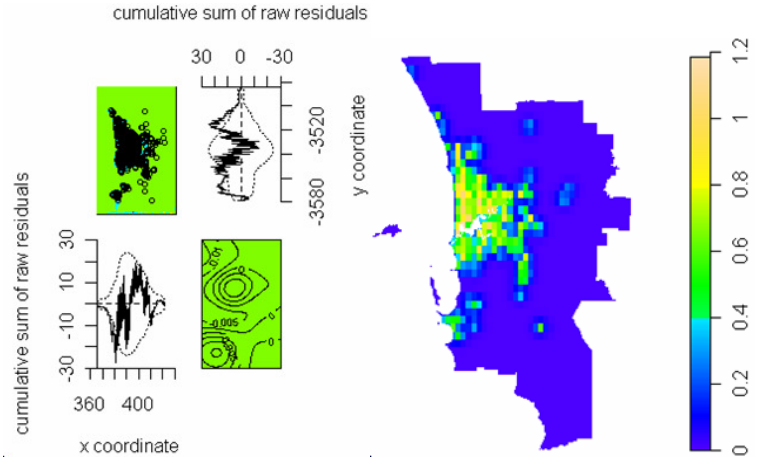


Figure 66 Diagnostic plots and the estimated rates (per square kilometres) map for model (6.2).

The parameters in the model (6.1) and (6.2) are shown in Table 23. There is slight difference between two models except intercept and transformation of coordinates in terms of coefficients and standard deviations. The coefficient of variation for the intercept is $\frac{0.4726}{0.7083} \approx 0.67$ in model (6.1) and it is $\frac{0.6152}{1.9799} \approx 0.31$ for the other model. The coefficient of variation for the spatial trends (transformation of coordinates) is $\frac{0.0979}{0.0601} \approx 1.63$ in model (6.1) and it is $\frac{0.1243}{0.3256} \approx 0.38$ for the other model. This indicates

that Kwinana industrial area is a better reference location for lung cancer incidence modelling than the Perth CBD.

Table 23 Standard deviation of the parameters of the fitted model when different spatial trends are used to account for the intensity of lung cancer incidence point patterns.

	Parameters for model (6.1)		Parameters for model (6.2)	
Parameter θ_i	Coefficient of θ_i	Std. Dev. $(\hat{\theta}_i)$	Coefficient of θ_i	Std. Dev. $(\hat{\theta}_i)$
intercept	-0.7083	0.4726	-1.9799	0.6152
mapc93	-0.9874	0.0539	-0.9834	0.0534
popdens	0.0005	0.0002	0.0010	0.0002
Transformation of coordinates	0.0601	0.0979	0.3256	0.1243

6.1.3 Analysis of lung cancer mortality point patterns

We will analyse the 2005 mortality lung cancer data and will use them to explore whether slight differences in point patterns have influence on model fitting, and if yes, what kind of data treatment will produce better model fitting. Turning to lung cancer mortality in 2005, there are 531 cases with five locations without coordinate records in the original data. In this section, the data to be analysed and covariates used (based on incidence point patterns) are treated using two methods respectively.

First, the data treated using the first method are analysed. Survival rates (Figure 3) for lung cancer are relatively low compared with other cancers. About 40% of lung cancer patients were alive one year and 28% were alive two years after diagnosis. It has been found that year 2003 incidence point patterns are more suitable to be used as a covariate in modeling mortality data based on the survival plots. The fit is very good when the model only incorporates the spatial trend and the spatially susceptible population (see left plot of Figure 67). The fit is tolerably good when `aged` is taken into consideration (see right plot of Figure 67). However, the fitting becomes worse when `popdens` is added as covariate (see middle plot of Figure 67). Converting back to the form (2.39), the fitted model for the left plot of Figure 67 is

$$\lambda_\theta(\mathbf{u}) = \frac{\exp(0.998 - 1.511 \times Z_1(\mathbf{u}))}{((x - 391.063)^2 + (y + 3536.509)^2)^{0.181}} \qquad (6.3)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 2003 incidence point patterns.

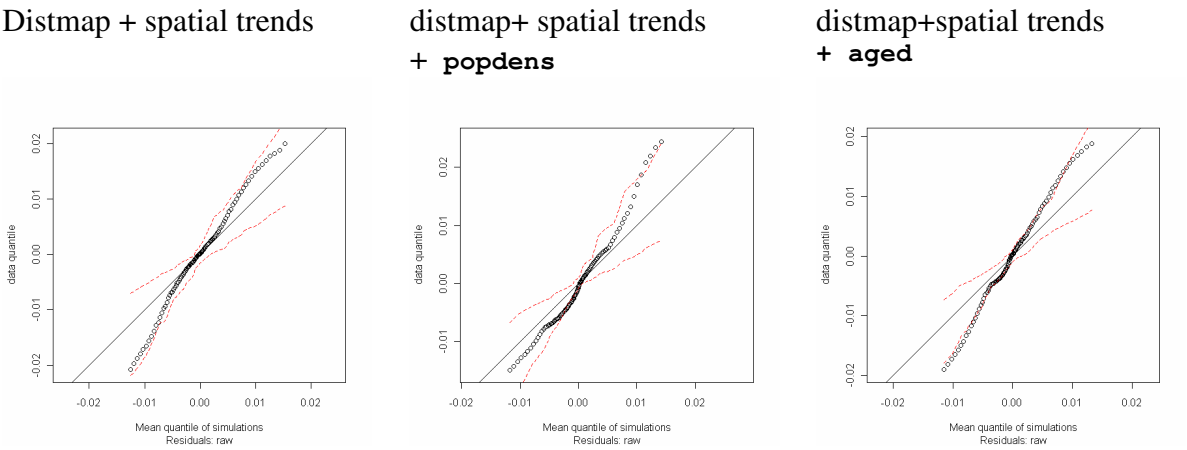


Figure 67 The quantile-quantile plots for the inhomogeneous Poisson model fitted to the year 2005 mortality data. The model involves transformations of coordinates and the `distmap` of year 2003 incidence data. The middle plot and the right plot incorporate `popdens` and `aged` respectively.

Next, the data are treated using the second method. The fitting is adequate when the model incorporates spatial trend and the spatially susceptible population in 2003. Converting back to the form (2.39), the fitted model is

$$\lambda_\theta(\mathbf{u}) = \frac{\exp(1 - 1.543 \times Z_1(\mathbf{u}))}{((x - 391.063)^2 + (y + 3536.509)^2)^{0.179}} \qquad (6.4)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 2003 modified incidence point patterns. The quantile-quantile plot for this model is shown in the left plot of Figure 68. When the covariate population density (middle plot) or the proportion of the people aged 50 or over (right plot) is considered respectively, the fit does not improve.

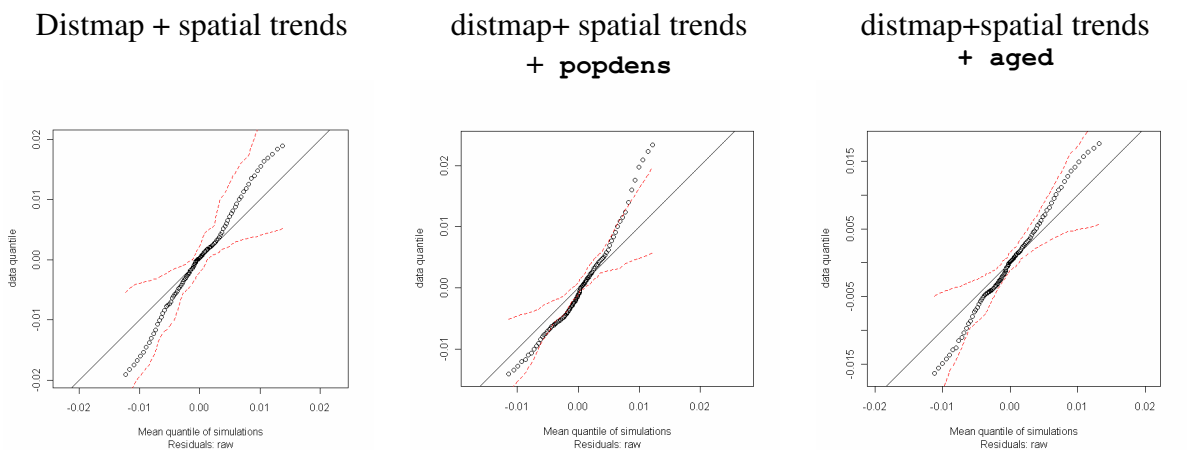


Figure 68 The quantile-quantile plots for the inhomogeneous Poisson model fitted to the year 2005 modified mortality data plotted in Figure 60. The model involves transformations of coordinates and the `distmap` of year 2003 modified incidence data. The middle and the right plot incorporate `popdens` and `aged` respectively.

The diagnostic plots for models (6.3) and (6.4) are shown in Figure 69. In bottom right panel, the colour scheme represents the value of residuals: negative residuals are in blue, positive residuals are in red and zero residuals are in green. The plots indicate clearly that *y* is a better explanatory variable compared with *x*. The two plots are nearly the same and they do not show distinct trend, so the fitted models are accepted.

When the spatial covariates play an indispensable part in the model fitting, the lurking variable plots can be used to investigate how much the intensity function is dependent on the covariates. The total residuals for the entire observation window must equal zero. The plots start and finish at the horizontal axis where the residuals are zero.The lurking variable plots (see Figure 70) illustrate that the fitted function for mortality data captures the dependence of intensity on these covariates. The proportion of the population age 50 or above (`aged`) accounts for mortality data more than the population density.

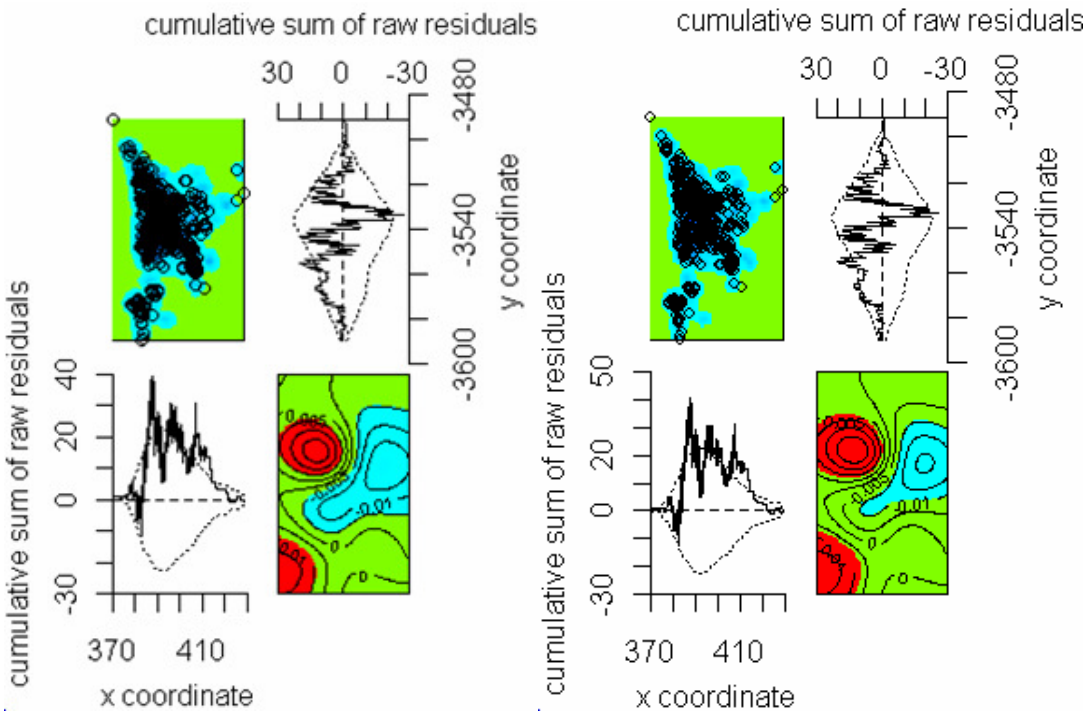


Figure 69 Diagnostic plots (left) for the model (6.3) and the right for the model (6.4).

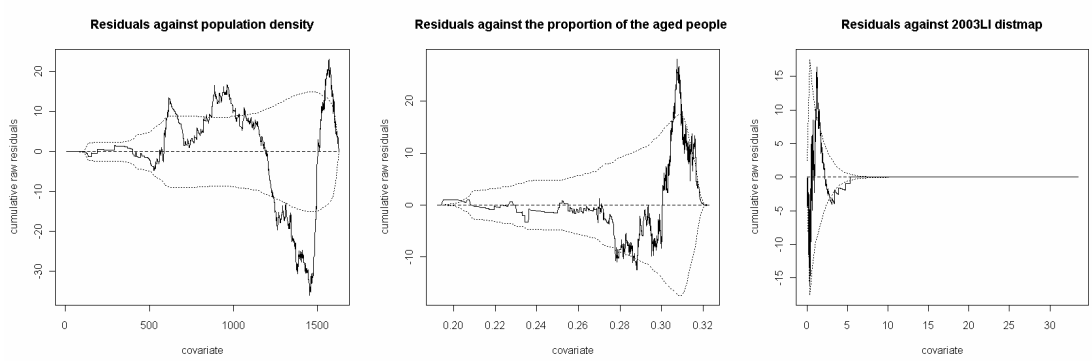


Figure 70 Lurking variable plots of lung cancer mortality for the spatial covariates: `popdens` (left), `aged` (middle) and the `distmap` of year 2003 lung cancer incidence data (right).

The coefficients and standard deviations of parameters are compared when the data analysed and the data used as covariates (`distmap`) are treated in two ways individually (see Table 24). According to this table, the coefficient and standard deviations are comparable for the two methods, with standard deviations for method 2 slightly lower.

Table 24 Standard deviation of the parameters of the fitted model when 2005 lung cancer mortality data and covariates (lung cancer incidence case locations) are treated using two methods.

	Method 1		Method 2	
Covariates	Coefficient of θ_i	Std. Dev. ($\hat{\theta}_i$)	Coefficient of θ_i	Std. Dev. ($\hat{\theta}_i$)
intercept	0.9984	0.1605	1.0000	0.1588
Mapc2003	-1.5105	0.0867	-1.5427	0.0867
Transformation of coordinates	-0.3614	0.0621	-0.3575	0.0614

Unless otherwise stated, in the further discussions method 2 will be used to accommodate data with missing coordinates, but known SLA. From now on the data treated with this method are called point patterns rather than modified point patterns for the remaining cancers.

6.2 Melanoma analysis

For spatial point pattern analysis of melanoma cancer data, the available covariates used in the model fitting are population density (`Popdens`), the percentage of people aged 30 or over (`aged30+`) shown in Figure 19 and susceptible population intensity (Figure 23). Looking back at section 5.2.1, kriging analysis has shown that melanoma cancer incidence is relatively high along the coast, and so the coastline will be used as reference as well for incidence modelling to explore whether the coast accounts for melanoma cancer incidence. The mortality data are analysed in the last section.

6.2.1 Exploratory data analysis

The point pattern in 2005 for melanoma cancer incidence and mortality (Figure 71) is not completely random and this has been verified in the beginning of Chapter 6. The inhomogeneous *K*-function (Baddeley et al. 2000) is applied in the data analysis in order to take into account the inhomogeneity of data characteristics.

The estimated *K*-function plot for the year 2005 incidence data is shown in left plot of Figure 72 and the right plot is for the mortality data. These two plots indicate that there is

no obvious attraction or inhibition for the year 2005 cancer data, so the point process is inhomogeneous Poisson process.

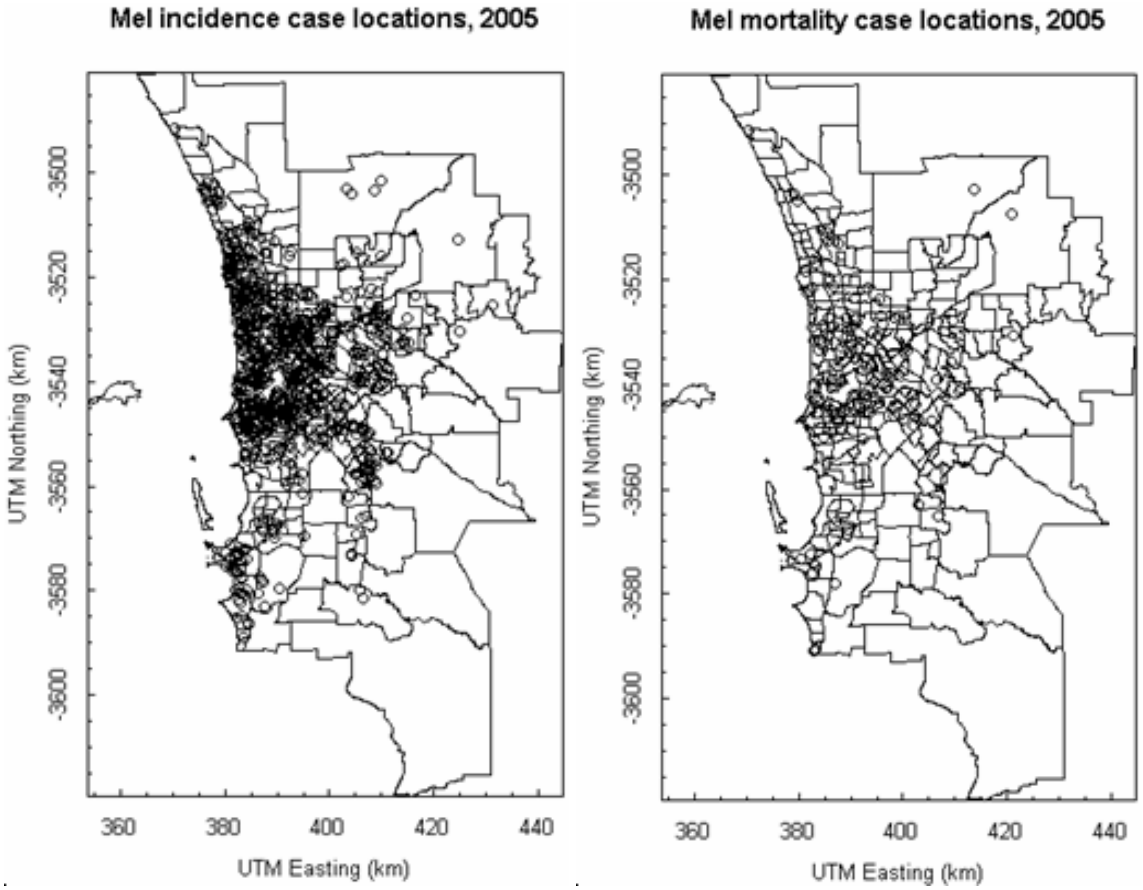
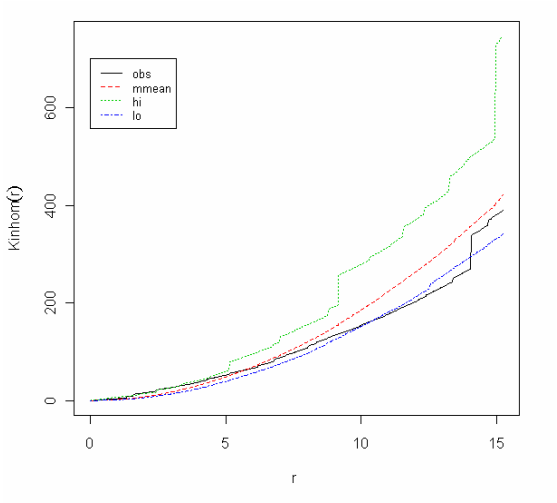


Figure 71 Melanoma cancer incidence (left) and mortality (right) case locations in year 2005 in Perth.

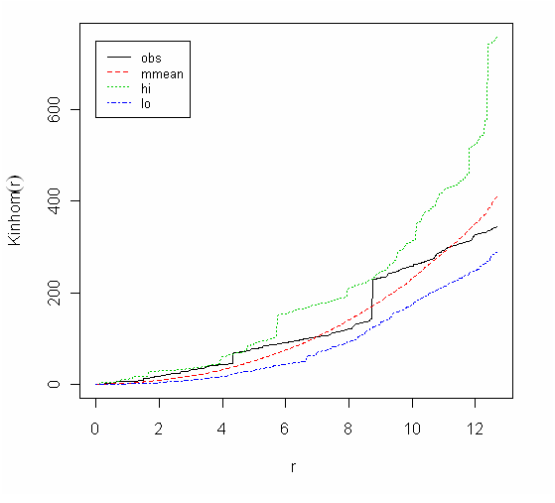
Incidence



Bandwith for smoothing intensity : 3.5

Bandwith for simulation from data: 5.5

Mortality



Bandwith for smoothing intensity: 7

Bandwith for simulation from data: 7

Figure 72 Estimates of the inhomogeneous K-function for the year 2005 melanoma cancer incidence and mortality data.

For the relative proportions of intensity of 16-year period, melanoma cancer incidence and mortality in year 1990, 1993, 1997, 2001 and 2005 are shown in Figure 73. It can be seen from Figure 73 that spatial trends are totally different in the selected years for incidence. It can be seen clearly that incidence cases increase in Perth during the period 1990-2005. However, the difference in the relative proportions for mortality is not significant.

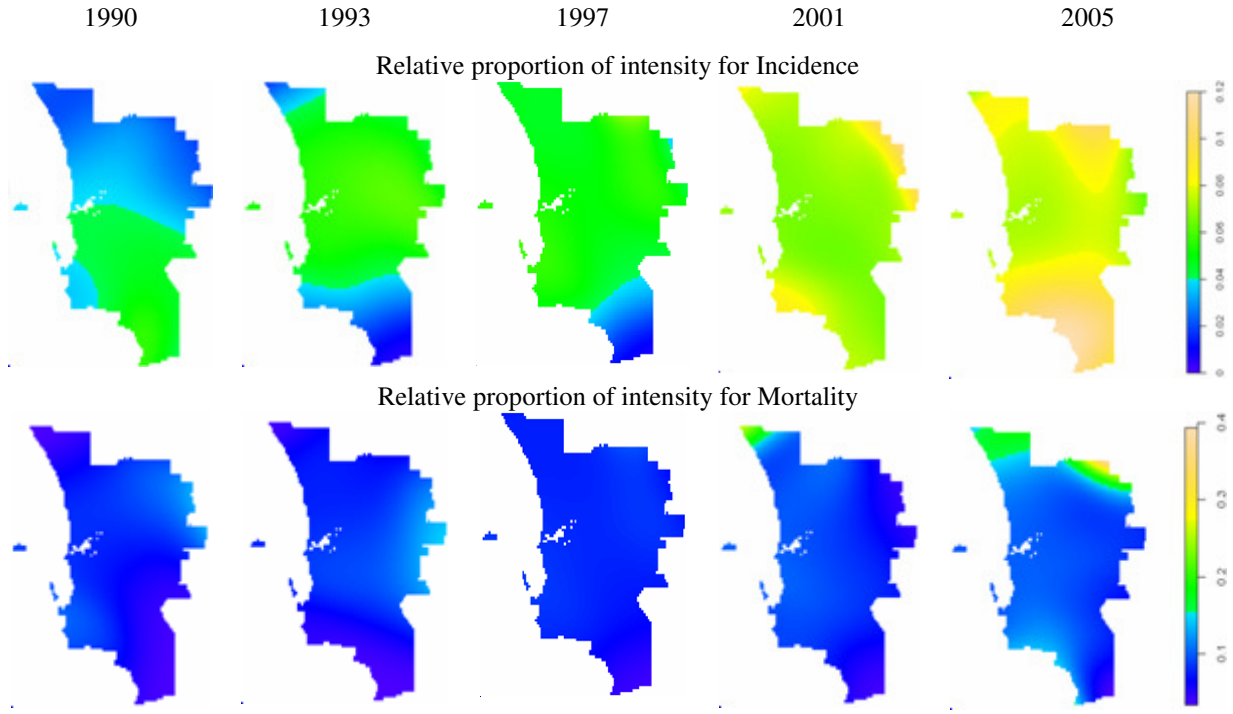


Figure 73 The relative proportion of intensity of 16-year period for melanoma cancer incidence (top) and mortality (bottom) in year 1990, 1993, 1997, 2001 and 2005.

6.2.2 Melanoma cancer incidence 2005 model fitting

In this section, Perth CBD is used first as the reference point in modelling. Then coastline is used as the reference line for incidence modelling to check whether proximity to the coast accounts for melanoma incidence.

It has been found that the year 2001 point pattern is more related to year 2005 data compared with point patterns in other years. An adequate model is obtained when susceptible population based on year 2001 and population density are used as covariates to model the year 2005 data. The fitted model is shown below:

$$\lambda_\theta(\mathbf{u}) = ((x-391.063)^2 + (y+3536.509)^2)^{0.057} \times \exp(-1.923 - 1.009\, Z_1(\mathbf{u}) + 0.001\, Z_2(\mathbf{u})) \quad (6.5)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 2001 melanoma cancer incidence point pattern and $Z_2(\mathbf{u})$ denotes population density (`popdens`).

The corresponding quantile-quantile plot, diagnostic plots and the estimation map for the model (6.5) are shown in Figure 74. The quantile-quantile plot just stays within the theoretical ideal line and this indicates that the agreement is good between the fitted model and the data. For diagnostic plots: mark plot shows no extreme outliers; nearly all the lurking variable plots for the coordinates stay within the 2σ-limits; contour plot does not appear to have a visible trend, so the fitted model is a good fit. The estimation map indicates that the highest estimated incidence rate per 100 square kilometres in 2005 is about 140 people. The proportion of people above 30 is not a good explanatory variable according to the quantile plot Figure 75.

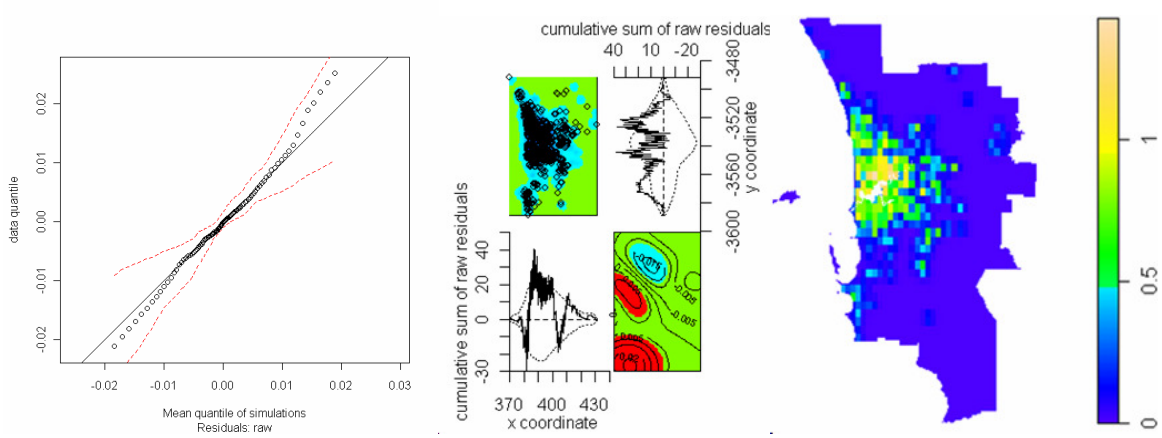


Figure 74 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.5) fitted to the year 2005 melanoma cancer incidence data.

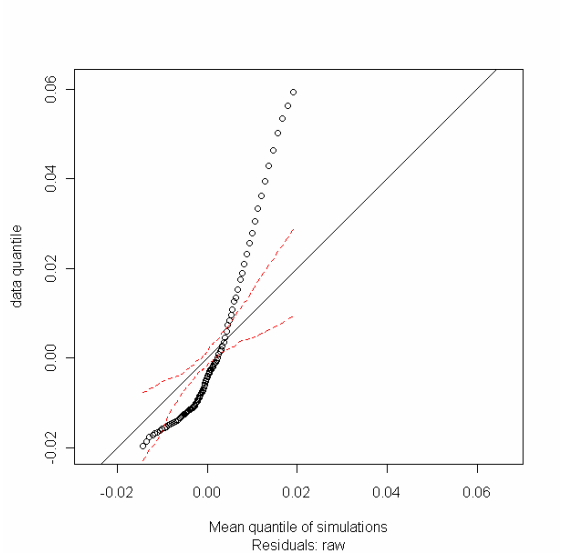


Figure 75 The quantile-quantile plot for the inhomogeneous Poisson model fitted to the year 2005 melanoma cancer incidence data. The models involve spatial trends in which Perth city center is the reference point, `aged30+` and the `distmap` of year 2001 melanoma cancer incidence data.

Next, the coastline is used as the reference line for the modelling. The coastline looks like a parabola which opens west with the horizontal line ($y=-3531.256$) as the axis of symmetry. This horizontal line passes the middle of metropolitan Perth. The point (381.8623, -3531.256) is used as the origin. Then the other point (373.796, -3499.855) on the northwestern border of metropolitan Perth is selected as the passing point of parabola. And so the equation for parabola is:

$$x = -0.0082 \times (y + 3531.256)^2 + 381.8623$$

An adequate model is obtained when susceptible population based on year 1997 and population density are used as covariates to model the year 2005 data. The fitted model is shown below:

$$\lambda_\theta(\mathbf{u}) = \frac{\exp(-1.202 - 0.931 \times Z_1(\mathbf{u}) + 0.001 \times Z_2(\mathbf{u}))}{\left| x + 0.0082 \times (y + 3531.256)^2 - 381.8623 \right|^{0.058}} \qquad (6.6)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 1997 melanoma cancer incidence point pattern and $Z_2(\mathbf{u})$ denotes population density (`popdens`). The distance between the point (x, y) and the coastline is denoted by $\left| x + 0.0082 \times (y + 3531.256)^2 - 381.8623 \right|$. The corresponding quantile-quantile plot, diagnostic plots and the estimation map for the model (6.6) are shown in Figure 76. The quantile-quantile plot indicates that the agreement is good between the fitted model and the data. The diagnostic plots do not show extreme outliers or visible trend. The estimation map indicates that the highest estimated incidence rate per 100 square kilometres in 2005 is about 135 people which is slightly lower than the estimation in model (6.5) although these two estimation maps are quite similar. Overall, the coastline accounts for melanoma incidence when it is used as the spatial trends. The two fitted intensity functions produce very similar estimates according to the estimation maps shown in Figure 74 and Figure 76. The intensity of the melanoma incidence point process is given by model (6.5) or (6.6) when different reference locations are used. However, the point patterns used as covariates are not the same: one is year 2001 in model (6.5) and the other is year 1997 in (6.6). These two models describe the point process properly using different spatial trends.

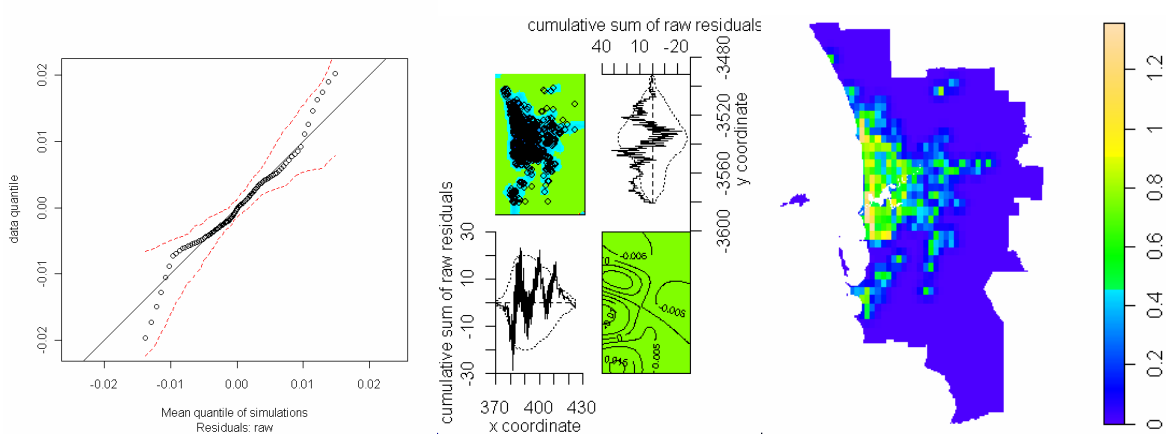


Figure 76 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.6) fitted to the year 2005 melanoma cancer incidence data.

6.2.3 Melanoma cancer mortality 2005 model fitting

The survival rates (Figure 3) for melanoma cancer are relatively high compared with other cancers. About 65% of people with melanoma were alive four years and 50% six years after diagnosis. To model mortality data in 2005, the susceptible population density in year 2001 is best suited for fitting compared with the point patterns in other years.Neither population density (`popdens`) nor the proportion of people aged 30+ (`aged30+`)are needed to obtain an adequate fit. The fitted model is given by:

$$\lambda_\theta(\mathbf{u}) = \frac{\exp(-0.617 - 1.397 \times Z_1(\mathbf{u}))}{((x-391.063)^2 + (y+3536.509)^2)^{0.195}} \qquad (6.7)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of the 2001 melanoma cancer incidence point pattern.

The quantile-quantile plot, diagnostic plots and estimation map for the model (6.7) are shown in Figure 77. The quantile-quantile plot indicates an adequate agreement between model (6.7) and mortality data. No extreme outliers or distinct spatial trends are detected from the diagnostic plots. Also, nearly all the lurking variable plots for the coordinates stay within the 2σ-limits. The estimation map indicates that the highest estimated mortality rate per 100 square kilometres in 2005 is 25 people. To obtain an adequate fit, modelling incidence needs covariates such as `popdens` or `aged30+`. However, these covariates are not necessary for modelling melanoma mortality.

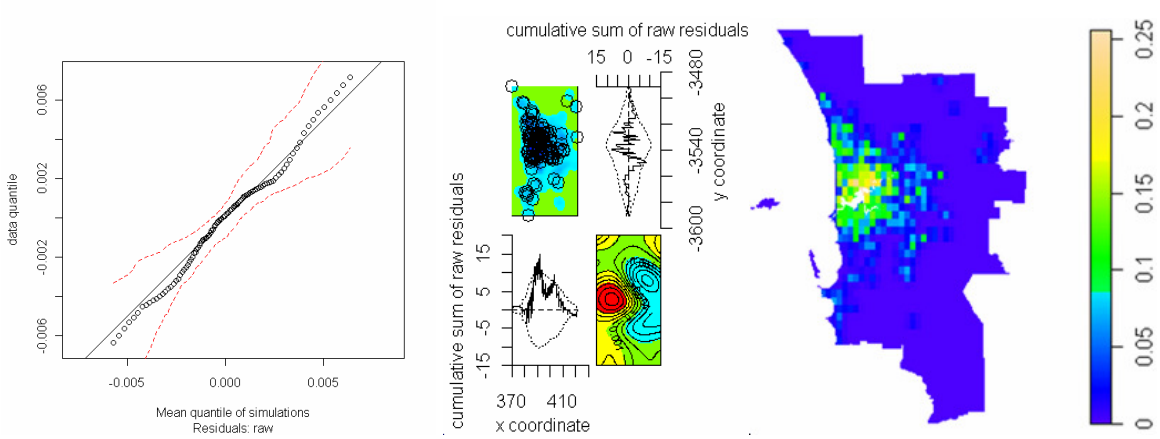


Figure 77 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.7) fitted to the year 2005 melanoma cancer mortality data.

The lurking variable plot (Figure 78) indicates that the fitted model is able to capture the dependence of intensity on the susceptible population intensity although there are values outside the envelope.

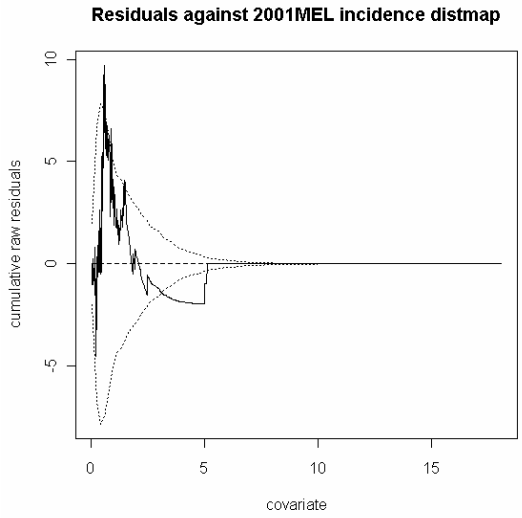


Figure 78 Lurking variable plot for the empty space distances of 2001 melanoma cancer incidence used in model (6.7).

6.3 Colorectal cancer analysis

6.3.1 Exploratory data analysis

The estimated *K*-function plots for the year 2005 colorectal cancer data (Figure 79) are shown in Figure 80. The left plot is for incidence data and the right plot is for mortality data.

These two plots indicate that there is no obvious attraction or inhibition for the year 2005 cancer poin patterns, so the point process is inhomogeneous Poisson process.

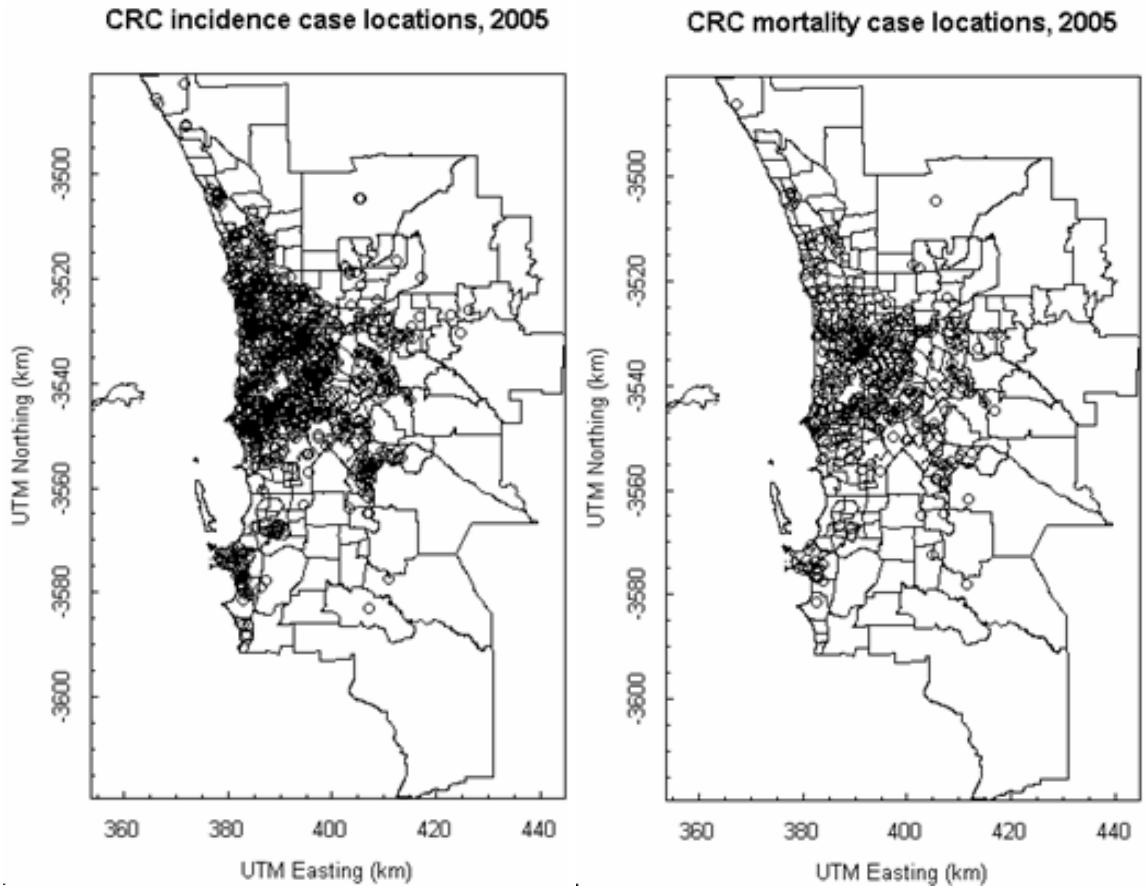
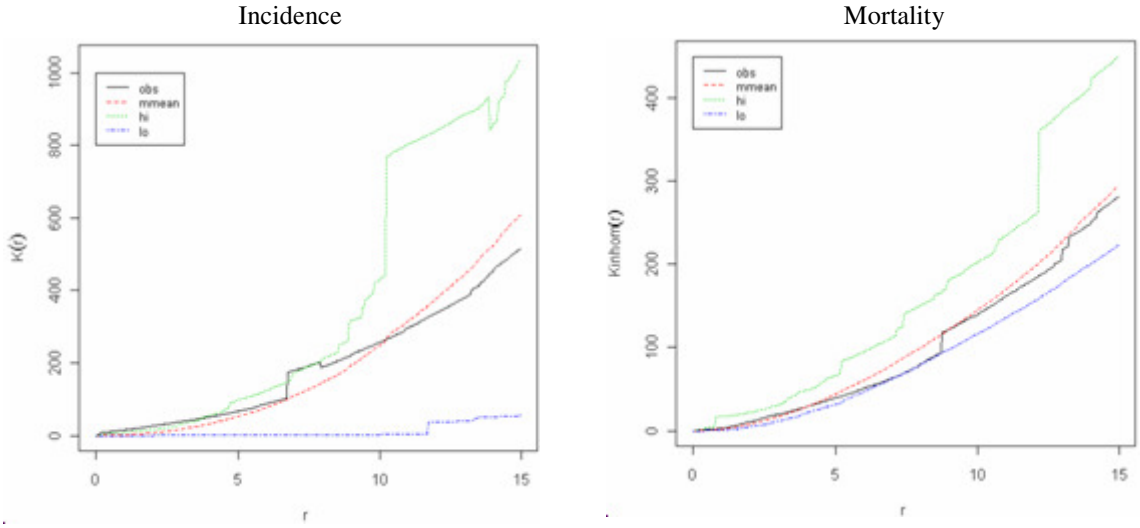


Figure 79 Colorectal cancer incidence (left) and mortality (right) case locations in year 2005 in Perth.



Bandwith for smoothing intensity :4.5
Bandwith for simulation from data: 4.5

Bandwith for smoothing intensity: 2.5
Bandwith for simulation from data: 6

Figure 80 Estimates of the inhomogeneous K-function for the year 2005 colorectal cancer incidence and mortality data.

For colorectal cancer incidence and mortality, the relative proportions of intensity of 16-year period in year 1990, 1993, 1997, 2001 and 2005 are shown in Figure 81. It can be seen that spatial trends in year 1990, 1993 and 1997 are quite similar but the trends are different from those in year 2001 and 2005 for both incidence and mortality. The geographical difference of incidence over the years is distinct compared with mortality.

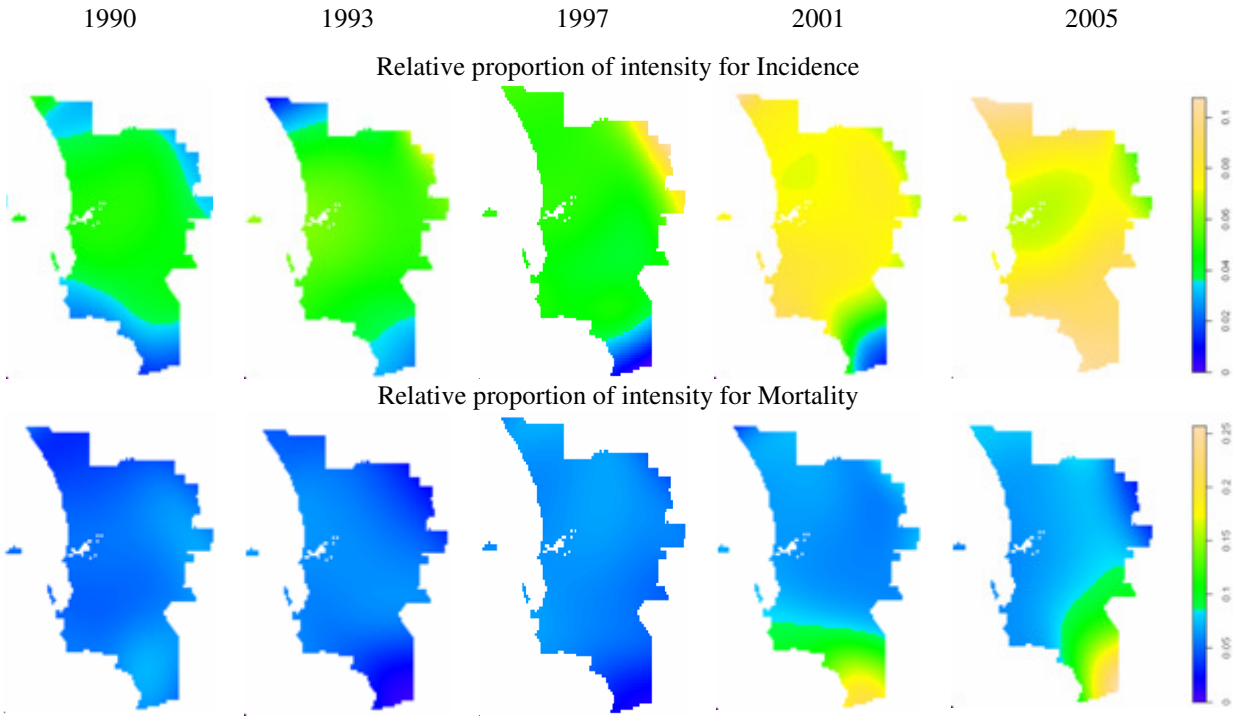


Figure 81 The relative proportion of intensity of 16-year period for colorectal cancer incidence (top) and mortality (bottom) in year 1990, 1993, 1997, 2001 and 2005.

6.3.2 Colorectal cancer incidence 2005 model fitting

It was ientified that the year 1998 point pattern is more similar to year 2005 colorectal incidence point pattern compared with point patterns in other years. An appropriate model is obtained when spatial trends, susceptible population based on year 1998 and population density (`popdens`) are used as covariates to model the year 2005 data. The fitted model is shown below:

$$\lambda_\theta(\mathbf{u}) = \frac{\exp(-1.079 - 0.956 \times Z_1(\mathbf{u}) + 0.001 \times Z_2(\mathbf{u}))}{((x - 391.063)^2 + (y + 3536.509)^2)^{0.011}} \qquad (6.8)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 1998 colorectal cancer incidence point pattern and $Z_2(\mathbf{u})$ denotes population density (`popdens`).

The corresponding quantile-quantile plot, diagnostic plots and estimation map for the model (6.8) are shown in Figure 82. For Q-Q plot, the line of observed value stays close to the line of theoretical value and this indicates that the agreement is adequate between the fitted model and the data. For diagnostic plots: no extreme outliers are found in mark plot; nearly all the lurking variable plots for the coordinates stay within the 2σ-limits; the contour plot does not appear to have any visible trend, so the fitted model is acceptable. When we replace the covariate `popdens` in model (6.8) using `aged`, the fit becomes worse. The estimation map indicates that the highest estimated incidence rate per 100 square kilometres in 2005 is 100 people.

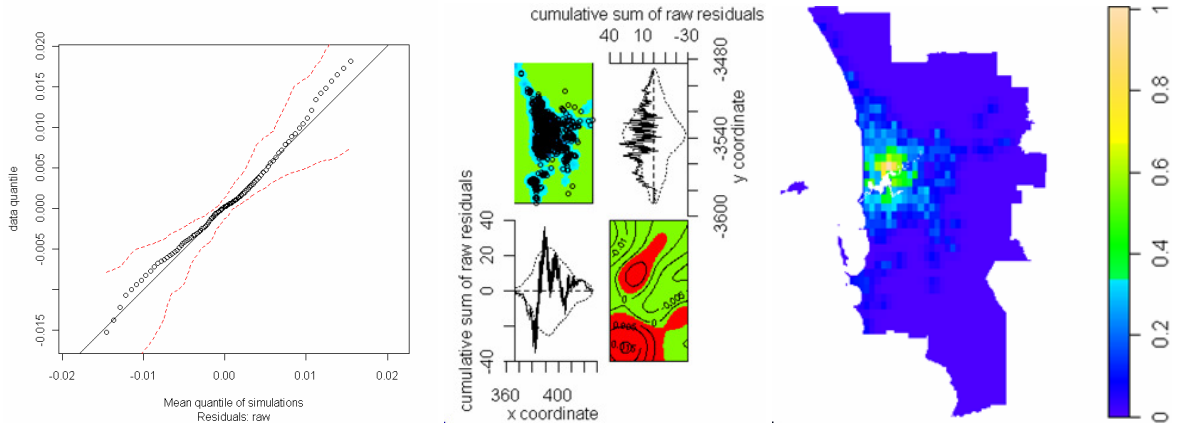


Figure 82 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.8) fitted to the year 2005 colorectal cancer incidence data.

6.3.3 Colorectal cancer mortality 2005 model fitting

The survival rates (Figure 3) for colorectal cancer patients are higher than lung cancer and lower than melanoma cancer during the sixteen-year period. About 60% of people diagnosed with colorectal cancer were alive after two years and 50% were alive for three years after diagnosis. To model mortality data in 2005, susceptible population density in

year 2003 is more suitable for the fitting compared with the point patterns in other years. An adequate fit is obtained even when population density (`popdens`) or the proportion of people aged 50+ (`aged`) is not included in the model. The fitted model is:

$$\lambda_\theta(\mathbf{u}) = \frac{\exp(0.473 - 1.357 \times Z_1(\mathbf{u}))}{((x - 391.063)^2 + (y + 3536.509)^2)^{0.231}} \tag{6.9}$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 2003 colorectal cancer incidence point pattern.

The quantile-quantile plot, diagnostic plots and estimation map for model (6.9) are shown in Figure 83. The Q-Q plot indicates an adequate agreement between the fitted model and mortality data. No "extreme" outliers or distinct spatial trends are detected from the four panel plots. Also, nearly all the lurking variable plots for the coordinates just stay within the 2σ-limits. The estimation map based on model (6.9) shows the spatial distribution of cancer rates and it indicates that the highest estimated mortality rate per 100 square kilometres in 2005 is about 65 people in the Perth metropolitan area. Compared with modelling colorectal cancer incidence, modelling mortality does not have to include covariates such as `popdens` or `aged` to obtain an appropriate fit.

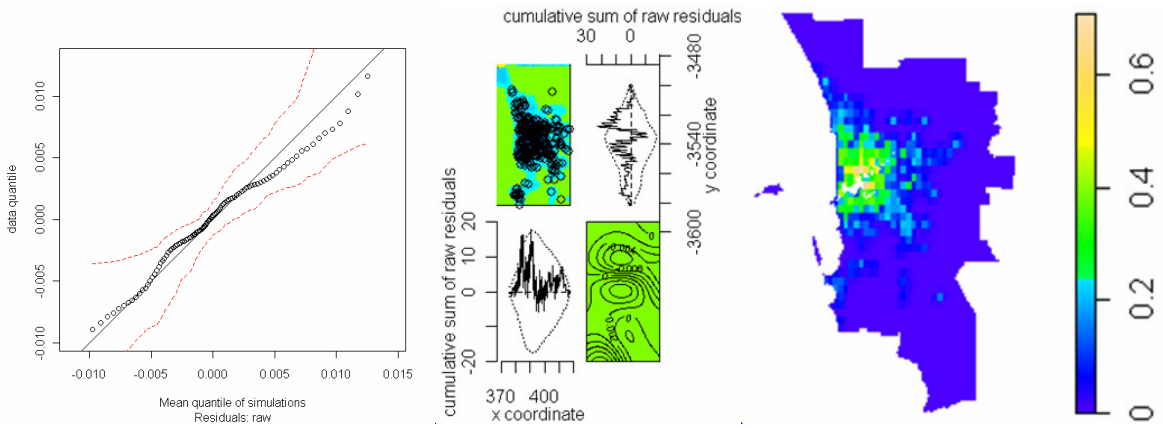


Figure 83 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.9) fitted to the year 2005 colorectal cancer mortality data.

The lurking variable plot (Figure 84) indicates that the fitted model (6.9) is able to capture the dependence of intensity on the covariate since the plot is not totally outside of the significance bands and also stays over and below the horizontal axis.

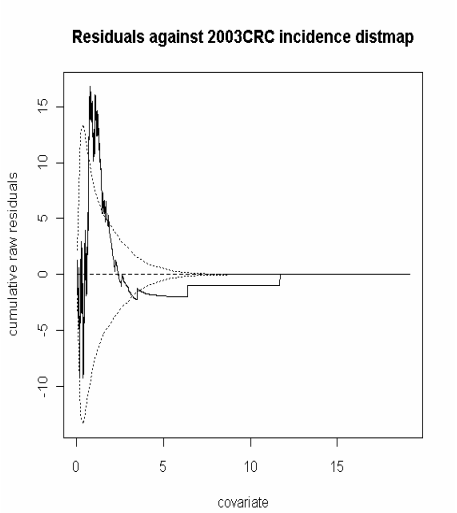


Figure 84 Lurking variable plots of colorectal cancer mortality in 2005 model fitting for covariates: empty space distances of 2003 colorectal cancer incidence based on model (6.9). The fitting involves transformations of coordinates.

6.4 Breast cancer analysis

For the spatial point pattern analysis of breast cancer data, the available covariates used in the model fitting are the female population density (`Popdensf`), the proportion of females aged 40 or over (`agedf`) shown in Figure 20, and susceptible population intensity (Figure 25).

6.4.1 Exploratory data analysis

In Australia, screening mammograms are used for the early detection of breast cancer. Women aged 40 or over are eligible for the testing. The development of screening mammograms contributes to an early diagnosis of breast cancer. The breast cancer risk increases with age and there is a marked increase when females are aged 40 or above (Figure 4) and so the proportion of females aged 40+ is considered as covariate for breast cancer model fitting.

Plots of an estimate of the inhomogeneous *K*-function, for the incidence and mortality data, are shown in
Figure 86. The sigma value for the left plot (

Figure 86) is 5 to calculate intensity and to simulate from the incidence data point pattern. The estimated *K*-function is in the range 0 to 1500. The plot indicates no obvious attraction or inhibition for the incidence data point process, so the point process is inhomogeneous. For the mortality point process (right plot), the estimated *K* values are slightly lower than the low bound of the critical envelope when the r values are in the range 9 to 12. Although there is evidence against inhomogeneous Poisson process, the mortality cases are more dispersed than the incidence cases from Figure 85. Also, the eastern part of the study region is sparsely populated (see Figure 20). This suggests "inhibition" indicated by the *K*-function plot.

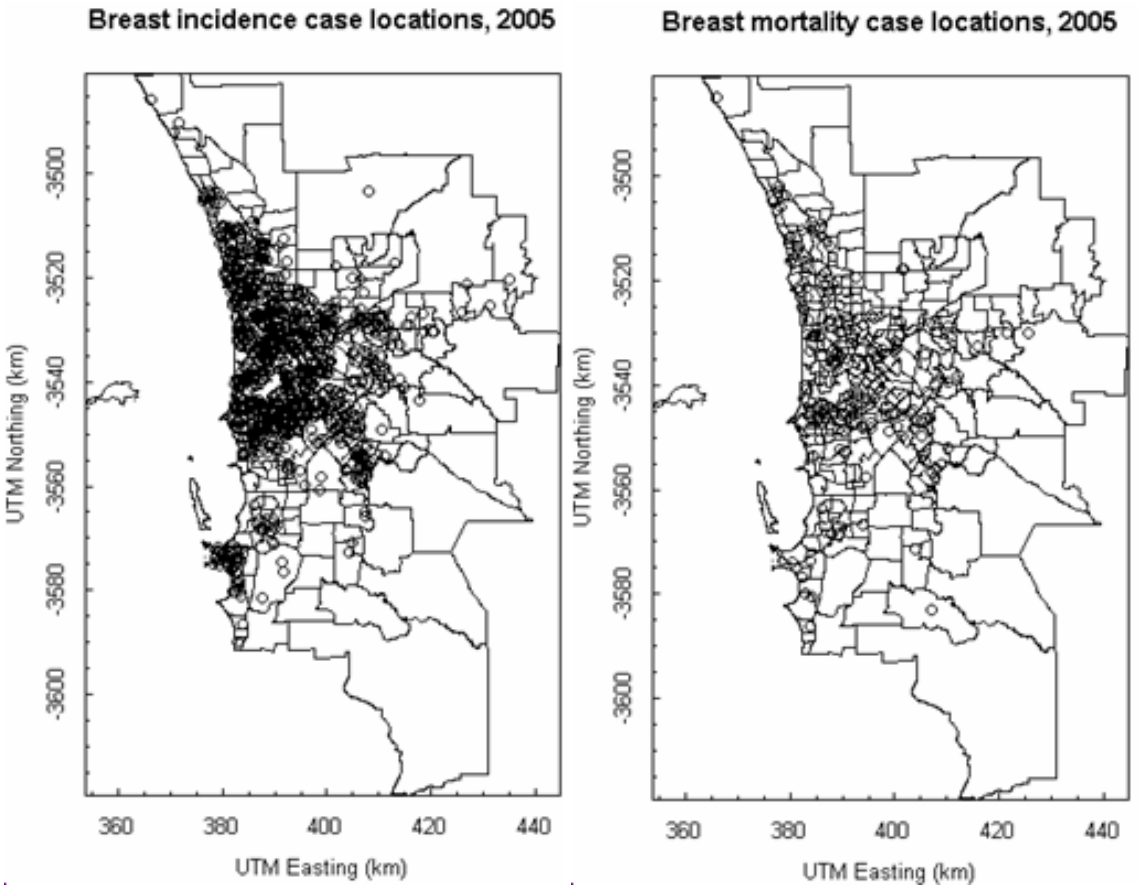
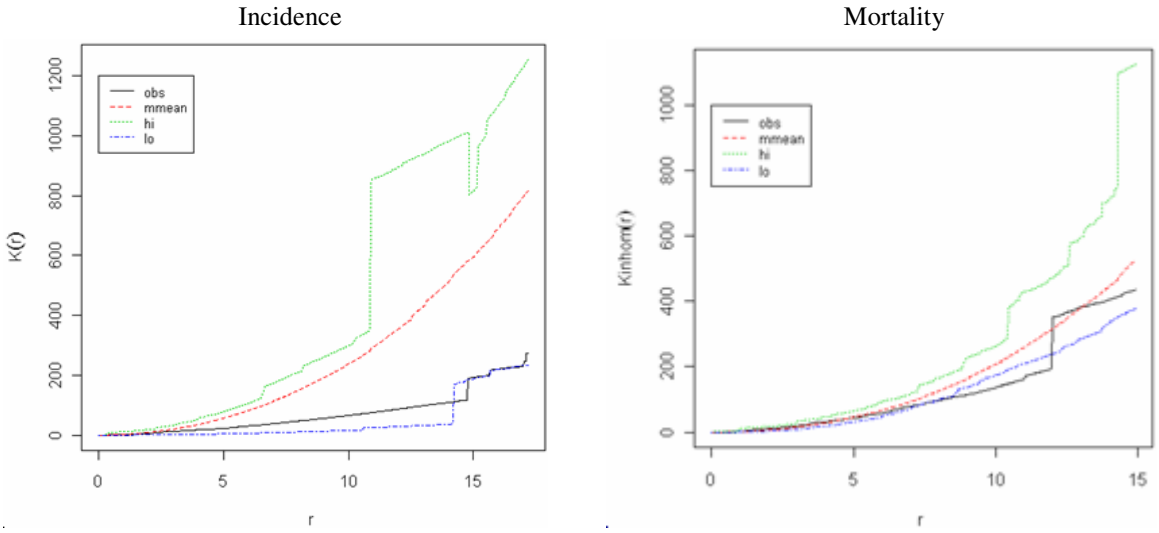


Figure 85 Breast cancer incidence (left) and mortality (right) case locations in year 2005 in Perth.



Bandwith for smoothing intensity : 5
Bandwith for simulation from data: 5

Bandwith for smoothing intensity: 6
Bandwith for simulation from data: 6

Figure 86 Estimates of the inhomogeneous K-function for the year 2005 breast cancer incidence and mortality data.

For the relative proportions of intensity of 16-year period, breast cancer incidence and mortality in year 1990, 1993, 1997, 2001 and 2005 are shown in Figure 87. Spatial trends are totally different in the selective years for incidence and mortality. There is a gradual increase with the year for incidence in most geographic areas, and this is mainly due to the fact that the number of incidence cases increased gradually with the year (Figure 2). Generally speaking, compared with the total number of cases in the period 1990-2005, cancer mortality cases increase in the north of the study area during the four selected years. Compared with the total number of cases in the period 1990-2005, the proportion of mortality cases in Perth centre is relatively high in 1990, 1997 and 2001.

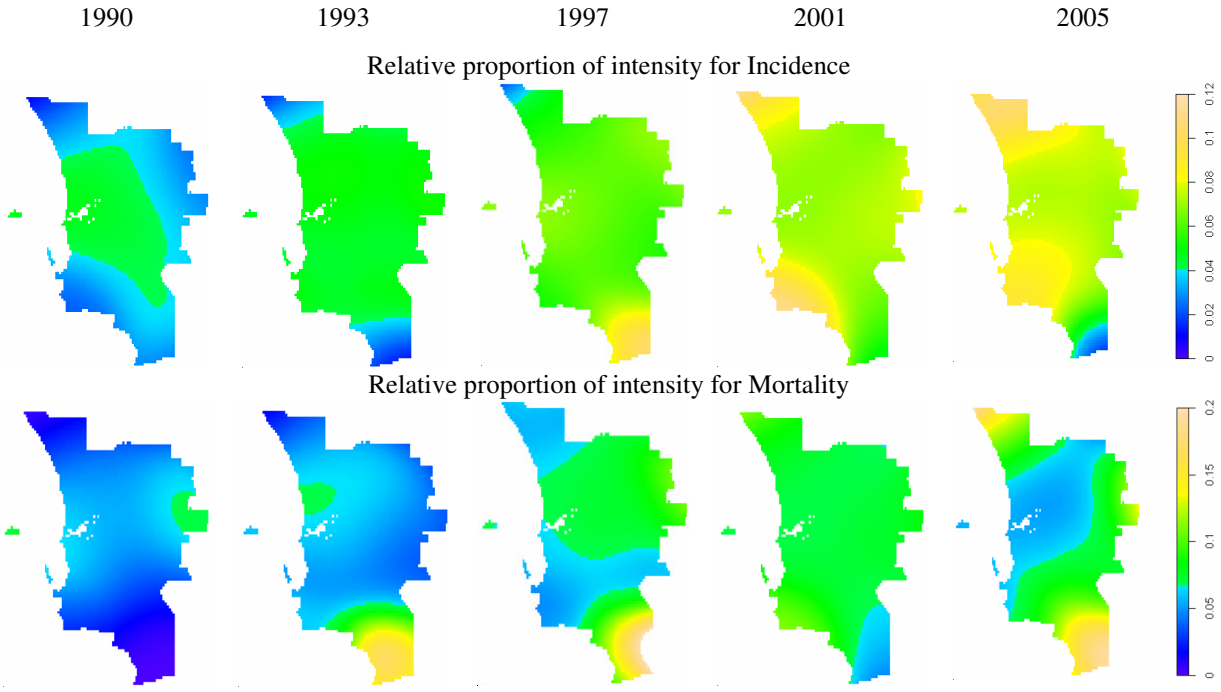


Figure 87 The relative proportion of intensity of 16-year period for breast cancer incidence (top) and mortality (bottom) in year 1990, 1993, 1997, 2001 and 2005.

6.4.2 Fitting models to 2005 breast incidence data

It has been found that the year 1992 point pattern is more similar to the year 2005 breast cancer incidence pattern than point patterns in other years (see Figure 88). Converting back to the form (2.39), the fitted model is:

$$\lambda_\theta(\mathbf{u}) = ((x-391.063)^2+(y+3536.509)^2)^{0.08} \times \exp\left(-1.618-0.942\,Z_1(\mathbf{u})+0.002\,Z_2(\mathbf{u})\right) \quad (6.10)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 1992 breast cancer incidence point pattern and $Z_2(\mathbf{u})$ represents the pixel image of female population density (see Figure 20).

The quantile-quantile plot, diagnostic plots and estimation map for the model (6.10) are shown in Figure 88. The quantile-quantile plot indicates that the agreement is tolerably good between the fitted model and the data. The smoothed residual field for the data has a lighter right-hand tail and lower variability than the smoothed residual field for simulations from the fitted model. The diagnostic plots illustrate that the trend is slightly overestimated

at which the lurking variable plots for the coordinates stay over the 2σ-limits. The estimation map indicates that the highest estimated incidence rate per 100 square kilometres in 2005 is about 210 people.

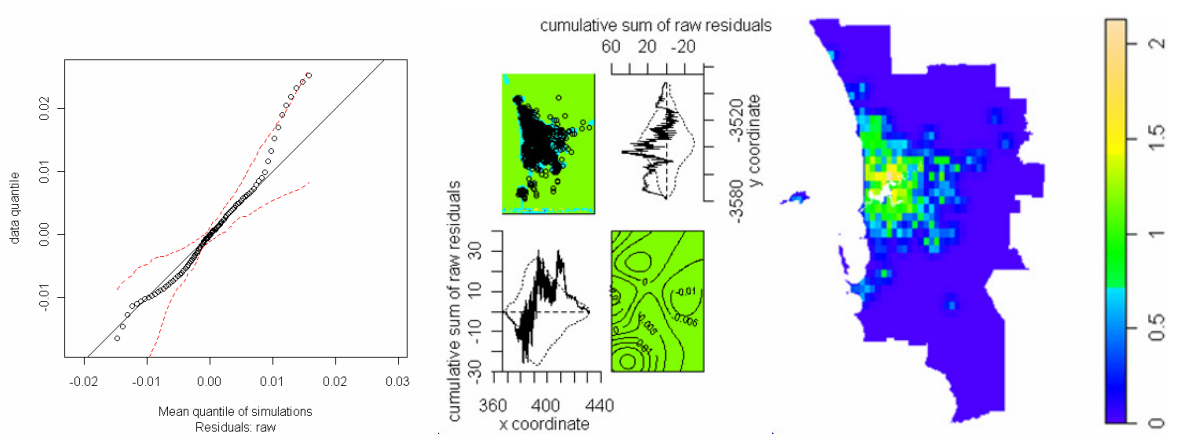


Figure 88 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.10) fitted to the year 2005 breast cancer incidence data.

6.4.3 Fitting models to 2005 breast mortality data

For the breast cancer patients survival rates (Figure 3), 58% of women diagnosed with breast cancer were alive for five years after diagnosis and about 40% were alive for at least six years. To model mortality data in 2005, susceptible population density in year 1999 is suitable for the fitting compared with other years in the period 1990-2005. When `popdensf` (see Figure 20 for female population density), `agedf` (see Figure 20 for the proportion of females aged 40 or over) and the `distmap` of point pattern (year 1999) work together as covariates to model mortality pattern, the fitting is adequate. The lurking variable plots in Figure 89 show that `agedf` accounts for data less than the other two covariates. If the `agedf` is not taken into consideration for modelling, the fit improves (see Figure 90). Converting back to the form (2.39), the fitted model with two covariates is:

$$\lambda_\theta(\mathbf{u}) = ((x-391.063)^2+(y+3536.509)^2)^{0.18} \times \exp(-4.236-1.089\,Z_1(\mathbf{u})+0.003\,Z_2(\mathbf{u})) \quad (6.11)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of breast cancer incidence point pattern (year 1999) and $Z_2(\mathbf{u})$ represents `popdensf` (see Figure 20).

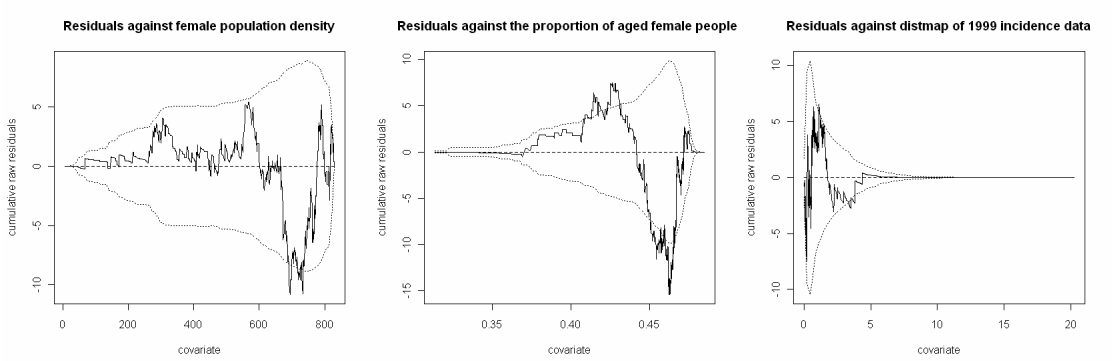


Figure 89 Lurking variable plots of 2005 breast cancer mortality for covariates: female population density (left), the proportion of the females aged 40 or over (middle) and the `distmap` of breast cancer incidence data (year 1999, right).

The fitted intensity function has a good capture of dependence on breast cancer incidence point pattern in 1999. Model (6.11) is the best model we have found which accounts for data adequately with the least number of parameters.

The quantile-quantile plot, diagnostic plots and estimation map for the model (6.11) are shown in Figure 90. The plots indicate a good agreement between model (6.11) and the data. No "extreme" outliers or distinct spatial trend is detected from four panel plots. The estimation map indicates that the highest estimated mortality rate per 100 square kilometres in 2005 is about 26 women in the centre of the study area.

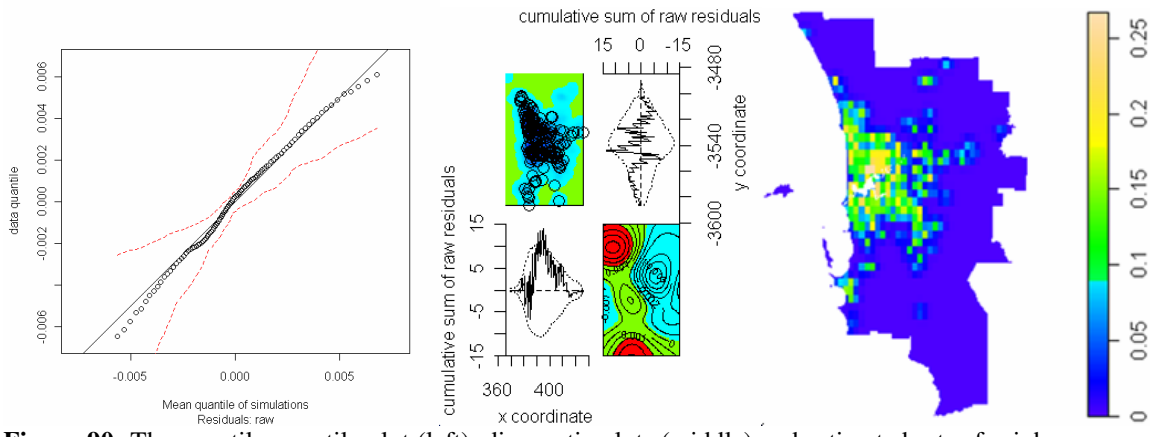


Figure 90 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.11) fitted to the year 2005 breast cancer mortality data.

6.5 Prostate cancer analysis

Prostate cancer only affects males and there is little risk to males under forty, so the available covariates are male population density (`Popdensm`) and the proportion of males

aged 40 above (`agedm`) shown in Figure 21, and susceptible population intensity (Figure 26).

6.5.1 Exploratory data analysis

The estimated inhomogeneous *K*-function plots for prostate cancer incidence and mortality data (Figure 91) are shown in Figure 92 . These two plots all indicate no obvious attraction or inhibition for the year 2005 prostate cancer data, so the point process is an inhomogeneous Poisson process.

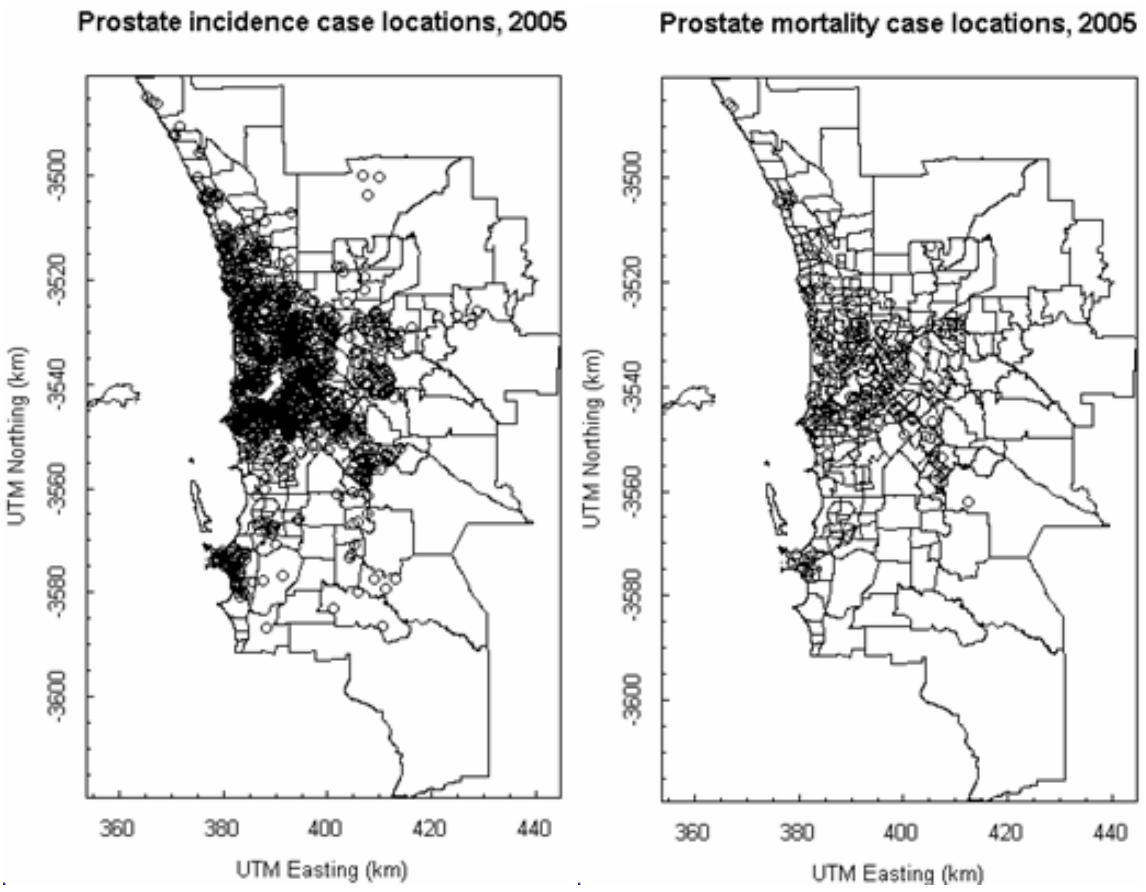
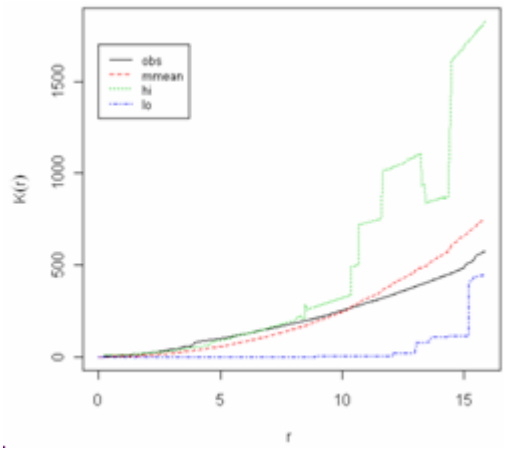


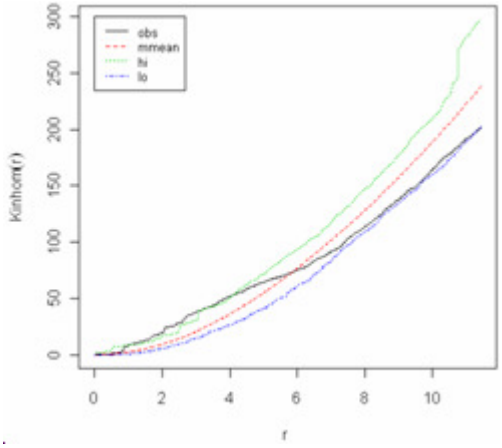
Figure 91 Prostate cancer incidence (left) and mortality (right) case locations in year 2005 in Perth.

Incidence



Bandwith for smoothing intensity : 5
Bandwith for simulation from data: 5

Mortality



Bandwith for smoothing intensity: 3.5
Bandwith for simulation from data: 7

Figure 92 Estimates of the inhomogeneous K-function for the year 2005 prostate cancer incidence and mortality data.

For prostate cancer incidence and mortality, the relative proportions of intensity of 16-year period in year 1990, 1993, 1997, 2001 and 2005 are shown in Figure 93. There is a distinct difference between year 2005 and other years except year 1997 for incidence. Compared with the total number of cases in the period 1990-2005, the proportion of incidence cases is relatively high in the southern part of the study area. However, the proportion in 2005 is much higher across the study area than other years. For mortality, the relative proportion of intensity is relatively high in the edge of the study area than inside except year 1990 based on the fill color of of the legend.

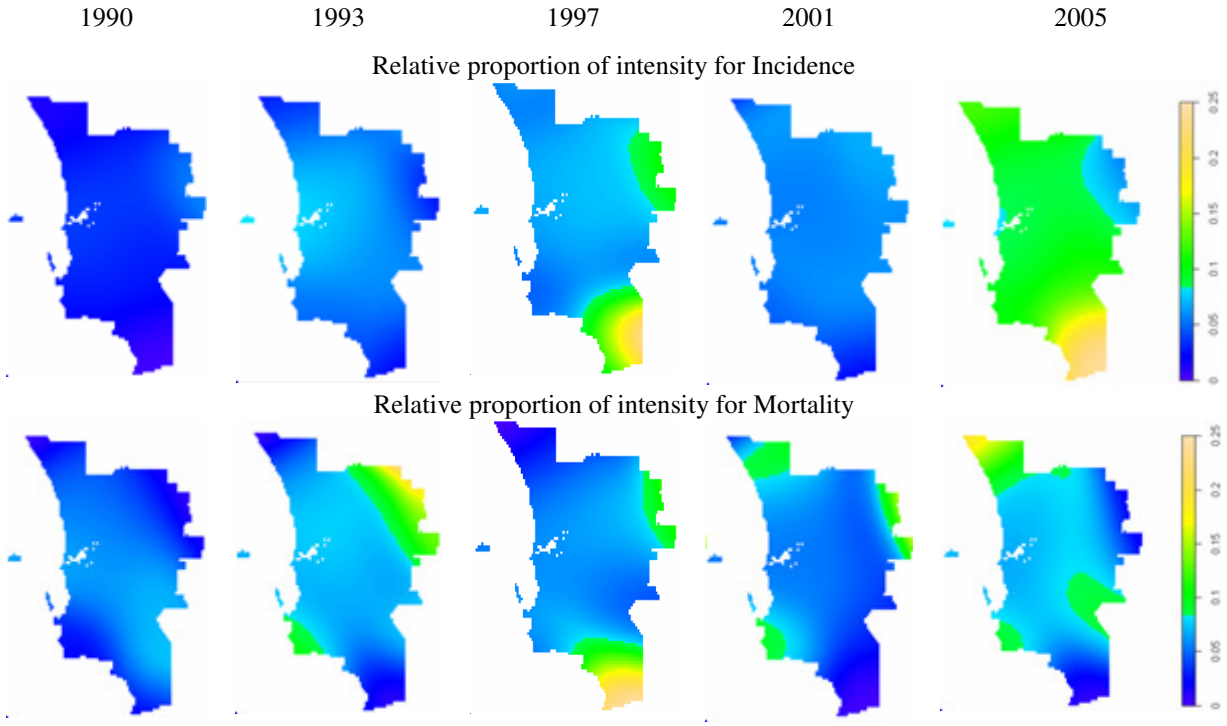


Figure 93 The relative proportion of intensity of 16-year period for prostate cancer incidence (top) and mortality (bottom) in year 1990, 1993, 1997, 2001 and 2005.

6.5.2 Prostate cancer incidence 2005 model fitting

We have shown that prostate cancer incidence point pattern in 2005 is an inhomogeneous Poisson process in Section 6.5.1. There are no appropriate covariates available to model the 2005 prostate cancer incidence point pattern. This indicates that the incidence point pattern in 2005 is not similar to the point pattern in other years.

6.5.3 Prostate cancer mortality 2005 model fitting

The survival rates (Figure 3) for prostate cancer patients show that about 65% of people diagnosed with prostate cancer were alive for three years after diagnosis and 55% were alive for four years. The susceptible population density based on prostate cancer incidence, `popdensm` and `agedm` are used as covariates to fit the 2005 mortality data. Yearly quantile-quantile plots (Figure 94-Figure 97) based on these covariates indicate clearly how much each fitted function captures the dependence of intensity on different covariates. Also, the

plots show the process to select an appropriate model. Three models are in three columns and all models involve spatial trends and susceptible population intensity (`distmap`) based on prostate cancer incidence. In addition, `popdensm` is incorporated in the first model (first column) and `agedm` is included in the second model (middle column). Susceptible population density (based on 2002 prostate cancer incidence) and spatial trends are able to make an adequate fitting. Converting back to (2.39), the fitted model is shown below:

$$\lambda_{\theta}(\mathbf{u}) = ((x-391.063)^2 + (y+3536.509)^2)^{0.186} \times \exp(-0.035 - 1.906 \times Z_1(\mathbf{u})) \quad (6.12)$$

where $Z_1(\mathbf{u})$ denotes the `distmap` of 2002 prostate cancer incidence point pattern.

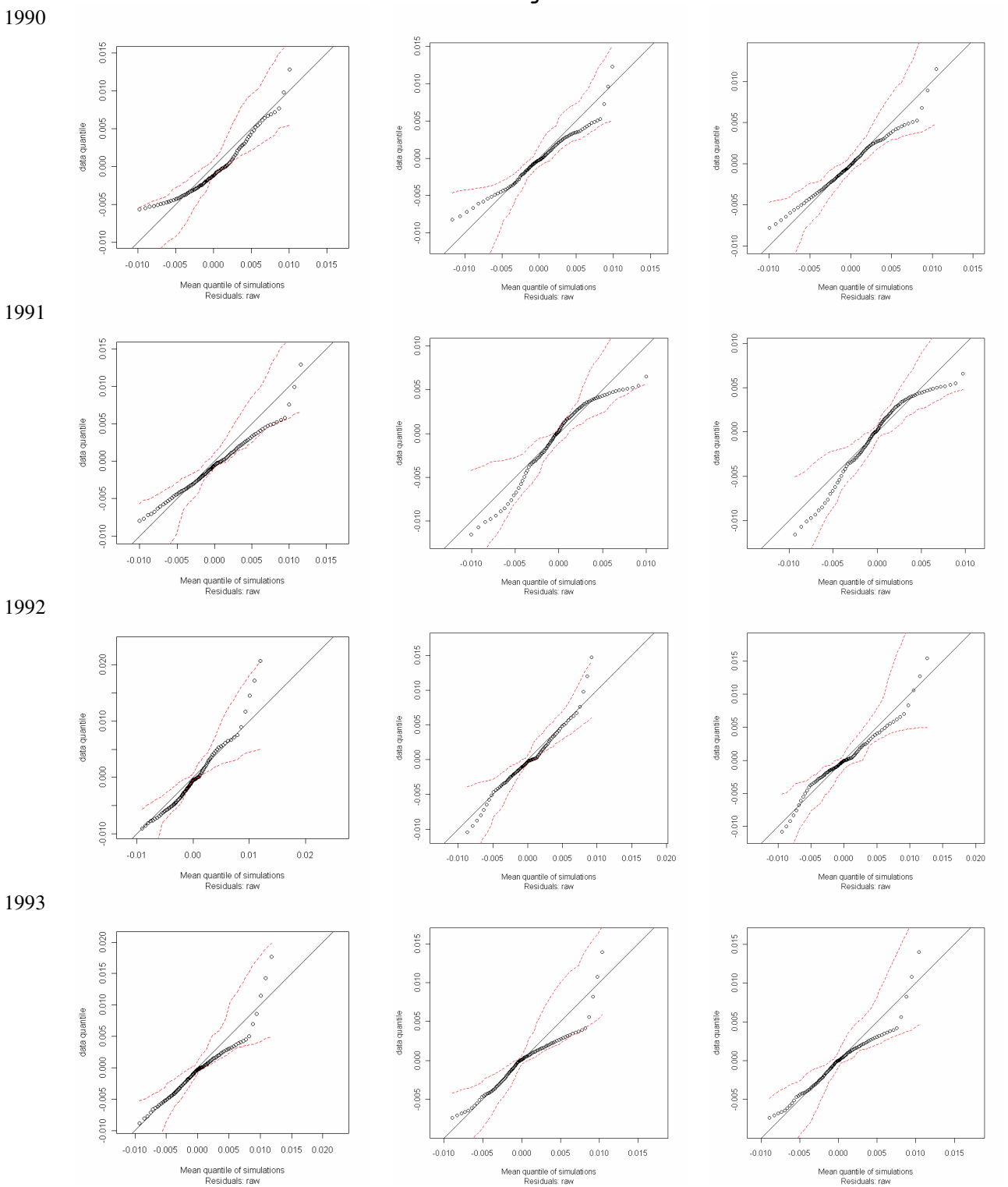


Figure 94 The quantile-quantile plots for the inhomogeneous Poisson models fitted to the year 2005 prostate cancer mortality data.

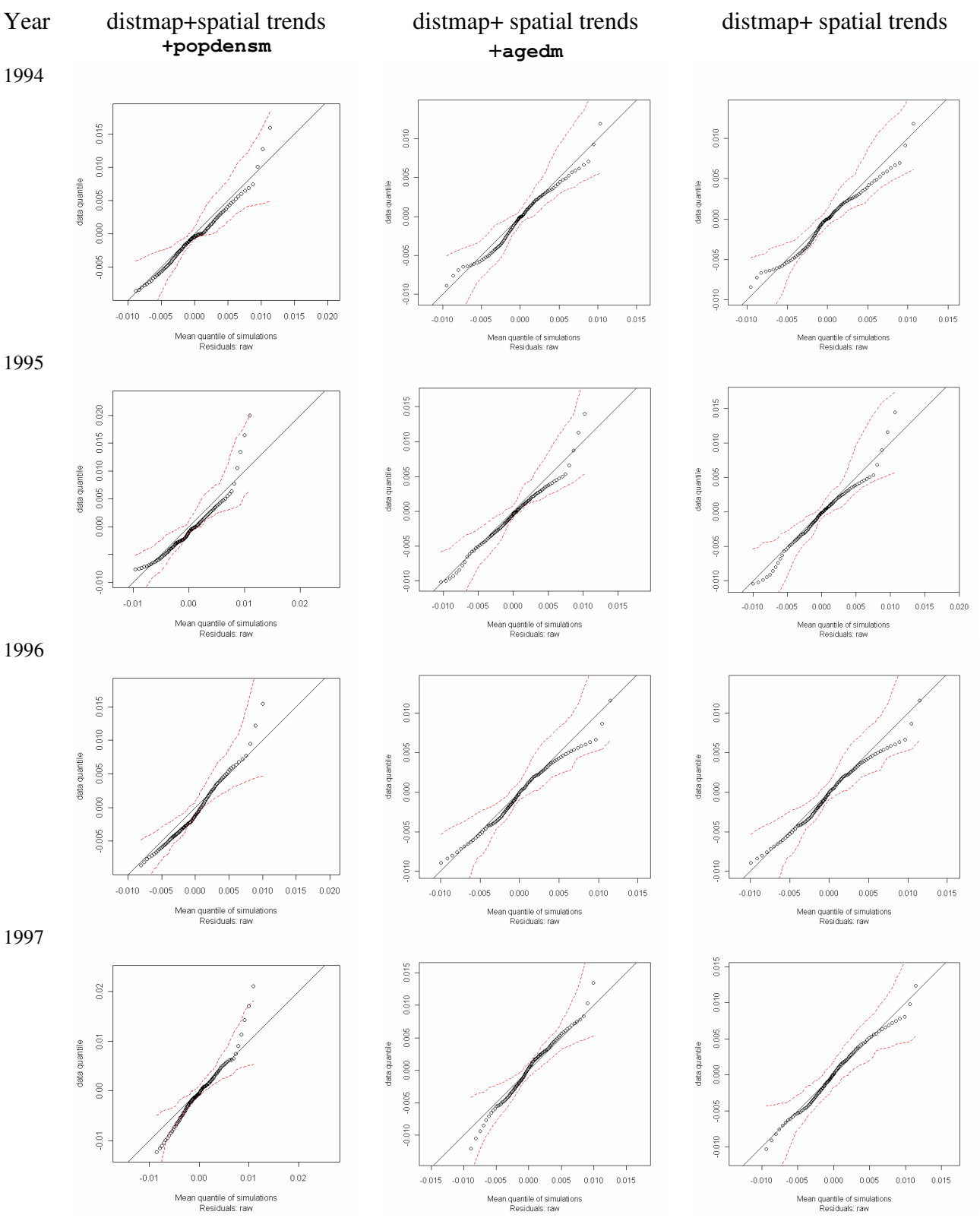


Figure 95 The quantile-quantile plots for the inhomogeneous Poisson models fitted to the year 2005 prostate cancer mortality data.

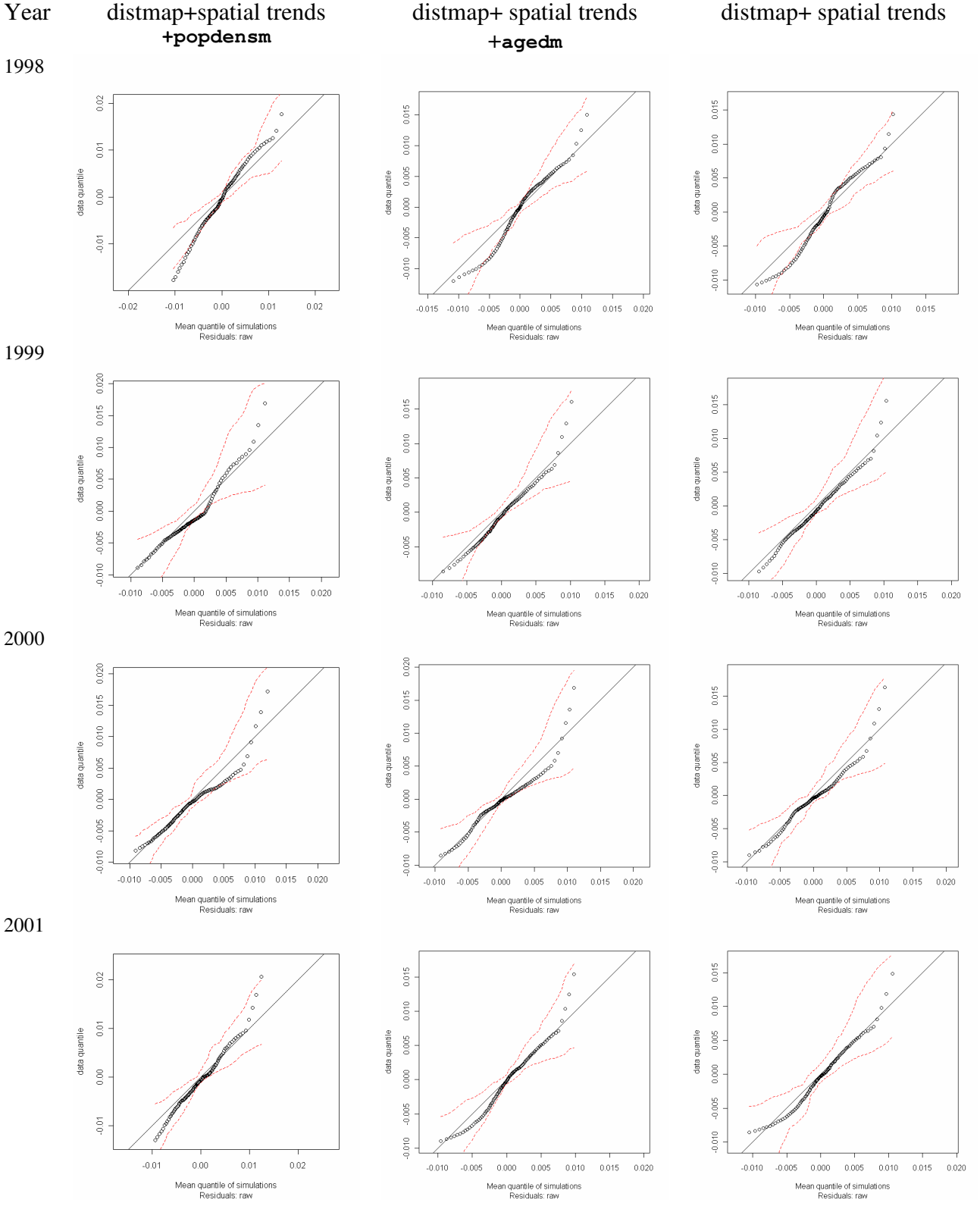


Figure 96 The quantile-quantile plots for the inhomogeneous Poisson models fitted to the year 2005 prostate cancer mortality data.

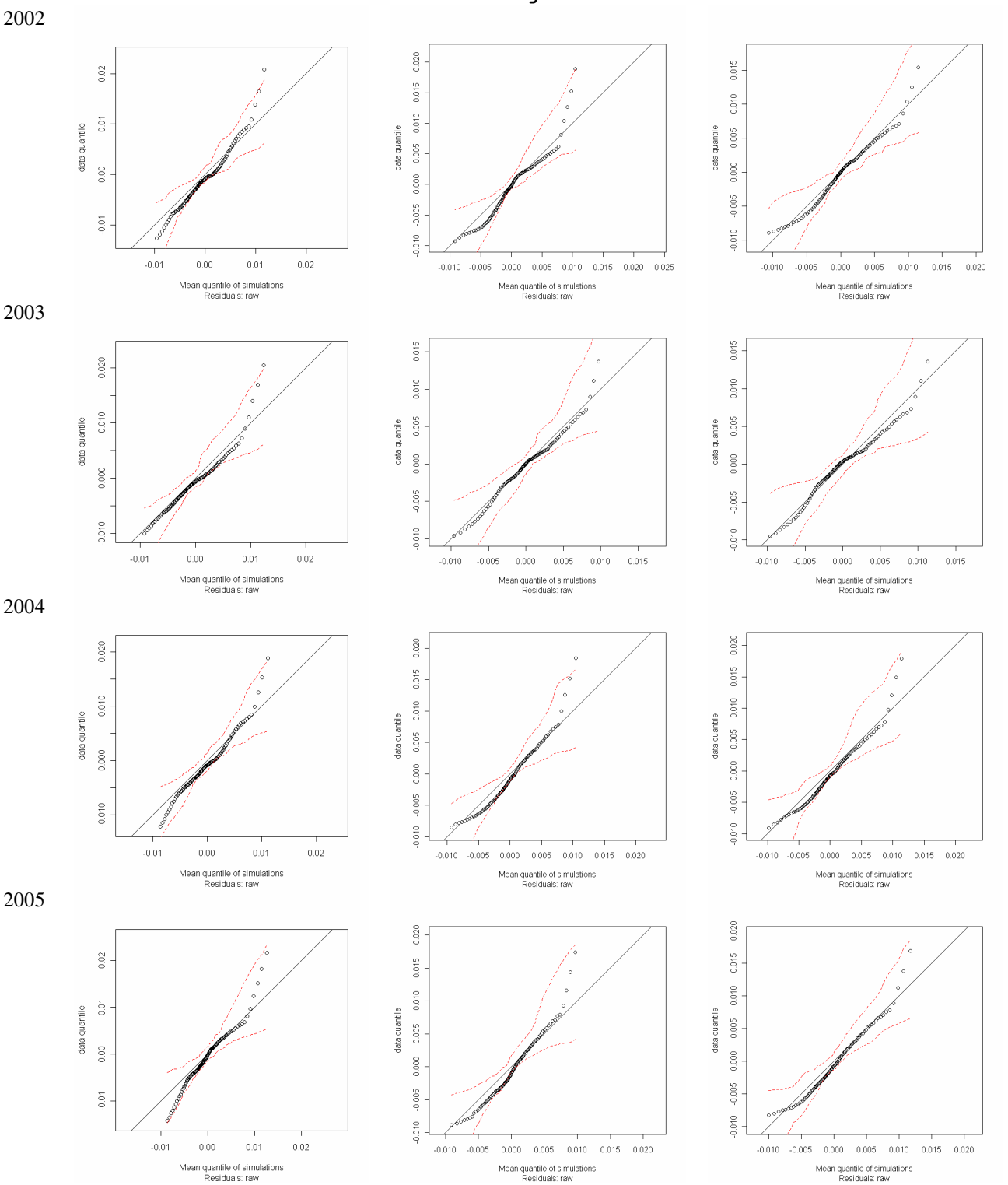


Figure 97 The quantile-quantile plots for the inhomogeneous Poisson models fitted to the year 2005 prostate cancer mortality data

The quantile-quantile plots (Figure 94-Figure 97) indicate that the models in the middle column are good in some years and models in the third column look adequate in years such as year 2001, 2002 and 2003. Although the fit is good for year 1997 (third column), it will not be considered for the modelling because only about 25% of people diagnosed with prostate cancer were alive after diagnosis according to Figure 3. They all indicate that the fitted intensity function captures the dependence on the covariates. However, the models in the middle column involve one more covariate `agedm` compared with the third column. The aim in statistical modelling is to investigate an adequate model for data, and this model should also be simple. The lurking variable plot (left plot of Figure 98) indicates that the fitted intensity function (6.12) perfectly captures dependence on prostate cancer incidence point pattern. In fact, the lurking variable plots have been tested for the years in which the model is adequate (third column) and it has been found that only the intensity function captures more dependence on year 2002 incidence point pattern than other years such as year 2001 and 2003 (right plot of Figure 98).

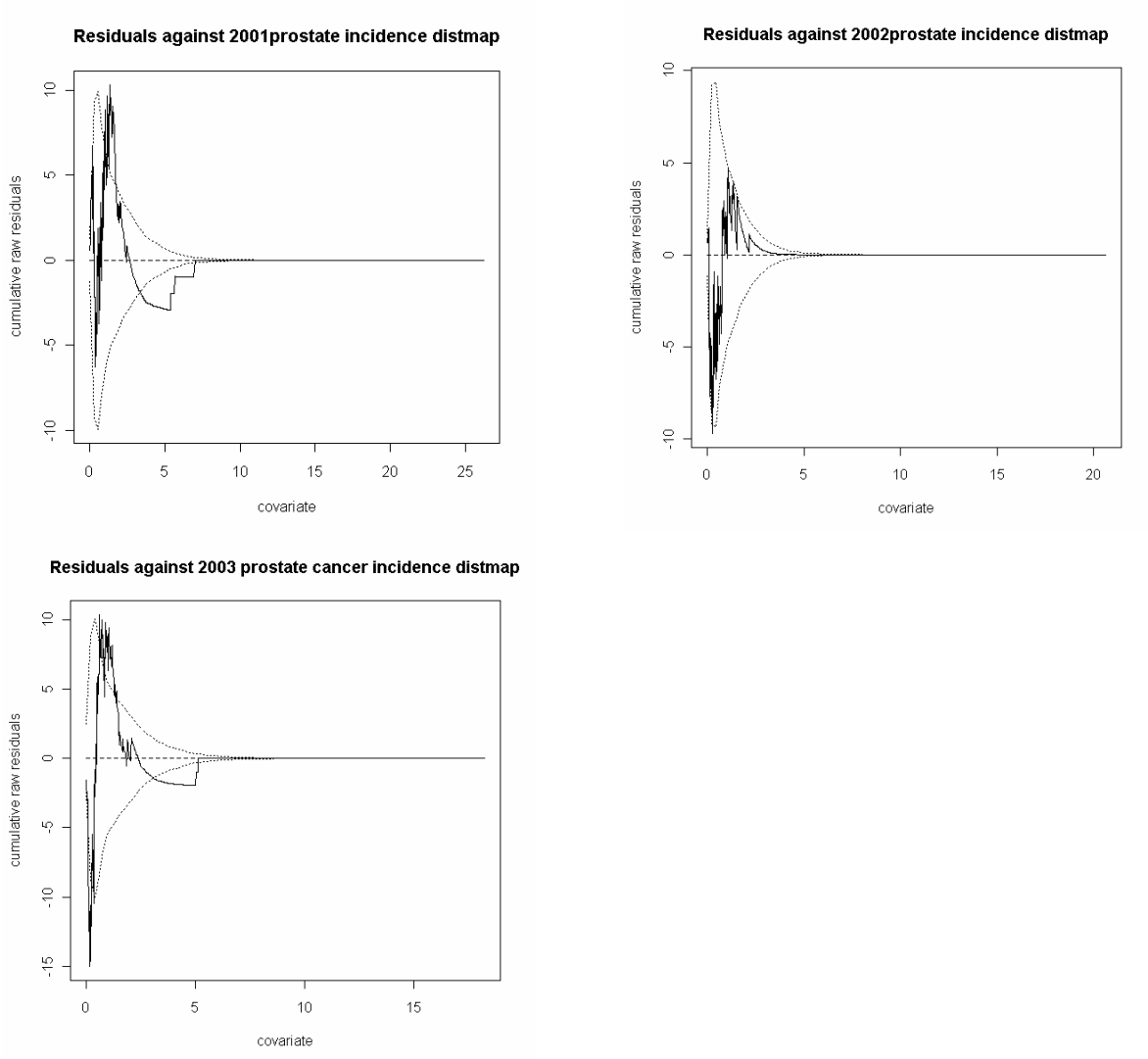


Figure 98 Lurking variable plot of 2005 prostate cancer mortality model fitting. The fitted intensity function involves transformations of coordinates and spatially susceptible population based on the `distmap` of 2001 (top left), 2002 (top right) and 2003 (bottom) prostate cancer incidence.

The quantile-quantile plot, diagnostic plots and estimation map for model (6.12) are shown in Figure 99. The Q-Q plot indicates an adequate agreement between the fitted model and mortality data. No "extreme" outliers or distinct spatial trend is detected from four panel plots. Also, nearly all the lurking variable plots for the coordinates stay within the 2σ-limits. The fitted model only involves covariates susceptible population density and spatial trends. The estimation map is based on model (6.12) and it shows the spatial distribution of cancer mortality rates. The estimation map indicates that the highest estimated mortality rate per 100 square kilometres in 2005 is about 41 men in the Perth metropolitan area.

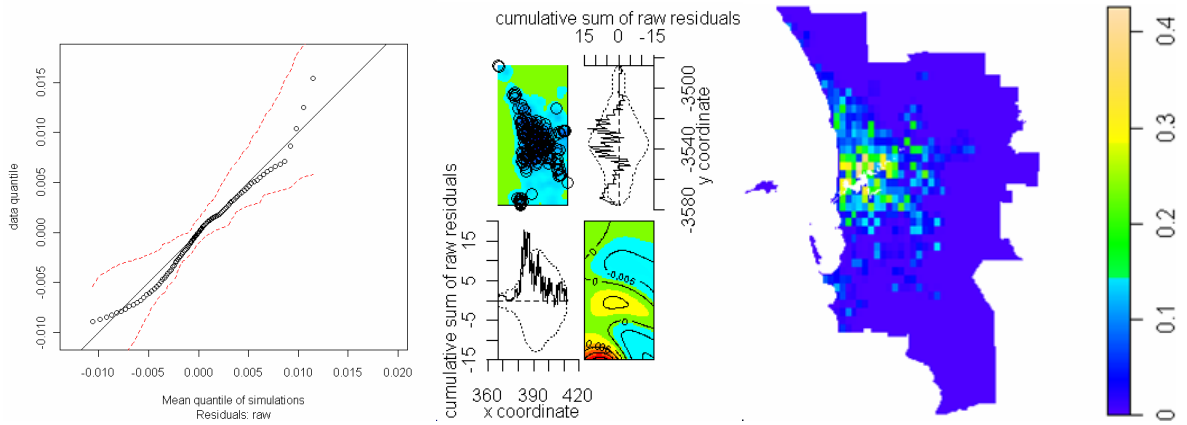


Figure 99 The quantile-quantile plot (left), diagnostic plots (middle) and estimated rates for inhomogeneous Poisson model (6.12) fitted to the year 2005 prostate cancer mortality data.

6.6 Summary

This chapter has provided the spatial point pattern analysis for cancers recorded in the Perth metropolitan region in 2005. We have shown that cancer incidence and mortality data in 2005 for five types of cancer are inhomogeneous Poisson processes using the inhomogeneous *K*-function. Based on the Poisson nature of the data, the Berman-Turner algorithm was used for the model fitting. All the covariates available have been investigated for modelling. However, no covariates have been found to account for 2005 prostate cancer incidence point pattern appropriately.

Two treatments of missing values have been applied to lung cancer mortality for model fitting and it has been illustrated that point patterns have influence on the spatial point pattern analysis. The fitting is slightly better when the data are treated using method 2 than method 1.

Perth CBD is selected as the key reference point as spatial trends for modelling. For lung cancer incidence, Kwinana industrial area is also used as reference point for modelling. It has been found that the distance from Kwinana is a better spatial trend for lung cancer incidence than the distance from Perth CBD. Kriging analysis has shown that melanoma cancer incidence is high in the western Perth metropolitan area. As a result, coastline is used as the reference line as well for modelling and we have found that the coastline accounts for melanoma incidence. Two models indicate that both of them are adequate to

account for the lung cancer and melanoma cancer incidence data respectively from different perspectives.

Quantile-quantile plots have been used to check whether the fitted model is acceptable and lurking variable plots have been applied to show whether the fitted function captures the dependence of intensity on the covariates. For lung cancer incidence analysis, it has been found that the point patterns in 1993 and 2005 are quite similar. A good fit is obtained when `popdens` is considered for the modelling. For mortality analysis, only one covariate is enough to get an adequate fit: the susceptible population density based on year 2003. Although the fitted intensity function captures the dependence of intensity both on the percentage of population above age 50 (`aged`) and the population density overall, it seems that `aged` contributes to lung cancer death more than population density.

For year 2005 melanoma incidence modelling, population density (`popdens`) accounts for more than the proportion of the people aged 30 or above (`aged30+`). This is probably due to the fact that melanoma cancer incidence affects any age group of people except children under the age of 10. The incidence point pattern in 2005 is quite similar to the point pattern in 2001 when population density and spatial trends are considered. For melanoma mortality model fitting, lurking variable plots indicate that the fitted intensity function captures dependence on susceptible population intensity. This covariate is enough to obtain a good fit for mortality data when spatial trend is considered.

For colorectal cancer analysis, we have shown that there are less covariates involved in fitting mortality data than fitting incidence data. For year 2005 colorectal incidence modelling, population density (`popdens`) accounts for data more than `aged`. This is probably due to the fact that colorectal cancer incidence affects any people aged 10 or over. Spatially susceptible population density and spatial trends are not enough to obtain a good fitting for incidence point pattern in 2005. For colorectal cancer mortality model fitting, susceptible population and spatial trends are enough to get a good fitting. The adequacy of the fit is confirmed by the least variability in the image of the smoothed residual field.

For breast cancer incidence analysis, it has been found that the point pattern in 2005 is quite similar to the point pattern in 1992. A good fit is obtained when popdens is considered for the modelling. For mortality analysis, female population density (popdensf) and spatially susceptible population intensity (1999) are able to acount for the point pattern in order to get an adequate fit. Although the fitted intensity function captures the dependence of intensity on popdensf and agedf, it seems that popdensf contributes to breast cancer death more than agedf(proportion of females aged 40 above).

For prostate cancer mortality, it has been indicated that the proportion of people aged 40+ (agedm) is a better covariate than male population density (popdensm) in most cases. These two covariates only improve the fitting when working with susceptible population intensity. The spatial trends (transformations of coordinates) and susceptible population intensity are enough to make an appropriate fitting. The estimated mortality number of prostate cancer is about 43 per 100 square kilometres in Perth 2005. Nearly all prostate cancer death is linked to prostate cancer incidence because the fitted model has been verified by the perfect capture of dependence on susceptible population intensity through lurking variable plots.

In conclusion, incidence modelling is more complex than mortality modelling. For mortality, it is often sufficient to use the susceptible population intensity in addition to the spatial trends. In contrast, for incidence modelling, the population density (or the proportion of selected age groups) is also required. Moreover, the agreement between fitted model and mortality data is usually better than that between fitted model and incidence data.

This chapter has provided the spatial point pattern analysis for cancer incidence and mortality data in Perth and the correlation between each type of cancer and covariates has been investigated. These findings will be informative to health service on the spatial distribution of cancer risks for each individual cancer incidence and mortality.

7 CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

7.1 Conclusions

In this study, geostatistical model has been applied to cancer incidence and motality data including lung, melanoma, colorectal, breast and prostate cancers. Spatial point pattern model has been applied to these data from a different perspective. The innovative application of these two models to cancer data in this study has presented the geostatistical analysis at SLA scale and a point pattern analysis at finer scale.

ATP Poisson kriging analysis takes into account the size and shape of administrative units (SLAs) and population within each SLA. This analysis is based on the adjusted rates to remove bias from the population structure and the rates also need to be corrected for the small number problem through the assigned weights which account for variability resulting from the population size. ATP Poisson kriging analysis is on the spatial distribution of cancer rates per 100,000 person-years for Perth SLAs based on the adjusted rates during 16-year period. To avoid the use of temporal structure in the model setup, the study period was divided into three contiguous periods to see how cancer mortality varies over time. To ensure the accuracy of the estimates, a more refined analysis for each individual year cannot be conducted in this study since there are not enough data available for the estimation. For breast and prostate cancer incidence, the spatial distribution in the three contiguous periods was explored. For non-sex specific cancers (lung, colorectal and melanoma cancer incidence), we have provided an analysis by sex and investigate how these cancers affect males and females differently.

Spatial point pattern analysis is based on spatial trends (through the selection of reference locations) and susceptible population intensity. Population density or the proportion of the selected age groups is also considered in the analysis. The covariates used in the spatial point pattern analysis are interpolated at every location in the study region. Spatial point pattern theory focuses on the analysis of the incidence (mortality) rate as the number of people per square kilometres in the Perth metropolitan area in year 2005. Environmental

covariates were also considered for lung and melanoma cancer incidence analysis. Two models indicate that lung cancer incidence risk is relatively high in Kwinana and melanoma incidence risk is relatively high along the coast.

For ATP Poisson kriging analysis, the estimates are not very accurate when there are not enough data available and the number of cases varies greatly with time (prostate cancer mortality in 2001-2005). For spatial point pattern analysis, the diagnostic plots can indicate the departure from the trend in the fitted model but it is often hard to interpret.

Conclusions from geostatistical analysis

In this study, ATP Poisson kriging was used to analyse incidence and mortality rates of the top five cancers in Perth by means of age-sex-adjustment. Geostatistical analysis of non-sex specific cancers is based on adjusted rates by taking into account age and sex. For sex specific cancers, the analysis only incorporates age. Spatial continuity and variability of adjusted rates have been investigated based on omnidirectional semivariogram plots because they are well-behaved and show good structure compared with directional semivariograms. Due to the distinct difference in the population density between SLAs in Perth, all kriging estimates have been conducted at the SLA level rather than a point-support semivariogram model within this study.

Breast cancer and prostate cancer are the most common cancers in women and men respectively. After these two cancers, males are more susceptible to lung cancer than females. Males also have slightly higher melanoma and colorectal cancer risk than females.

There are differences in the estimates between age-adjusted rates and age-sex-adjusted rates. Overall the risk from age and sex is possibly more continuous than the risk from age only but spatial correlation is slightly stronger for age-adjusted rates across Perth SLAs.

Spatial correlation is relatively strong for mortality compared with the corresponding incidence across Perth SLAs. For cancer incidence, the spatial correlation is relatively strong for colorectal cancer and weak for lung cancer. For mortality, the spatial correlation is strong for breast cancer but weak for prostate cancer.

Cancer mortality varies more continuously than the corresponding cancer incidence. For cancer incidence, prostate cancer shows the greatest variability and colorectal cancer varies relatively smoothly compared with other cancers. For mortality, lung cancer displays the greatest variability but melanoma cancer varies relatively smoothly compared with other cancers. The nugget effect is relatively low for melanoma and colorectal cancer mortality and this indicates that the cancer mortality case locations are probably close to each other compared with other cancers.

The mortality of most cancers has an increasing rate with time except prostate cancer. Age-adjusted rates and age-sex-adjusted rates have shown the difference in spatial variation. The values for the kriging variance are normally low in densely populated areas and high in sparsely populated areas. Higher rates arise mainly in SLAs in the vicinity of the Perth metropolitan area and the estimates have lower variance values compared with other areas.

Conclusions from spatial point pattern analysis

In this study, the Berman-Turner algorithm has been applied for model fitting and the intensity of the point pattern for cancers has been investigated. The model fitting is based on the inhomogeneous Poisson nature of the cancer data and this was verified via the inhomogeneous K-function.

The characteristics of the population distribution in the Perth metropolitan area (see Figure 19) play an important role in model fitting, and these spatial trends (a function of spatial coordinates using Perth city centre as the reference location) are indispensable to make an adequate fit. Except for the spatial trends, the fitted intensity function for the point process is also related to the spatial variability in susceptible population. The spatial intensity of the point patterns for different cancers is shown by the risk maps of estimated rates. The intensity is generally high in the centre of the study area which is densely populated. The fitted Poisson process model was validated by quantile-quantile plots and diagnostic plots.

Lung cancer data have been used as an example to illustrate how to treat missing values properly to get an improved model fit. The raw data are treated by assigning the mean of

16-year period cancer case locations in the SLA to those locations without latitude or longitude recorded.

Spatial point pattern analysis has shown that there are fewer covariates needed for fitting mortality data than fitting incidence data. In addition, the agreement between the fitted model and the mortality data is generally better than that between incidence data and the fitted model . Covariates available in this study were the population density, the distribution of the proportion of selected age groups and the susceptible population intensity. The covariates vary between cancers. Spatially susceptible population (based on cancer incidence) and spatial trends are not enough to obtain a good fit for incidence point patterns. When they are combined with population density, a good fit is obtained. This confirms a fact that cancer incidence in a specific year has little influence on incidence in another year. The incidence point pattern in 2005 is quite similar to the point pattern in another individual year when population density is included. A good fit is obtained for lung cancer incidence data when population density is considered for the modelling. For melanoma incidence modelling, population density is a better covariate than the proportion of the people aged 30 or above. For colorectal incidence modelling, population density accounts for data more than the proportion of the selected age groups, and spatially susceptible population density and spatial trends are not enough to obtain a good fit. When population density is considered for breast cancer incidence modelling an adequate fit is obtained. However, no covariates have been found to account for prostate cancer incidence point pattern appropriately.

For mortality data fitting, one covariate (susceptible population intensity) is enough to achieve an adequate fit in most cases. This probably reflects a fact that cancer mortality is mainly due to the corresponding cancer incidence. In fact, nearly all the cancer patients died of the corresponding cancer rather than other causes according to mortality data provided by the Cancer Registry, WA Department of Health. For lung cancer mortality modelling, it seems that the proportion of the selected age groups contributes to lung cancer death more than population density. In fact, one covariate (the susceptible population density) is enough to get an adequate fit. For melanoma and colorectal mortality model fitting, susceptible population intensity is enough to obtain a good fit when spatial trend is considered. For breast cancer mortality analysis, female population density and spatially

susceptible population intensity are able to account for the point pattern to get an adequate fit. For prostate cancer mortality, the spatial trends and susceptible population intensity are enough to make an appropriate fitting.

In most cases population density is a more reasonable covariate than the percentage of selected age groups. However, prostate cancer and lung cancer are an exception. This is mainly because these cancer incidence risks are influenced greatly by the age compared with other cancers.

7.2 Recommendations for future work

There are some limitations of this study, and so further research is recommended. A number of areas are worthy of further investigation such as the link between estimates and possibly sources of the cancer. For example, smoking patterns by SLAs for males and females will be recommended for lung cancer analysis in the future work if such data are available. The smoking patterns at finer scales such as suburb are recommended in spatial point pattern analysis in the future work.

The data modelled in the present spatial point pattern study are for the year 2005. At this stage, the population data at suburb scale is only available for year 2006. It is recommended that modelling cancer data in other years should be studied in future work if the population data by suburb are available for other census years.

The present spatial point pattern study only focuses on modelling cancer data using covariates such as population density, the percentage of people at higher risk and the susceptible population intensity. It is recommended that the correlation between different cancers should be studied in the future work to see whether people with one cancer are possibly to develop another cancer as a secondary cancer.

Geostatistical modelling and spatial statistical modelling are restricted to two-dimensional space in the present study. Therefore further investigation of these methods to incorporate time as a third dimension in the model should be considered.

8 REFERENCES

ABS (2009). 2009 Australia Bureau of Statistics estimated population.

Ali, M., P. Goovaerts, et al. (2006). "Application of poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of Bangladesh." International Journal of Health Geographics **5**: 45.

Baddeley, A. (2008). Analysing spatial point patterns in R.

Baddeley, A., J. Moller, et al. (2000). "Non- and semi-parametric estimation of interaction in inhomogeneous point patterns." Statistical Neerlandica **54**(3): 329-350.

Baddeley, A. and R. Turner (2000). "Practical maximum pseudolikelihood for spatial point patterns (with discussion)." Australian and New Zealand Journal of Statistics **42**(3): 283-322.

Baddeley, A. and R. Turner (2005). "An R package for analyzing spatial point patterns." Journal of statistical software **12**(6): 1-42.

Baddeley, A., R. Turner, et al. (2005). "Residual analysis for spatial point processes." Royal Statistical Society **67**: 617-666.

Berman, M. and R. Turner (1992). "Approximating point process likelihoods with GLIM." Applied Statistics **41**(1): 31-38.

Box, G. E. P. (1966). "Use and abuse of regression." Technometrics **8**(4): 625-629.

Christakos, G. and J. Lai (1997). "A study of the breast cancer dynamics in North Carolina." Soc.Sci.Med **45**(10): 1503-1517.

CRAN (2009). Consolidated R Archive Network, http://www.r-project.org/.

CRUK Available at:http://info.cancerresearchuk.org/healthyliving/harmfulsubstances/airpollutionandradon/.

Diggle, P. J. (1983). Statistical analysis of point processes. New York, Chapman and Hall.

Diggle, P. J. (1990). "A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point." Journal of the Royal Statistical Society, Series A **153**(3): 349-362.

Diggle, P. J. (2003). Statistical analysis of spatial point patterns, London: Arnold.

Gatrell, A. C., T. C. Bailey, et al. (1996). "Spatial point pattern analysis and its application in geographical epidemiology." Royal Geographical Society.

Goovaerts, P. (1997). Geostatistics for Natural Resources Evalution, Oxford University Press.

Goovaerts, P. (2005a). Analysis and detection of health disparities using geostatistics and a space-time information system. The case of prostate cancer mortality in the United States, 1970-1994. In Proceedings of GIS planet 2005 Congress. Paper available at http://home.comcast.net/~goovaerts/Paper148PierreGoovaerts.pdf.

Goovaerts, P. (2005b). "Detection of spatial clusters and outliers in cancer rates using geostatistics filters and spatial neutral models." Geostatistics for Environment Applicaion: 149-160.

Goovaerts, P. (2005c). "Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using poisson kriging." International Journal of Health Geographics **4**: 31.

Goovaerts, P. (2005d). "Simulation-based assessment of a geostatistical approach for estimation and mapping of the risk of cancer." Geostatistics Banff 2004 **2**: 787-796.

Goovaerts, P. (2006a). "Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point poisson kriging." International Journal of Health Geographics **5**: 52.

Goovaerts, P. (2006b). "Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using poisson kriging and p-field simulation." International Journal of Health Geographics **5**: 7.

Goovaerts, P. (2008). "Kriging and semivarioram deconvolution in the presence of irregular geographical units." Math Geosci **40**: 101-128.

Goovaerts, P. and G. Jacquez (2004). "Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York." International Journal of Health Geographics **3**: 14.

Goovaerts, P., G. Jacquez, et al. (2005). "Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms." Geographical Analysis **37**: 152-182.

Kaiser, M. and N. Cressie (1997). "Modeling Poisson variables with positive spatial dependence " Statistical & Probability Letters **35**(4): 423-432.

Klein, R. and C. Schoenborn (2001). Age adjustment using the 2000 projected U.S. population.

Kyriakidis, P. (2004). "A geostatistical framework for area-to-point spatial interpolation " Geographical Analysis **36**(3).

Lai, D. (2004). "Geostatistical analysis of Chinese cancer mortality: variogram, kriging and beyond." Journal of Data Science **2**: 177-193.

Marcon, E. and F. Puech (2003). "Generalizing Ripley's *K* function to inhomogeneous populations."

Monestiez, P., L. Dubroca, et al. (2004). "Comparison of model based geostatistical methods in ecology: application to fin whales spatial distribution in northwestern Mediterranean Sea." Geostatistics Banff **2**: 777-786.

Monestiez, P., L. Dubroca, et al. (2006). "Geostatistical modelling of spatial distribution of Balaenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts." Ecological Modelling **193**: 615-628.

Oliver, M., R. Webster, et al. (1998). "Binomial cokriging for estimating and mapping the risk of childhood cancer." IMA Journal of Mathematics Applied in Medicine and Biology **15**: 279-297.

Papangelou, F. (1974). On the Palm probabilities of processes of points and processes of lines. Stochastic Geometry. E.F.Harding and D.G.Kendall, London: Wiley**:** pp.114-147.

Podur, J., D. Martell, et al. (2003). "Spatial patterns of lightening-caused forest fires in Ontario, 1976-1998." Ecological Modelling **164**: 1-20.

Redfearn, J. C. B. (1948). "Traverse Mercator Formulae." Empire Survey Review **69**.

Ripley, B. D. (1977). "Modelling spatial patterns (with discussion)." Journal of the Royal Statistical Society **39**: 172-212.

Schabenberger, O. and C. A. Gotway (2005). Statistical Methods for Spatial Data Analysis, Chapman & Hall/CRC.

Senn, S. and M. Scott (2008). Statistical analysis and modelling of spatial point patterns, John Wiley & Sons, Ltd.

Shao, C., U. Mueller, et al. (2009). Area-to-point Poisson kriging analysis for lung cancer incidence in Perth areas. 18th World IMACS/MODSIM Congress. Cairns, Australia.

Shibuya, K., M. Inoue, et al. (2005). "Statistical modeling and projections of lung cancer mortality in 4 industrialized countries." Int. J. Cancer **117**: 476-485.
Sparks, L., J. Blanch, et al. (2011). "Estimating ionospheric delay using kriging." Radio Science **46**.
TCCWA (2007). The Cancer Council Western Australia (2007). Statistics: cancer risk. Available at: http://www.cancerwa.asn.au/resources/statistics/#wa.
Threlfall, T. and J. Thompson (2004). Threlfall TJ, Thompson JR. 2004. Cancer incidence and mortality in Western Australia, 2002. Department of Health, Western Australia, Perth. Statistical series number 71.
Threlfall, T. and J. Thompson (2007). Threlfall TJ, Thompson JR. 2007. Cancer incidence and mortality in Western Australia, 2005. Department of Health, Western Australia, Perth. Statistical series number 81. Available at: http://www.health.wa.gov.au/wacr/downloads/rep05all.pdf.
Turner, R. (2009). "Point patterns of forest fire locations." Environ Ecol Stat **16**: 197-223.
Waagepetersen, R. (2008). "Estimating functions for inhomogeneous spatial point processes with incomplete covariate data." Biometrika **95**(2): 351-363.
WACR (1997). Cancer incidence and mortality in Western Australia, 1997.
WACR (2008). Western Australian Cancer Registry (2008). Common cancers by type. Available at: http://www.health.wa.gov.au/wacr/statistics/stats_quick_type.cfm.
Waller, L. A. and C. A. Gotway (2004). Applied Spatial Statistics for Public Health data, A John Wiley & Sons, INC., Publication.
WHO (2003). World Health Organization (2003) Global cancer rates could increase by 50% to 15million by 2020. Available at: http://www.who.int/mediacentre/news/releases/2003/pr27/en/.
WHO (2007). World Health Organization (2007) Cancer scales turn against children as obesity problem grows. Available at: http://www.uicc.org/index2.php?option=com_content&do_pdf=1&id=15769.
WHO (2011). World Health Organization (2011) Cancer . Available at: http://www.who.int/mediacentre/factsheets/fs297/en/

9 APPENDICES

9.1 Appendix A-Crude rates of cancers for geostatistical analysis

In this section, the crude rates for the five types of cancer are provided for the reference. The crude rates for lung, melanoma and colorectal cancers are presented in Figure 100, Figure 101 and Figure 102. The crude rates for sex specific cancers are given in Figure 103 and Figure 104. Generally speaking, the maps for crude rates show slight more variability than those of adjusted rates which are discussed in details in section 4.6.1.

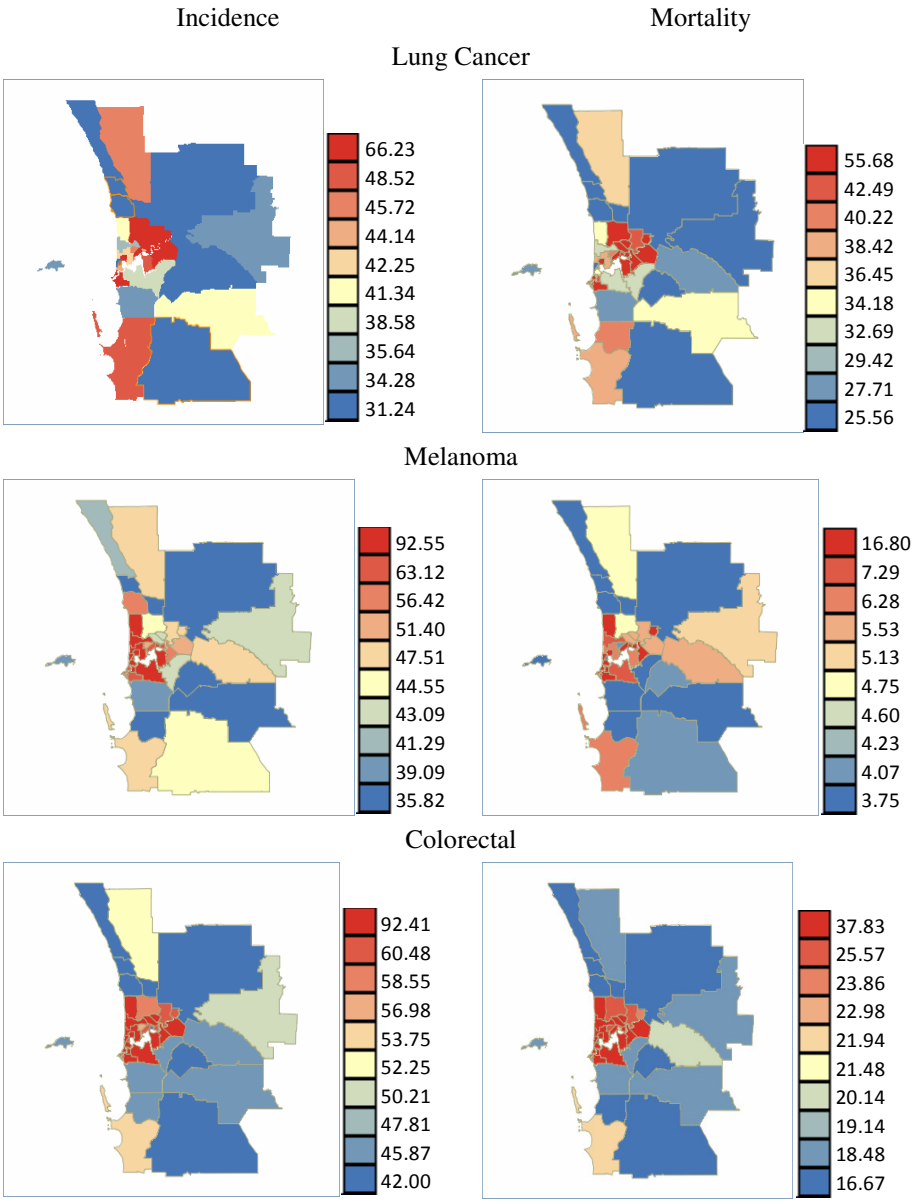


Figure 100 Crude rates for lung, melanoma and colorectal cancer.

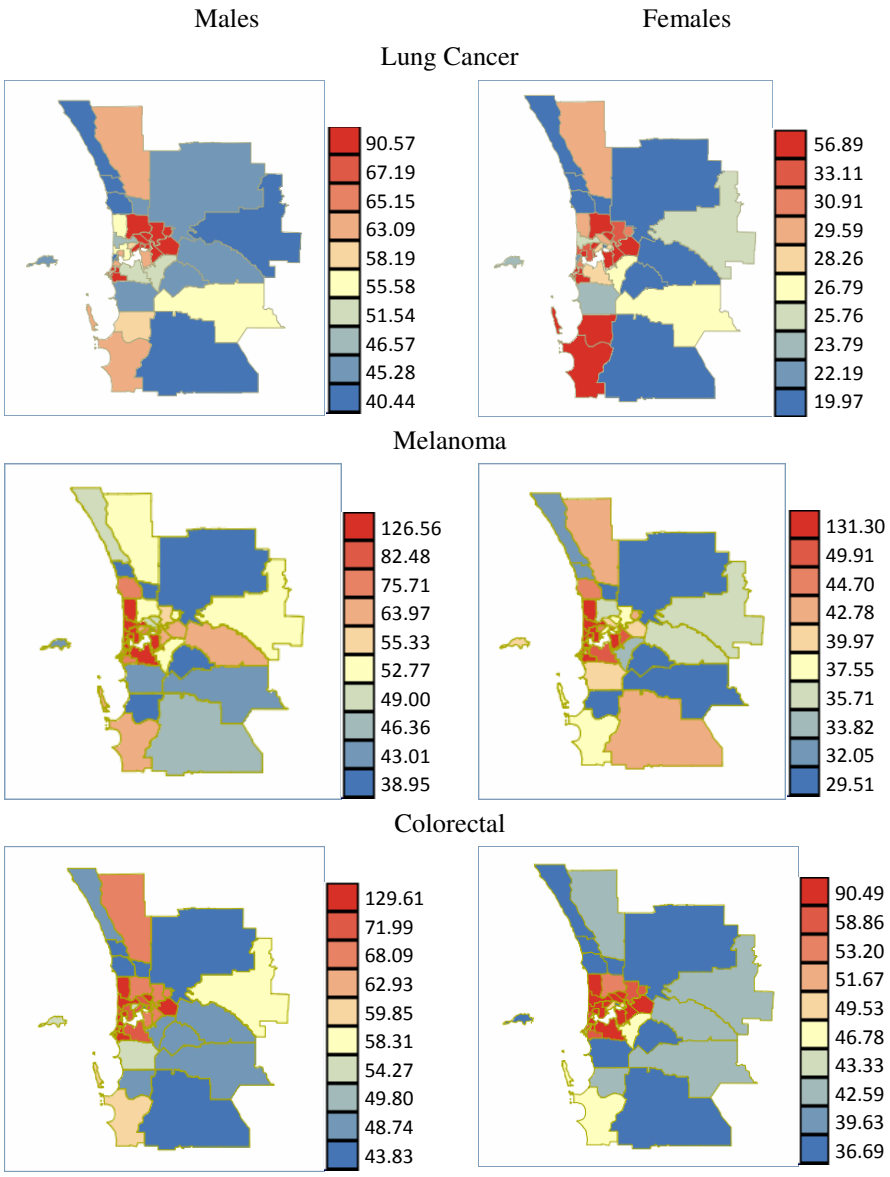


Figure 101 Crude rates by sex for lung, melanoma and colorectal cancer incidence.

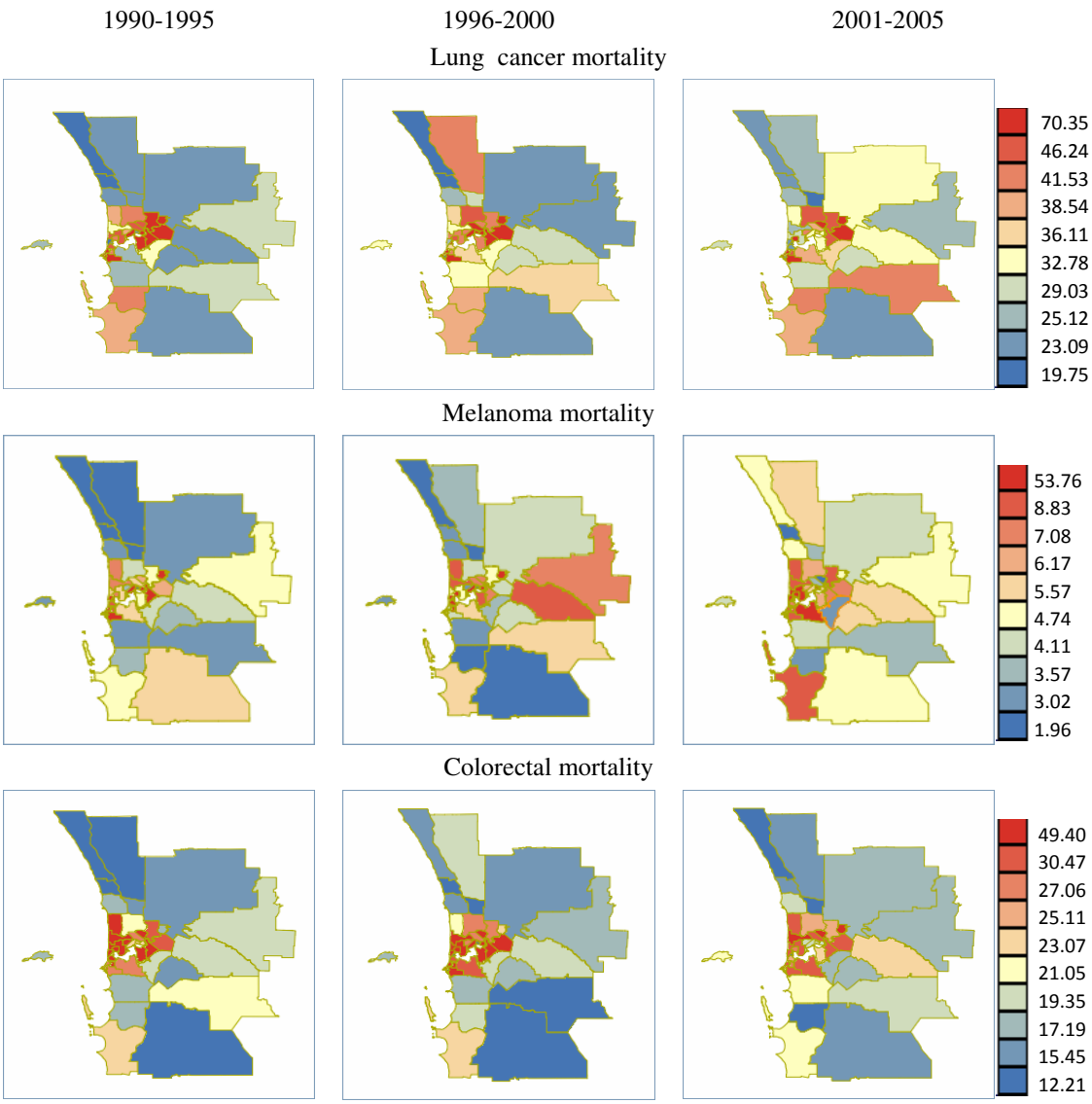


Figure 102 Crude rates for cancer mortality during three time periods.

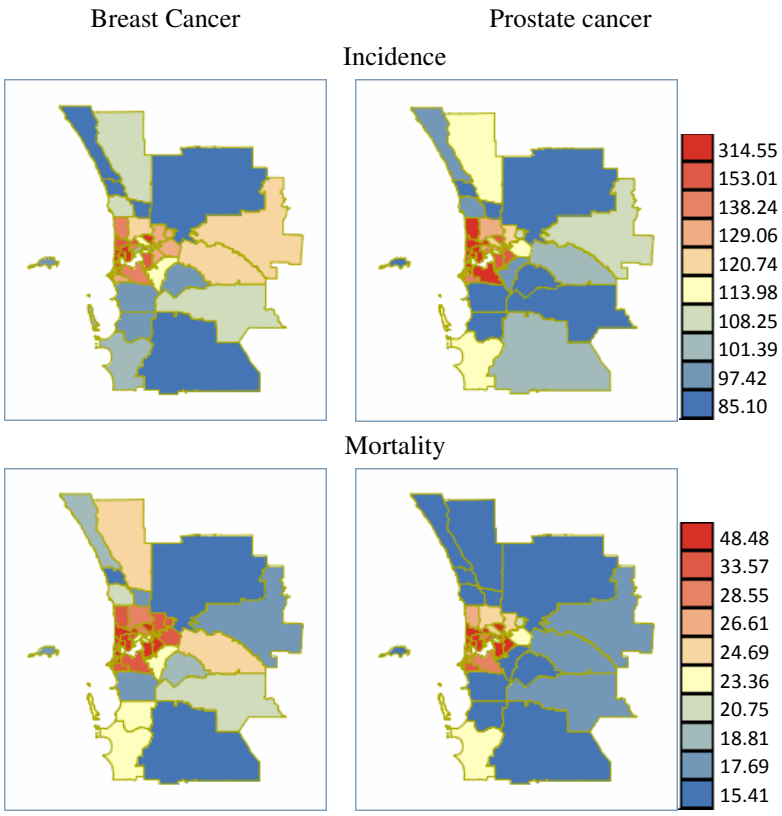


Figure 103 Crude rates for breast and prostate cancer.

Figure 104 Crude rates for breast and prostate cancer during three time periods.

9.2 Appendix B-Codes

9.2.1 Code to read "shapefiles" into R package

The commands of reading "shapefiles" into spatstat are:

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)
```

9.2.2 Code to produce empty space distances

Take lung cancer incidence in year 1990 as an example, the code to produce empty space distances is shown below:

```
esd = function(file="perth suburb06") {
        require(grDevices); require(maptools); require(spatstat); require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
        rutm = sapply(sutm,range)
        xlim=rutm[,"X"]; ylim=rutm[,"Y"]

        c90<-read.table("Lcoord90.txt",header=TRUE)
        attr(  c90,"projection")<-"LL"
        c90<-convUL ( c90, km=TRUE)
        c90.ppp = ppp(c90$X,c90$Y,window=owin(xlim,ylim))
        c90.im  = distmap(c90.ppp)

        bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
        if (!exists("land",envir=.GlobalEnv)) {
                land = joinPolys(sutm,operation="UNION")
```

```
            assign("land",land,pos=1) }
        water = joinPolys(bbox,land,"DIFF")

        plot(c90.im, main="1990",zlim=c(0,66))
        addPolys(water,col="white",border=FALSE)

 browser();return()
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,
        c90=c90,c90.ppp=c90.ppp,c90.im=c90.im,bbox=bbox,land=land,water=water))}
out=esd()
```

9.2.3 Code to produce population density image

The code shown below is used to produce the pixel image `popdens` (see Table 3 for the file 2006popsuburb.txt). This image is used as covariate for lung, melanoma and colorectal cancer model fitting.

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)

burbs<- read.table("2006popsuburb.txt", header=TRUE)
avepopden <- with(burbs, Total/Area)
burb <- with(burbs, ppp(X, Y, range(X),range(Y), marks=avepopden))

window<-as.owin(z)
burbs$window<-window

popdens <- smooth.ppp(burbs)
plot(popdens,main="popdens")
```

The following code is used to produce the pixel image `popdensf` (see Table 4 for the file 2006popsuburbsex.txt). This image is used for breast cancer model fitting.

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)

burbsf<- read.table("2006popsuburbsex.txt", header=TRUE)
avepopdenf<- with(burbsf, Fpop/Area)
burbsf <- with(burbsf, ppp(X, Y, range(X),range(Y), marks=avepopdenf))

window<-as.owin(z)
burbsf$window<-window

popdensf <- smooth.ppp(burbsf)
plot(popdensf)
```

The following code is used to produce the pixel image `popdensm` (see Table 4 for the file 2006popsuburbsex.txt). This code is usd for prostate cancer model fitting.

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)

burbsm<- read.table("2006popsuburbsex.txt", header=TRUE)
avepopdenm<- with(burbsm, Mpop/Area)
burbsm <- with(burbsm, ppp(X, Y, range(X),range(Y), marks=avepopdenm))

window<-as.owin(z)
```

```
burbsm$window<-window

popdensm <- smooth.ppp(burbsm)
plot(popdensm)
```

9.2.4 Code to produce the percentage of selected age group

The following code is used to produce the pixel image `aged` (see Table 3 for the file 2006popsuburb.txt) for lung and colorectal cancer model fitting.

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)

burbs<- read.table("2006popsuburb.txt", header=TRUE)
burbs<- with(burbs, ppp(X, Y, range(X),range(Y), marks=Fiftypp))

window<-as.owin(z)
burbs$window<-window

aged<- smooth.ppp(burbs)
plot(aged,main="aged")
```

The following code is used to produce the pixel image `aged30+` (see Table 3 for the file 2006popsuburb.txt) for melanoma cancer model fitting.

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)
```

```
burbs<- read.table("2006popsuburb.txt", header=TRUE)
burbs<- with(burbs, ppp(X, Y, range(X),range(Y), marks=Thirtypp))

window<-as.owin(z)
burbs$window<-window

aged<- smooth.ppp(burbs)
plot(aged,main="aged(30+)")
```

The following code is used to produce the pixel image `agedf` (see Table 4 for the file 2006popsuburbsex.txt) for breast cancer model fitting.

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)

burbsf<- read.table("2006popsuburbsex.txt", header=TRUE)
burbsf<- with(burbsf, ppp(X, Y, range(X),range(Y), marks= Ffortypp))

window<-as.owin(z)
burbsf$window<-window

agedf<- smooth.ppp(burbsf)
plot(agedf,main="agedf")
```

The following code is used to produce the pixel image `agedm40+` (see Table 4 for the file 2006popsuburbsex.txt) for prostate cancer model fitting.

```
x<-readShapeSpatial("perth suburb06.shp")
y <- as(x, "SpatialPolygons")
```

```
spatstat.options(checkpolygons = FALSE)
z <- as(y, "owin")
spatstat.options(checkpolygons = TRUE)

burbsm<- read.table("2006popsuburbsex.txt", header=TRUE)
burbsm<- with(burbsm, ppp(X, Y, range(X),range(Y), marks= Mfortypp))

window<-as.owin(z)
burbsm$window<-window

agedm<- smooth.ppp(burbsm)
plot(agedm,main="agedm (40+)")
```

9.2.5 Code to produce relative proportion of intensity

Take breast cancer incidence as an example, the `spatstat` commands to obtain the relative proportions of intensity in year 1990 are shown below:

```
X<-density(split(BIyr))
attach(X)
a0<-X$"1990"
a1<-X$"1991"
a2<-X$"1992"
a3<-X$"1993"
a4<-X$"1994"
a5<-X$"1995"
a6<-X$"1996"
a7<-X$"1997"
a8<-X$"1998"
a9<-X$"1999"
a10<-X$"2000"
a11<-X$"2001"
```

```
a12<-X$"2002"
a13<-X$"2003"
a14<-X$"2004"
a15<-X$"2005"
Relative90<-
eval.im(a0/(a0+a1+a2+a3+a4+a5+a6+a7+a8+a9+a10+a11+a12+a13+a14+a15))
plot(Relative90,main="Relative proportion of intensity 1990")
detach(X)
```

where BIyr is a file containing all breast cancer incidence cases during the period 1990-2005.

9.2.6 Code to produce estimated inhomogeneous K-plot

9.2.6.1 Lung cancer

The `spatstat` commands to get the estimated inhomogeneous *K*-plot for lung cancer incidence (Lcoord05.txt) and mortality (Mor2005L.txt) point patterns in 2005 are:

```
LI05<-read.table("Lcoord05.txt",header=TRUE)
attr(LI05,"projection")<-"LL"
LI05<-convUL (LI05, km=TRUE)
as.EventData(LI05, projection = NULL, zone = NULL)
LI05<-ppp(LI05$X,LI05$Y,window=owin(range(LI05$X),range(LI05$Y)))

smo<-density.ppp(LI05,sigma=7)
Ken<- envelope(LI05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=7, correction="border")
plot(Ken)

mb<-read.table("Mor2005L.txt",header=TRUE)
attr(mb,"projection")<-"LL"
mb<-convUL (mb, km=TRUE)
as.EventData(mb, projection = NULL, zone = NULL)
```

```
mb<-ppp(mb$X,mb$Y,window=owin(range(mb$X),range(mb$Y)))
smo<-density.ppp(mb,sigma=4.9)
Ken<- envelope(mb, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=4.9, correction="trans")
plot(Ken)
```

9.2.6.2 Melanoma cancer

The `spatstat` commands to get the estimated inhomogeneous *K*-plot for melanoma cancer incidence (Mcoord05.txt) and mortality (Mor2005M.txt) point patterns in 2005 are:

```
MI05<-read.table("Mcoord05.txt",header=TRUE)
attr(MI05,"projection")<-"LL"
MI05<-convUL (MI05, km=TRUE)
as.EventData(MI05, projection = NULL, zone = NULL)
MI05<-ppp(MI05$X,MI05$Y,window=owin(range(MI05$X),range(MI05$Y)))
smo<-density.ppp(MI05,sigma=3.5)

Ken<-envelope(MI05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=5.5, correction="trans")
plot(Ken)

mel05<-read.table("Mor2005M.txt",header=TRUE)
attr(mel05,"projection")<-"LL"
mel05<-convUL (mel05, km=TRUE)
as.EventData(mel05, projection = NULL, zone = NULL)
mel05<-ppp(mel05$X,mel05$Y,window=owin(range(mel05$X),range(mel05$Y)))
smo<-density.ppp(mel05,sigma=7)

Ken<- envelope(mel05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=7, correction="trans")
plot(Ken)
```

9.2.6.3 Breast cancer

The `spatstat` commands to get the estimated inhomogeneous *K*-plot for breast cancer incidence (Bcoord05.txt) and mortality (Mor2005B.txt) point patterns in 2005 are:

```
b05<-read.table("Bcoord05.txt",header=TRUE)
attr(b05,"projection")<-"LL"
b05<-convUL (b05, km=TRUE)
as.EventData(b05, projection = NULL, zone = NULL)
b05<-ppp(b05$X,b05$Y,window=owin(range(b05$X),range(b05$Y)))
smo<-density.ppp(b05,sigma=5)

Ken<- envelope(b05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=5, correction="border")
plot(Ken)



mb05<-read.table("Mor2005B.txt",header=TRUE)
attr(mb05,"projection")<-"LL"
mb05<-convUL (mb05, km=TRUE)
as.EventData(mb05, projection = NULL, zone = NULL)
smo<-density.ppp(mb05,sigma=6)

Ken<- envelope(mb05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=6, correction="trans")
plot(Ken)
```

9.2.6.4 Prostate cancer

The `spatstat` commands to get the estimated inhomogeneous *K*-plot for prostate cancer incidence (Pcoord05.txt) and mortality (Mor2005P.txt) point patterns in 2005 are:

```
p05<-read.table("Pcoord05.txt",header=TRUE)
attr(p05,"projection")<-"LL"
p05<-convUL (p05, km=TRUE)
as.EventData(p05, projection = NULL, zone = NULL)
p05<-ppp(p05$X,p05$Y,window=owin(range(p05$X),range(p05$Y)))

smo<-density.ppp(p05,sigma=5)
Ken<- envelope(p05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=5, correction="border")
plot(Ken)



mp05<-read.table("Mor2005P.txt",header=TRUE)
attr(mp05,"projection")<-"LL"
mp05<-convUL (mp05, km=TRUE)
mp05<-ppp(mp05$X,mp05$Y,window=owin(range(mp05$X),range(mp05$Y)))
as.EventData(mp05, projection = NULL, zone = NULL)

smo<-density.ppp(mp05,sigma=3.5)
Ken<- envelope(mp05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=7, correction="trans")
plot(Ken)
```

9.2.6.5 Colorectal cancer

The `spatstat` commands to get the estimated inhomogeneous *K*-plot for colorectal cancer incidence (Rcoord05.txt) and mortality (Mor2005R.txt) point patterns in 2005 are:

```
r05<-read.table("Rcoord05.txt",header=TRUE)
attr(r05,"projection")<-"LL"
r05<-convUL (r05, km=TRUE)
as.EventData(r05, projection = NULL, zone = NULL)
r05<-ppp(r05$X,r05$Y,window=owin(range(r05$X),range(r05$Y)))
```

```
smo<-density.ppp(r05,sigma=4.5)
Ken<- envelope(r05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=4.5, correction="border")
plot(Ken)

mr05<-read.table("Mor2005R.txt",header=TRUE)
attr(mr05,"projection")<-"LL"
mr05<-convUL (mr05, km=TRUE)
as.EventData(mr05, projection = NULL, zone = NULL)
mr05<-ppp(mr05$X,mr05$Y,window=owin(range(mr05$X),range(mr05$Y)))

smo<-density.ppp(mr05,sigma=2.5)
Ken<- envelope(mr05, Kinhom,
nsim=99,simulate=expression(rpoispp(smo)),sigma=6, correction="trans")
plot(Ken)
```

9.2.7 Code to produce estimation map

9.2.7.1 Melanoma cancer

The code to produce the estimation map for melanoma cancer incidence based on model (6.5) is:

```
estmap = function(file="perth suburb06",
        col.im=rgb(colorRamp(c("royalblue","green","yellow","white"))(seq(0,1,len=300)),
max=255)) {
        #col.im=terrain.colors(300)) {
require(grDevices); require(maptools); require(spatstat); require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
        rutm = sapply(sutm,range)
```

```
xlim=rutm[,"X"]; ylim=rutm[,"Y"]

m01<-read.table("Mcoord01.txt",header=TRUE)
attr( m01,"projection")<-"LL"
m01<-convUL ( m01, km=TRUE)
m01.ppp<-ppp( m01$X, m01$Y,window=owin(xlim,ylim))
m01.im<-distmap(m01.ppp)

burbs<- read.table("2006popsuburb.txt", header=TRUE)
attr(burbs,"projection")<-"LL"
burbs<-convUL (burbs, km=TRUE)
avepopden <- with(burbs, Total/Area)
burb <- with(burbs, ppp(X, Y, range(X),range(Y), marks=avepopden))
popdens <- smooth.ppp(burb)

mel05<-read.table("Mcoord05.txt",header=TRUE)
attr(mel05,"projection")<-"LL"
mel05<-convUL (mel05, km=TRUE)
mel05.ppp<-ppp(mel05$X,mel05$Y,window=owin(xlim,ylim))
mel05.ppm<- ppm(mel05.ppp, ~m01.im+popdens+
log(sqrt((x-391.0632)^2+(y+3536.509)^2)),
covariates=list(m01.im=m01.im))

        bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
        if (!exists("land",envir=.GlobalEnv)) {
                land = joinPolys(sutm,operation="UNION")
                assign("land",land,pos=1) }
        water = joinPolys(bbox,land,"DIFF")

        plot(predict(mel05.ppm), main="Estimated rates")

        addPolys(water,col="white",border=FALSE)
```

```
#browser();return()
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,

        m01=m01,m01.ppp=m01.ppp,m01.im=m01.im,popdens=popdens,mel05=mel05,me
l05.ppp=mel05.ppp,mel05.ppm=mel05.ppm,bbox=bbox,land=land,water=water))}

out=estmap()
```

The susceptible population density based on melanoma cancer incidence in 2001 (Mcoord01.txt) is applied as covariate to model melanoma cancer mortality in 2005 (Mor2005M.txt). The code to produce the estimation map of melanoma mortality in 2005 based on model (6.7) is shown below:

```
estmap = function(file="perth suburb06",
col.im=rgb(colorRamp(c("royalblue","green","yellow","white"))(seq(0,1,len=300)),max=2
55)) {#col.im=terrain.colors(300)) {require(grDevices); require(maptools); require(spatstat);
require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
        rutm = sapply(sutm,range)
        xlim=rutm[,"X"]; ylim=rutm[,"Y"]

m01<-read.table("Mcoord01.txt",header=TRUE)
attr( m01,"projection")<-"LL"
m01<-convUL ( m01, km=TRUE)
m01.ppp<-ppp( m01$X, m01$Y,window=owin(xlim,ylim))
m01.im<-distmap(m01.ppp)
mel05<-read.table("Mor2005M.txt",header=TRUE)
attr(mel05,"projection")<-"LL"
mel05<-convUL (mel05, km=TRUE)
```

```
mel05.ppp<-ppp(mel05$X,mel05$Y,window=owin(xlim,ylim))
mel05.ppm<- ppm(mel05.ppp, ~m01.im+log(sqrt((x-391.0632)^2+(y+3536.509)^2)),
covariates=list(m01.im=m01.im))

bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
if (!exists("land",envir=.GlobalEnv)) {
land = joinPolys(sutm,operation="UNION")
assign("land",land,pos=1) }
water = joinPolys(bbox,land,"DIFF")
plot(predict(mel05.ppm), main="Estimated rates")
addPolys(water,col="white",border=FALSE)
#browser();return()
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,
        m01=m01,m01.ppp=m01.ppp,m01.im=m01.im,mel05=mel05,mel05.ppp=mel05.pp
p,mel05.ppm=mel05.ppm,bbox=bbox,land=land,water=water)) }

out=estmap()
```

9.2.7.2 Breast cancer

The code to produce the estimation map for breast cancer incidence based on model (6.10) is:

```
estmap = function(file="perth suburb06",
        col.im=rgb(colorRamp(c("royalblue","green","yellow","white"))(seq(0,1,len=300)),
max=255)) {
        #col.im=terrain.colors(300)) {
require(grDevices); require(maptools); require(spatstat); require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
        rutm = sapply(sutm,range)
```

```
        xlim=rutm[,"X"]; ylim=rutm[,"Y"]

b92<-read.table("Bcoord92.txt",header=TRUE)
attr( b92,"projection")<-"LL"
b92<-convUL ( b92, km=TRUE)
b92.ppp<-ppp( b92$X, b92$Y,window=owin(xlim,ylim))
b92.im<-distmap(b92.ppp)

burbsf<- read.table("2006popsuburbsex.txt", header=TRUE)
attr(burbsf,"projection")<-"LL"
burbsf<-convUL (burbsf, km=TRUE)
avepopdenf <- with(burbsf, Total/Area)
burbf<- with(burbsf, ppp(X, Y, range(X),range(Y), marks=avepopdenf))
popdensf <- smooth.ppp(burbsf)

b05<-read.table("Bcoord05.txt ",header=TRUE)
attr(b05,"projection")<-"LL"
b05<-convUL (b05, km=TRUE)
b05.ppp<-ppp(b05$X,b05$Y,window=owin(xlim,ylim))
b05.ppm<- ppm(b05.ppp, ~b92.im+popdensf+log(sqrt((x-391.0632)^2+(y+3536.509)^2)),
covariates=list(b92.im=b92.im))

        bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
        if (!exists("land",envir=.GlobalEnv)) {
                land = joinPolys(sutm,operation="UNION")
        assign("land",land,pos=1) }
        water = joinPolys(bbox,land,"DIFF")

        plot(predict(b05.ppm), main="Estimated rates")

        addPolys(water,col="white",border=FALSE)

#browser();return()
```

```
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,
b92=b92,b92.ppp=b92.ppp,b92.im=b92.im,popdensf=popdensf,b05=b05,b05.ppp=b05.ppp,
b05.ppm=b05.ppm,bbox=bbox,land=land,water=water))}

out=estmap()
```

The susceptible population density based on breast cancer incidence in 1999 (Bcoord99.txt) is applied as covariate to model breast cancer mortality in 2005 (Mor2005B.txt). The code to produce the estimation map of breast mortality in 2005 based on model (6.11) is shown below:

```
estmap = function(file="perth suburb06",
        col.im=rgb(colorRamp(c("royalblue","green","yellow","white"))(seq(0,1,len=300)),
max=255)) {
        #col.im=terrain.colors(300)) {
require(grDevices); require(maptools); require(spatstat); require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
        rutm = sapply(sutm,range)
        xlim=rutm[,"X"]; ylim=rutm[,"Y"]

b99<-read.table("Bcoord99.txt",header=TRUE)
attr( b99,"projection")<-"LL"
b99<-convUL ( b99, km=TRUE)
b99.ppp<-ppp( b99$X, b99$Y,window=owin(xlim,ylim))
b99.im<-distmap(b99.ppp)

burbsf<- read.table("2006popsuburbsex.txt", header=TRUE)
attr(burbsf,"projection")<-"LL"
burbsf<-convUL (burbsf, km=TRUE)
```

```
avepopdenf <- with(burbsf, Fpop/Area)
burbsf <- with(burbsf, ppp(X, Y, range(X),range(Y), marks=avepopdenf))
popdensf <- smooth.ppp(burbsf)

mb05<-read.table("Mor2005B.txt",header=TRUE)
attr(mb05,"projection")<-"LL"
mb05<-convUL (mb05, km=TRUE)
mb05.ppp<-ppp(mb05$X,mb05$Y,window=owin(xlim,ylim))
mb05.ppm<- ppm(mb05.ppp, ~b99.im+popdensf+log(sqrt((x-
391.0632)^2+(y+3536.509)^2)), covariates=list(b99.im=b99.im,popdensf=popdensf))

        bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
        if (!exists("land",envir=.GlobalEnv)) {
                land = joinPolys(sutm,operation="UNION")
                assign("land",land,pos=1) }
        water = joinPolys(bbox,land,"DIFF")

        plot(predict(mb05.ppm), main="Estimated rates")

        addPolys(water,col="white",border=FALSE)

#browser();return()
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,

        b99=b99,b99.ppp=b99.ppp,b99.im=b99.im,popdensf=popdensf,mb05=mb05,mb05.
ppp=mb05.ppp,mb05.ppm=mb05.ppm,bbox=bbox,land=land,water=water))}

out=estmap()
```

9.2.7.3 Prostate cancer

The susceptible population density based on prostate cancer incidence in 2002 (Pcoord02.txt) is applied as covariate to model prostate cancer mortality in 2005 (Mor2005P.txt). The code to produce the estimation map of prostate mortality in 2005 based on model (6.12) is shown below:

```
estmap = function(file="perth suburb06",
        col.im=rgb(colorRamp(c("royalblue","green","yellow","white"))(seq(0,1,len=300)),
max=255)) {
        #col.im=terrain.colors(300)) {
        require(grDevices); require(maptools); require(spatstat); require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
        rutm = sapply(sutm,range)
        xlim=rutm[,"X"]; ylim=rutm[,"Y"]

p02<-read.table("Pcoord02.txt",header=TRUE)
attr( p02,"projection")<-"LL"
p02<-convUL ( p02, km=TRUE)
p02.ppp<-ppp( p02$X, p02$Y,window=owin(xlim,ylim))
p02.im<-distmap(p02.ppp)

mp05<-read.table("Mor2005P.txt",header=TRUE)
attr(mp05,"projection")<-"LL"
mp05<-convUL (mp05, km=TRUE)
mp05.ppp<-ppp(mp05$X,mp05$Y,window=owin(xlim,ylim))
mp05.ppm<- ppm(mp05.ppp, ~p02.im+log(sqrt((x-391.0632)^2+(y+3536.509)^2)),
covariates=list(p02.im=p02.im))
```

```
        bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
        if (!exists("land",envir=.GlobalEnv)) {
                land = joinPolys(sutm,operation="UNION")
        assign("land",land,pos=1) }
        water = joinPolys(bbox,land,"DIFF")

        plot(predict(mp05.ppm), main="Estimated rates")

        addPolys(water,col="white",border=FALSE)

#browser();return()
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,

        p02=p02,p02.ppp=p02.ppp,p02.im=p02.im,mp05=mp05,mp05.ppp=mp05.ppp,mp0
5.ppm=mp05.ppm,bbox=bbox,land=land,water=water))}

out=estmap()
```

9.2.7.4 Colorectal cancer

The code to produce the estimation map for colorectal cancer incidence based on model (6.8) is:

```
estmap = function(file="perth suburb06",
        col.im=rgb(colorRamp(c("royalblue","green","yellow","white"))(seq(0,1,len=300)),
max=255)) {
        #col.im=terrain.colors(300)) {
require(grDevices); require(maptools); require(spatstat); require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
```

```
        rutm = sapply(sutm,range)
        xlim=rutm[,"X"]; ylim=rutm[,"Y"]

r98<-read.table("Rcoord98.txt",header=TRUE)
attr( r98,"projection")<-"LL"
r98<-convUL ( r98, km=TRUE)
r98.ppp<-ppp( r98$X, r98$Y,window=owin(xlim,ylim))
r98.im<-distmap(r98.ppp)

burbs<- read.table("2006popsuburb.txt", header=TRUE)
attr(burbs,"projection")<-"LL"
burbs<-convUL (burbs, km=TRUE)
avepopden <- with(burbs, Total/Area)
burb <- with(burbs, ppp(X, Y, range(X),range(Y), marks=avepopden))
popdens <- smooth.ppp(burbs)

mr05<-read.table("Mor2005R.txt",header=TRUE)
attr(mr05,"projection")<-"LL"
mr05<-convUL (mr05, km=TRUE)
mr05.ppp<-ppp(mr05$X,mr05$Y,window=owin(xlim,ylim))
mr05.ppm<- ppm(mr05.ppp, ~r98.im+popdens+log(sqrt((x-391.0632)^2+(y+3536.509)^2)),
covariates=list(r98.im=r98.im))

        bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
        if (!exists("land",envir=.GlobalEnv)) {
                land = joinPolys(sutm,operation="UNION")
                assign("land",land,pos=1) }
        water = joinPolys(bbox,land,"DIFF")

        plot(predict(mr05.ppm), main="Estimated rates")

        addPolys(water,col="white",border=FALSE)
```

```
#browser();return()
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,
                r98=r98,r98.ppp=r98.ppp,r98.im=r98.im,popdens=popdens,
mr05=mr05,mr05.ppp=mr05.ppp,mr05.ppm=mr05.ppm,bbox=bbox,land=land,water=water
))}

out=estmap()
```

The susceptible population density based on colorectal cancer incidence in 2003 (Rcoord03.txt) is applied as covariate to model colorectal cancer mortality in 2005 (Mor2005R.txt). The code to produce the estimation map of colorectal mortality in 2005 based on model (6.9) is shown below:

```
estmap = function(file="perth suburb06",
        col.im=rgb(colorRamp(c("royalblue","green","yellow","white"))(seq(0,1,len=300)),
max=255)) {
        #col.im=terrain.colors(300)) {

require(grDevices); require(maptools); require(spatstat); require(PBSmapping)
        sll = importShapefile ("perth suburb06.shp")
        attr(sll,"projection")<-"LL"
        sutm = convUL (sll, km=TRUE)
        plotMap(sutm, projection = TRUE, tckLab = TRUE)
        rutm = sapply(sutm,range)
        xlim=rutm[,"X"]; ylim=rutm[,"Y"]

r03<-read.table("Rcoord03.txt",header=TRUE)
attr( r03,"projection")<-"LL"
r03<-convUL ( r03, km=TRUE)
r03.ppp<-ppp( r03$X, r03$Y,window=owin(xlim,ylim))
r03.im<-distmap(r03.ppp)
```

```
mr05<-read.table("Mor2005R.txt",header=TRUE)
attr(mr05,"projection")<-"LL"
mr05<-convUL (mr05, km=TRUE)
mr05.ppp<-ppp(mr05$X,mr05$Y,window=owin(xlim,ylim))
mr05.ppm<- ppm(mr05.ppp, ~r03.im+log(sqrt((x-391.0632)^2+(y+3536.509)^2)),
covariates=list(r03.im=r03.im))

        bbox=makeGrid(xlim,ylim,addSID=FALSE,projection="UTM")
        if (!exists("land",envir=.GlobalEnv)) {
                land = joinPolys(sutm,operation="UNION")
                assign("land",land,pos=1) }
        water = joinPolys(bbox,land,"DIFF")

        plot(predict(mr05.ppm), main="Estimated rates")
        #plot(mr05.ppm, main="Estimated rates")

        addPolys(water,col="white",border=FALSE)
#browser();return()
        invisible(list(sll=sll,sutm=sutm,xlim=xlim,ylim=ylim,

        r03=r03,r03.ppp=r03.ppp,r03.im=r03.im,mr05=mr05,mr05.ppp=mr05.ppp,mr05.pp
m=mr05.ppm,bbox=bbox,land=land,water=water))}

out=estmap()
```