Cardiovascular Disease Mortality, Socio-Economic Status and Air Pollution: Understanding the Geographical Patterns.

Ivan Hanigan 2005

A Thesis Submitted for the Degree of Bachelor of Arts with Honours in Environmental Health,

The School of Resources, Environment and Society, and

The National Centre for Epidemiology and Population Health,

The Australian National University, Canberra.

Supervisors: Mr. Ken Johnson (SRES) and Dr. Rennie M. D'Souza (NCEPH)

Adviser: Prof. Tord Kjellstrom (NCEPH).

Contact details:

Email: ivan.hanigan@anu.edu.au or konivanivan@yahoo.com

Phone: (02) 6278 3965

Mobile: 0407 1389 15

Statement of Originality

Except where otherwise acknowledged, this thesis is my own original work.

Acknowledgements

I am grateful to the input of the following:

Prof. Tony McMichael, Dr. Mark Clements, Dr. Keith Dear, Ms. Melissa Goodwin, Dr. Simon Hales, Ms. Rosemary Korda, Dr. Rosalie Woodruff, Dr. Hilary Bambrick, Mr. Rupendra Shrestha, Dr. Bruce Doran; Mr. Chris Eiser, Mr. Alan Betts and Mr. Matthew Riley, Mr. Frank Blanchfield, Mr. Alec Bamber, the Australian Bureau of Statistics Geography Department, the NSW Department of Environment and Conservation, Radha, Peter and Robyn Hanigan, and all my friends

Table of Contents

List of Tables

List of Figures

Glossary

$\mu g/m^3$	Micrograms per cubic metre
ABS	Australian Bureau of Statistics
AP	Air Pollution
ASGC	Australian Standard Geographical Classification System. Includes CCD, SLA and SD.
CCD	Census Collector's Districts
CVD	Cardio-Vascular Disease
Georeference	The spatial coordinates that allow objects and attributes to be mapped
MAUP	Modifiable Areal Unit Problem
NMD	National Mortality Database
NO_2	Nitrogen dioxide
O_3	Ozone
PM_{10}	Particulate Matter with aerodynamic diameter less than 10um
POA	Postal Areas derived from CCDs
pphm	Parts per hundred mullion
SD	Statistical Division
SEIFA	Socio-Economic Indexes for Areas from the ABS
SES	Socio-Economic Status
SLA	Statistical Local Areas
TEOM	Tapered Element Oscillating Microbalance. Instrument for PM_{10} measurement.

Abstract

There is a considerable amount of epidemiological literature identifying relationships between Socio-Economic Status (SES), ambient Air Pollution (AP) and Cardiovascular Disease (CVD). Processes determining these relationships are elements of the physical and social systems in human ecology. There are numerous studies that show a negative gradient of this category of disease between populations with increasing levels of SES. The influence of air pollutants on CVD mortality is not yet conclusive, however results are suggestive and plausible cardiovascular mechanisms are known. This study investigates the geographical pattern of socio-economic status and air pollution across Sydney and the relationships with CVD mortality.

The mortality data are available from the Australian National Mortality Database (NMD). These data aggregate the locations of usual residence at time of death to spatial units. There are two kinds of spatial unit currently available: Postal Areas (POA) and Statistical Local Areas (SLA). A central problem of geographical analysis of environmental and health data is that analyses using information collected at an inappropriate scale may obscure or distort the relationships investigated. The observed relationships may differ substantially between types of spatial units because of the issues of scale and zonation. These issues are known collectively as the Modifiable Areal Unit Problem (MAUP) in geographical terminology. The scale issue has an effect by varying the size of the aggregation units used, while the zonation issue has an effect by changing the shape of spatial units at the same scale. A key aspect of this study is the exploration of the effect of the scale of the spatial units used in these small area data, aiming to improve understanding of the geographical patterns of the relationships.

CVD mortality data were extracted and environmental data were processed and integrated for Sydney for the period 1996 until 1998. Four consecutive six-month seasons were used to account for the strong seasonal variation of these diseases. Regression tree models and stratified linear regression were used to explore relationships shown by these data. Regression trees are a data-mining approach that allows many variables to be assessed and key relationships identified.

The tree model results identified similar variables in the analysis at the two levels of aggregation. The winter season, ABS Index of Relative Socio-Economic Disadvantage and particulate matter with aerodynamic diameter less than 10 micrometers (PM_{10}) defined regions with different rates of CVD mortality in both sets of spatial units. These variables were explored using linear regression models that were stratified to control for the interaction between these such that only the winter rates were regressed against the disadvantage index. Then the winter rates for the most disadvantaged areas were analysed for a relationship with PM_{10}. It seems that the mortality rates in the most disadvantaged areas are influenced by this pollutant more than the less disadvantaged areas.

The relationships observed are slightly different between the two types of spatial units. The POA-level regression slope for winter rates against the disadvantage index was not as steep, but more statistically significant than that found at SLA-level. The relationship with PM_{10} observed in the disadvantaged areas in winter at the POA-level was not significant whereas those at the SLA-level were. From these results it can be concluded that scale of analysis does influence the understanding of geographical patterns of Socio-Economic Status, Air Pollution and Cardiovascular Disease mortality.

Introduction

Cardiovascular Disease (CVD) is an important category of diseases causing death in Australia. The trend in CVD mortality throughout the 20th century showed an increase from 15% of all Australian deaths in 1907 to the peak at 56% in 1968, and despite a steady decline since then was the leading cause of death in the year 2000, accounting for 39% of deaths (Australian Bureau of Statistics 2002). The study of the causes of CVD is an important field of public health research with the potential to prevent a large proportion of these deaths in the future.

The CVD category can be defined as:

> Ischemic heart disease (heart tissue damage due to obstruction of blood flow to the heart), cerebrovascular disease (blood vessels supplying the brain), atherosclerosis (hardening of the arteries), hypertension (high arterial blood pressure), and rheumatic heart disease (reduced function of the heart because of heart inflammation and scarring due to previous rheumatic fever) (Meade and Earickson 2000: 234).

Socio-Economic Status (SES) and ambient Air Pollution (AP) have been implicated in disparities between rates of CVD mortality between populations (Bennett 1996, World Health Organization 2000). The majority of air pollution studies have used time-series analysis, focused on daily variations in mortality rates. For the most part these studies aggregate the spatial aspects of the health and pollution phenomena to the level of the entire city. With the advent of Geographical Information Systems (GIS), there is now a new wave of analyses using spatially organised data.

In Australia privacy regulations are enforced to restrict access to sensitive information with fine spatial resolution. There are ethical guidelines and the study of these data requires approval from ethics committees (National Health and Medical Research Council 1999). To protect confidentiality the data are aggregated to large spatial units: Australia Post Postcodes and Australian Bureau of Statistics Statistical Local Areas (SLA). However, Postcodes do not match the ABS census geography and so Census Collector's Districts (CCD) are aggregated to approximate the population of Postcodes. These CCD agglomerations are termed Postal Areas (POA) and despite some boundaries that do not match those of real Postcodes, they are the best estimate of Postcode populations available. SLAs are medium level units in the Australian Standard Geographical Classification System (ASGC) and are used to publish a wide range of demographic data from the five-yearly census of population and housing. These two types of spatial unit represent quite different levels of spatial aggregation.

The research questions to be investigated using these data can be articulated as:

1. What relationships are observable between SES, AP and CVD mortality across Sydney?
2. Do these differ depending on the scale of the spatial units used?

The relationship between health and environment is a complex system that can be conceptualised using the hierarchically structured ecosystem paradigm of ecological theory (Allen and Hoekstra 1992, Kay et al. 1999, Gunderson and Holling 2002). This paradigm considers ecosystems as sets of components (and their relationships) at levels in a hierarchy from local, up to the global level. Lower or higher levels of the ecosystem have differently scaled components and systems operating. The systems at these different levels are driven by internal feedbacks, as well as the relationships and function of the system at higher levels. For example individuals and their exposures are important for conceptualising

the local level of the system, whilst sub-populations are important at more macro levels and constrain the behaviour of the individuals at the lower level.

Urban ecosystems are especially complex and large cities have many layers of interconnected socio-economic and physical systems operating. These include population movement, cultural mixing and genetic backgrounds. The social context may vary substantially across small areas of the city and there is also large variation in the physical environment. The temporal element of these systems complicates understanding as the dynamic processes of exposure and effect may span days, weeks, years or even decades. To understand the ecology of CVD we must consider the processes whereby individual risk factors interplay with social and environmental contexts.

In environmental epidemiology the conceptual framework has been described as a multifactorial "web of causation" that includes both proximal and distal components of the health-environment system (Krieger 1994). The analysis of CVD mortality is inevitably confounded by the many relationships of this web. Therefore understanding of the complex system will always be limited depending on availability of data on the necessary variables, observed at an appropriate scale.

A methodology used to assess health disparities across sub-populations is the aggregate-level study of health and environmental data. This method is sometimes referred to as the "ecological method" by epidemiologists and allows investigation of different levels of human ecosystems. The health and environment data are aggregated in space and time to delineate meaningful population sub-groups whose health and exposure attributes can be compared. The aim of many aggregate-level studies is to discern spatial patterns of the variables of interest.

CVD mortality exhibits broad regional trends across population groups as most sub-populations in Australia have a large number of individuals at risk due to age and lifestyle factors. The spatial patterns of SES and AP are known to respond to local variations in social and physical environments. As these local patterns vary more than population attributes (such as age structure) the effect of these will overlay the broad trends as fine resolution spatial pattern.

There is a well-recognised issue regarding the analysis of aggregate-level data known as the Modifiable Areal Unit Problem (MAUP) (Openshaw 1983). This problem consists of two sub-problems; the scale issue and the zonation issue. The scale issue is a problem when the aggregation units used are an inappropriate scale for the phenomenon under investigation, while the zonation issue is a problem when aggregation units produce different results to other spatial units at the same scale due to their shape, therefore obfuscating the understanding of relationships. The effect of the MAUP confounds cross-level inference between levels of observation and the term "ecological fallacy" has been used to describe the inappropriate inference of individual-level associations from aggregate-level results. This can introduce errors, bias and contradictions to the understanding of processes causing disease.

The approach taken here is to conduct an aggregate level study of mortality rates and environmental context while assessing the issues of scale and zonation in the spatial units used to georeference mortality data in Sydney. The methodology used is essentially reductionist as it tries to simplify the complexity of the system and investigate only a specific set of relationships between socio-economic, atmospheric and disease subsystems. It is apparent that any reductionist analysis will give limited explanation of the variation observed and it is conceded that many variables not included also have extraneous effects on

these relationships. These may have independent effects on the diseases, or modify the effect of SES and AP on CVD (Blakely and Woodward 2000).

For this study the data are sourced from existing collections. Tree-based models were used first to explore these data. This method is a relatively assumption free, flexible, non-parametric, and non-normal approach to multivariate models (Breiman 1984). The tree algorithm recursively partitions a dataset into more homogenous subsets based on multiple variables. The order of importance of the different variables and optimum split levels in these are identified. The rules identified were used in Stratified Linear Regression to quantify the relationships while controlling for the other covariates which also influence CVD mortality (Shannon et al. 2002).

Structure

The background section describes the ecology of CVD, and the influences of SES and AP on CVD mortality are identified from epidemiological literature. Then the context of the Sydney mortality data, landscape and population context are described before discussion turns to the issues of scale and zonation identified from geographical literature.

The spatiotemporal framework for analysing health and environmental data at aggregated levels in spacetime is explored in the methodology section.

The data section outlines the attributes of the CVD, SES and AP data before the regression tree and linear regression modelling techniques are described.

The results section presents the quantification of the health and environment datasets for the two types of spatial units. Then the results of the regression tree models are presented. These models identified key variables such as season, disadvantage and PM_{10} that were then visually explored using scatter plots and Locally Weighted Smoothing (loess) lines. These relationships appear linear and so linear regression was used to quantify and compare them.

The discussion section has special focus on the differences between the relationships of SES and AP with CVD depending on the type of spatial units used. First, focus is on characterisation of the tree model's partitioning rules and presents the patterns found for the different types of spatial units. Then the results of the stratified linear regressions are discussed.

The conclusion offers a synthesis of the results of this study. There were relationships observed between SES, AP and CVD in the Sydney data. These did differ slightly between the types of spatial unit used. Ultimately the judgement whether one set of spatial units are better or worse for understanding the geographical patterns of disease and exposure depends on the scale of the phenomenon under investigation. In the case of CVD it appears that both sets of spatial units are useful.

Background

The Ecology of Cardiovascular Disease

CVD is an important category of diseases in countries that have high levels of industrialisation; sedentary activity patterns; consumptive lifestyles; inequities in distribution of socio-economic benefits; and atmospheric pollution.

The mechanisms by which SES and AP influence CVD mortality can be described using the interconnected elements of a triangular conceptual framework consisting of population, behaviour, and habitat as the corners (Meade and Earickson 2000: 233-242).

The population element includes the age, sex and genetic characteristics of the people. The salient feature is that the cardiovascular system becomes clogged, inefficient or faulty with age. Genetic and metabolic processes influence this universal process creating some disparity between sub-population groups.

The behaviour element incorporates the way humans interact with one another and with their environment. These include smoking, physical activity, and diet. Cultural differences between different groups of people will drive different behaviour patterns and thus the exposure and susceptibility of individuals to these CVD health risks.

In the habitat element there are many environmental factors implicated with CVD including: SES; atmospheric pollution; meteorological systems; water quality; and aspects of landscape geochemistry (Meade and Earickson 2000).

The focus of this study is only on the SES and AP relationships as it is likely these will display geographical patterns across the city. The statistical modelling used to understand these relationships needs to accept the importance of individual level processes including exposure and susceptibility. The individual level however is not available for analysis using these administrative data and so aggregate-level data is used to describe the geographical contexts of the mortality rates.

Socio-Economic Status

The social epidemiology literature focuses on health inequalities caused by the differences in poverty, deprivation, and status (Gould-Ellen et al. 2001, Berkman and Kawachi 2003, O'Neill et al. 2003, Kawachi and O'Neill 2005). There is often a negative relationship evident in morbidity and mortality rates with increasing SES. This may be due to a variety of individual level risk factors such as diet or smoking, which are influenced by SES. Other explanations refer to the contextual attributes of disadvantaged neighbourhoods. This understanding is further complicated by findings that it is not just simple material wealth or deprivation, but relative status and hierarchy that are involved (Meade and Earickson 2000, Krieger et al. 2004)

In Australia studies have found differences between SES groups in the rates of various causes of death, including CVD (Bennett 1996, Taylor et al. 1999, Turrell and Mathers 2000, Turrell and Mathers 2001, Stocks et al. 2004).

The mechanisms whereby SES influences CVD mortality are complex and may include: social relationships; access to medical resources; harmful behaviour such as smoking; inadequate or

inappropriate diet; and polluted indoor atmospheric conditions. These mechanisms have been primarily investigated by individual risk-factor epidemiology, comparisons between groups in aggregate-level studies and recently by multi-level modelling that includes individual and aggregate-level data.

Air Pollution

AP occurs when harmful substances are present in the atmosphere in areas where human populations can be exposed to them (Yassi et al. 1998, World Health Organization 2000).

The three pollutants: coarse particulate matter (PM_{10}); ozone (O_3); and nitrogen dioxide (NO_2) selected for this study are identified in the literature as key pollutants associated with CVD in Sydney (Morgan et al. 2003).

In a meta-analysis by (Stieb et al. 2002) the CVD relationship was unclear, which surprised the authors who cited references which suggest:

> There are well-recognized cardiorespiratory pathophysiological mechanisms for the effects of certain pollutants, and, in particular, evidence has recently emerged of mechanisms observed in humans, which could explain the association of both particulate and gaseous pollutants with cardiovascular health outcomes, including changes in autonomic control, as reflected in changes in heart rate or heart rate variability, blood coagulability and viscosity, blood pressure, and bone marrow response (Stieb et al. 2002: 478).

Epidemiological literature that assesses the relationship between CVD and AP have primarily used time-series methods that study variations through time (usually daily variations) across whole cities and relate these to AP levels (Morgan et al. 1998, Morgan et al. 2003). There is now a growing number of spatial analyses that investigate the relationship of AP and disease (Burnett et al. 2001, Pikhart et al. 2001, Scoggins et al. 2004).

Interactions

Some studies have identified an interaction between SES and AP on health outcomes (Kawachi and O'Neill 2005). These can interact so that disadvantaged populations are badly affected by AP exposures while high SES groups are not. SES can theoretically modify the influence of AP because of: differences in AP exposure between SES groups or increased susceptibility to AP related health conditions (O'Neill et al. 2003).

The interactions between SES and AP demand modelling that deals with non-linearity and confounding between the variables. Therefore the framework employed in this study uses regression trees and stratified linear models that can handle these complicated relationships.

The Sydney Context

There is variation across Sydney in terms of CVD mortality rates, SES and AP. The following discussion will describe the temporal distribution of CVD in Sydney and then the spatial patterns.

Time

Temporal patterns in the crude daily CVD mortality rate for the Sydney statistical division (SD) through the period 1995 until 2002 are shown in figure 1. The seasonal pattern in the mortality rates is distinct, with an increase in the long-term smoothed rate of approximately 60% each winter. There also appears to be a general decline in rates over the period. There is some under reporting evident in the final months of the time series.

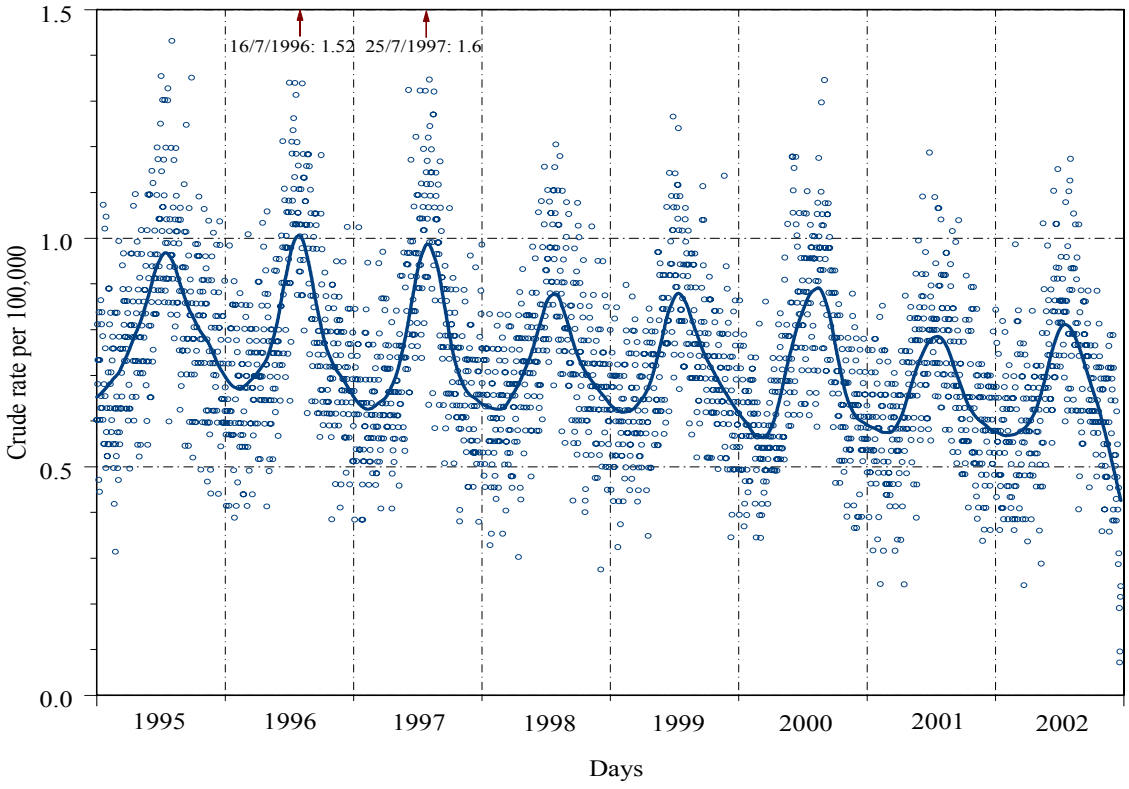


Figure 1: Daily Crude Rates per 100,000 for CVD Mortality in Sydney, 1995-2002.

These trends in the data suggest a strong influence of the seasons. Winter rates may be higher due to the different climatic conditions, or changed susceptibility in the population during those times.

AP has also been linked with increases in the death rate. Time-series studies in Sydney suggest a relationship between particulate AP with CVD (Morgan et al. 1998, Morgan et al. 2003). An increase in daily mean particulate concentration from the 10th to the 90th centile was associated with an increase of 2.68% (0.25 to 5.16) in daily cardiovascular mortality (Morgan et al. 1998).

Space

Sydney is ideal for AP and health study as it has a large population in excess of 4 million people and is situated in an airshed that is well monitored for pollutants.

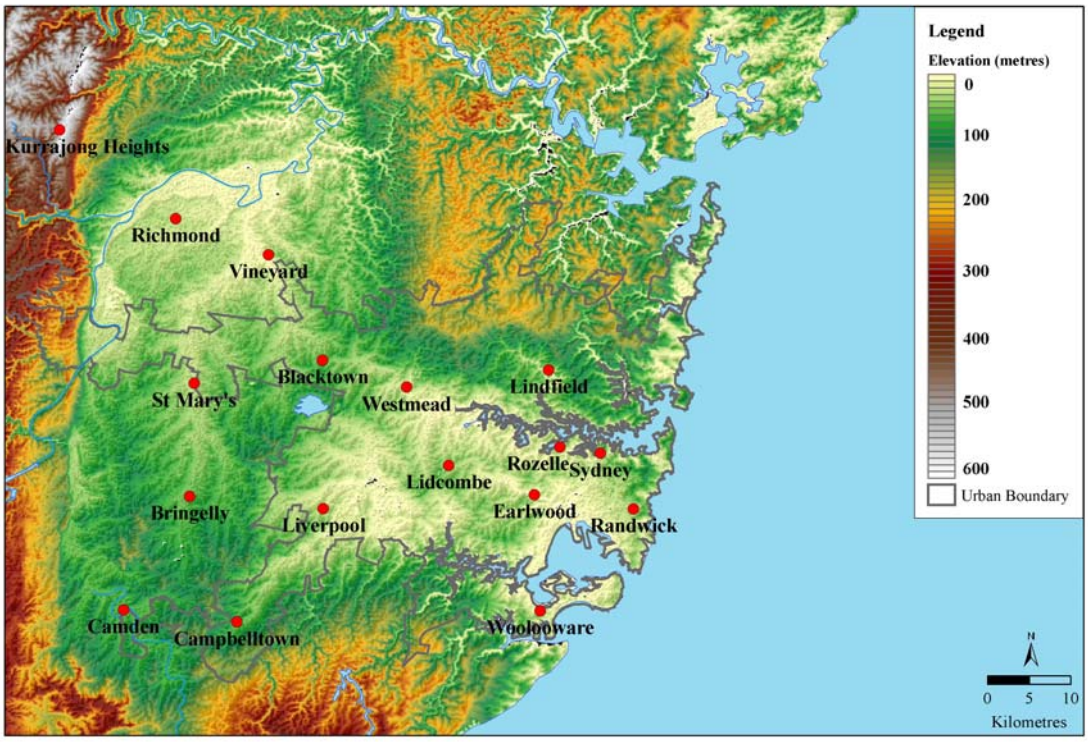


Figure 2: The Sydney Basin and Air Pollution Monitoring Sites.

The Sydney Basin topography and AP monitoring network can be seen in figure 2. The New South Wales Department of Environment air monitoring network is shown by the red dots. The urban area is delineated by the grey line. The central business district is located around the air pollution site labelled 'Sydney', and extends a short distance north, south and east around Sydney Harbour. Sydney is located on the coast of New South Wales and clusters around the Sydney harbour and the Parramatta River. The city extends north to Broken Bay, the Hawkesbury and Nepean Rivers in the northwest and the Royal National Park in the south. The western suburbs are generally medium density suburbia with some industrial areas.

The AP network is concentrated in the bottom of the basin and was designed to assess the exposures of this part of the population (NSW Department of Environment and Conservation 2001). However there is a weakness in coverage by the monitoring network for the northern periphery of the city and therefore estimating exposures from these will be a key limitation of this study (McPhail 1996).

Spatially Organised Data

In the NMD the data are aggregated to POA and SLA by the ABS to protect the confidentiality of the information. The ABS have only used POA in the NMD since 1995. The different scales of the available spatial units can be seen in figures 3 and 4. These maps depict the spatial pattern of the disadvantage index, one of the Socio-Economic Indexes For Areas (SEIFA) available from the ABS. The SEIFA indexes are constructed from indicators measured at the census. The ABS uses SLAs for the collection and dissemination of data from the Census of Population and Housing, and therefore SLAs have a great deal of data published for them. Australia Post Postcodes on the other hand are designed for delivery of mail and demographic data do not get collected specifically for these areas.

Postcodes and Postal Areas

The boundaries of Postcodes in Australia do not match those of the CCDs so there is some potential for spatial mismatch with Postal Areas (POA) units differing substantially from 'real' Postcode areas. Due to the spatial mismatch precise populations are not available for Postcodes. In addition some Postcodes are excluded because they are smaller than CCDs and some Postcodes consist of many multiple distinct areas, potentially several kilometres apart. These problems are much worse in rural than urban areas. These attributes make Postcodes problematic for use in geographical analysis of environmental exposures as it cannot be assumed that health indicators or exposure estimates have high accuracy despite the quality of the underlying health or environmental datasets.

Statistical Local Areas

Statistical Local Areas are a medium level spatial unit in the Australian Standard Geographical Classification (ASGC) system. The ASGC is a set of hierarchically nested non-overlapping spatial units used by the ABS for collecting and publishing the Australian Census since 1984. There are changes in boundaries, names and codes over time, sometimes twice in one year (e.g. editions 7 and 8 were both released in 1989).

The SLAs are usually large areas that tend to 'smooth' the detail of the social and demographic data they portray. This effect can be seen in figure 4 when the disadvantage index is aggregated to the POA and SLA levels. The broad regional trend is picked up by the SLAs and we can see that the northern areas are generally less disadvantaged than the inner southwest areas. However the POA show there is some fine detail pattern of high disadvantage between the central west and the inner southwest, and some heterogeneity in the north that is not displayed by the SLAs. This means the scale of the two spatial units give different views of the spatial pattern.

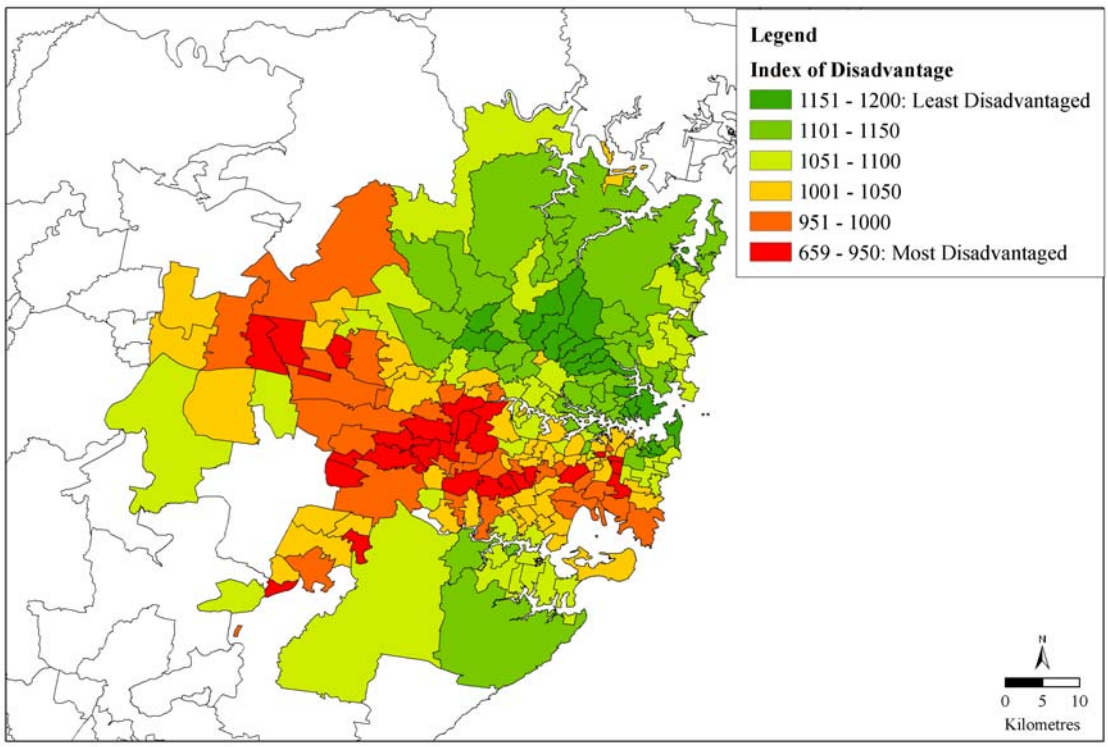


Figure 3: The Disadvantage Index at POA Level

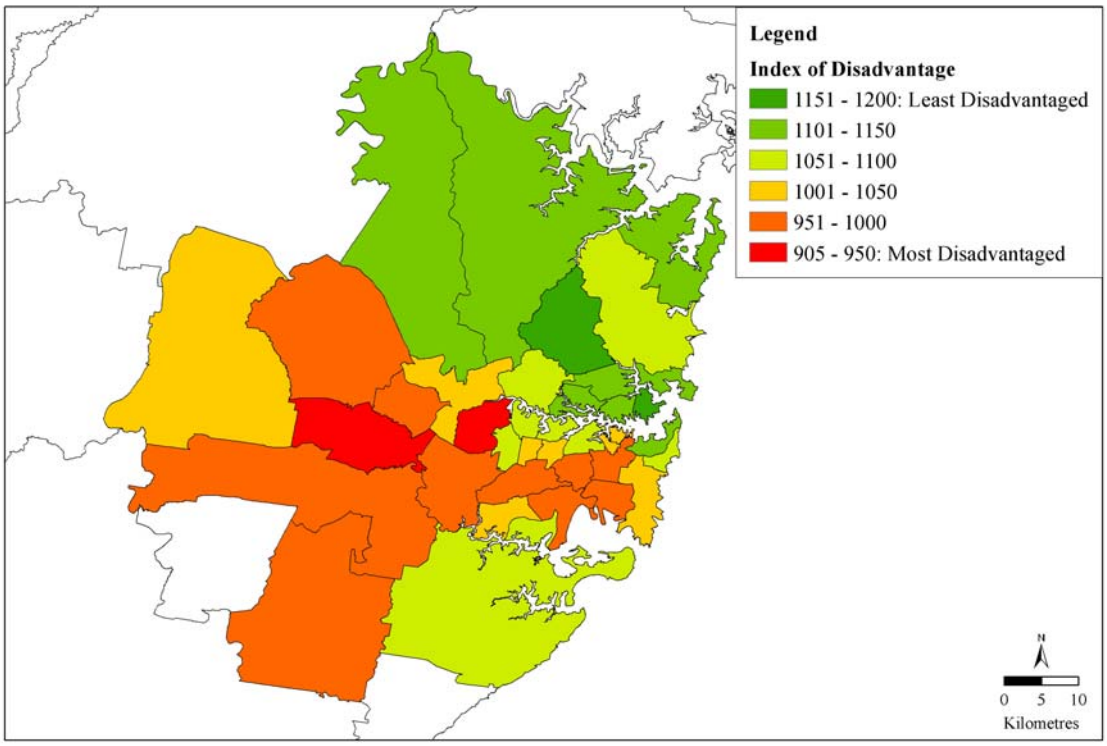


Figure 4: The Disadvantage Index at SLA Level

Summary of Spatial Units

The population attributes of the two types of spatial units may differ substantially. The box plots in figures 5 and 6 show the range in population and area for these two types of spatial units in Sydney. The grey box depicts the interquartile range and hides all data points except for the median value. The whiskers hide points within one half of the interquartile range in figure 5 while in figure 6 they hide those within one-and-a-half times this range.

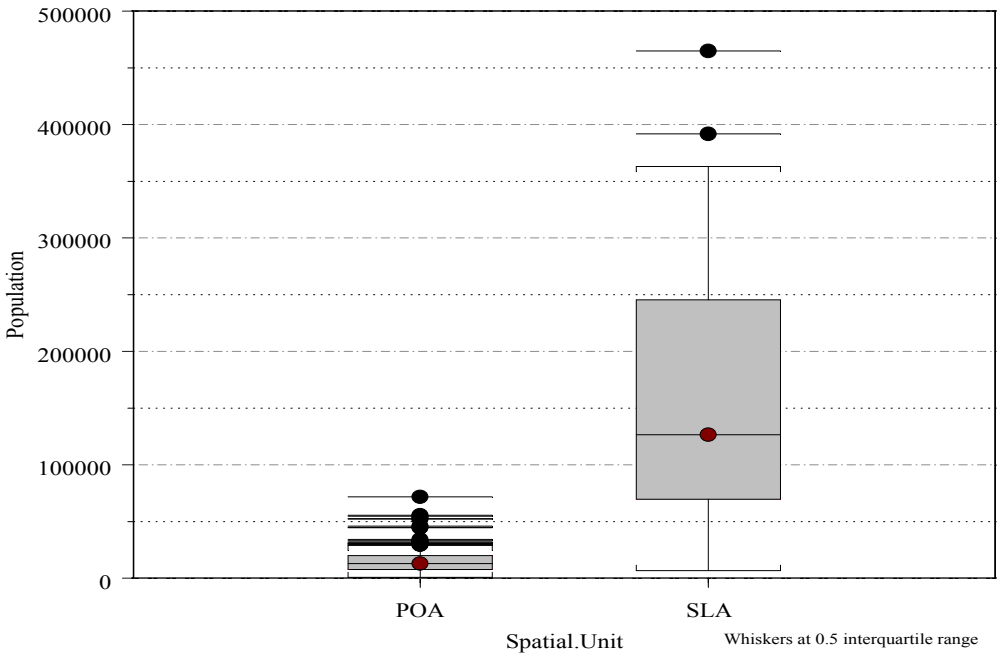


Figure 5: Box Plots of Population by Type of Spatial Units

SLAs have a median population of 125,000 compared with 20,000 persons for POA.

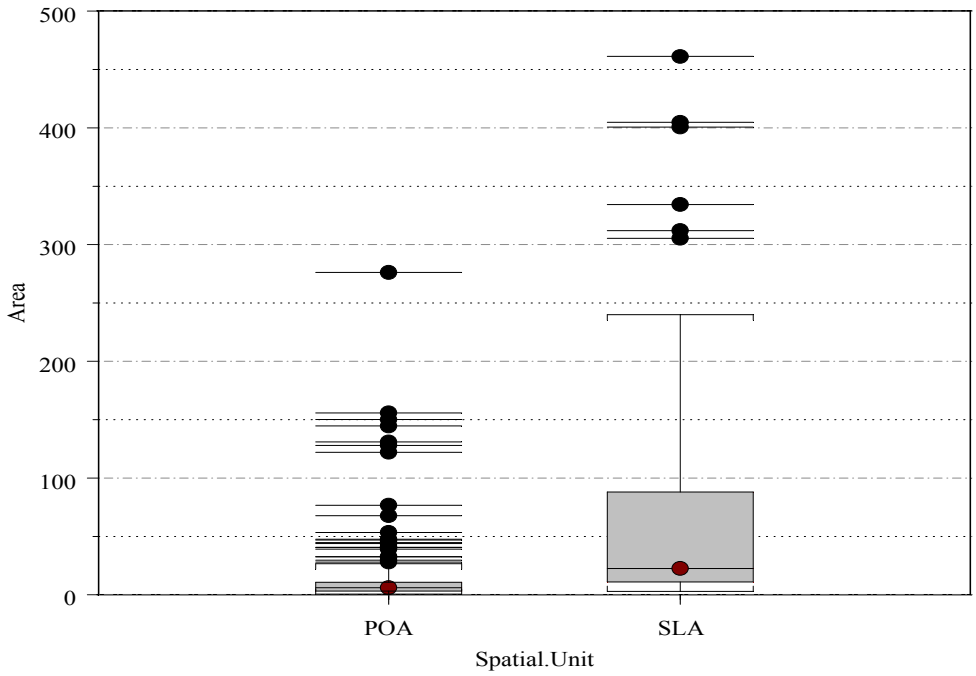


Figure 6: Box Plots of Area (km^2) by Type of Spatial Units

These box plots show that SLAs are much larger units with a median area of approximately 25km^2 as opposed to the POA median that is between 5 and 10 km^2.

Multiple Levels of Spacetime

The conceptual framework of ecological hierarchy theory (Allen and Hoekstra 1992) is used to understand the many levels of spacetime that phenomena operate at. The understanding of complex systems must take into account the many possible scales of analysis required by this multi-levelled framework. There is current development of a statistical framework for multilevel methods that can use data collected at different scales (Jones and Duncan 1996).

Aggregate-level studies will aggregate health events to the spatial unit in which they occurred, or in which the individual usually resided.

There is also a hierarchy of levels in time. The health event observations are aggregated to temporal units such as day, month, season or year.

Important features of this approach are: the recognition of a hierarchy in spacetime with patterns and processes organised at definable scales; the potential for different relationships to be observed between scales; and the implications of this for analyses of the health-environment complex (Elliott and Wartenberg 2004).

The development of hierarchy theory in ecology has focused on dealing with the issues of observing and analysing multi-scaled spacetime. It has been suggested that "time and space play something of a zero-sum game for control of our conceptions; tighter constraint in time appears to relax the constraint in space, and vice versa" (Allen and Hoekstra 1992: 205). This means that spatial analysis is hampered when attempting to discern dynamic spatial patterns through time. This is why larger temporal units are used for aggregation.

This framework is described in figure 7 (adapted from Kawachi 2004). The individual exposure is denoted by a lower case x and area level exposure measurements as capital X, with a range of possible levels of aggregation of exposure in between. Individual health outcomes are denoted as y and aggregate-level health indicators are Y. The hierarchy of possible levels of aggregation are shown by the different sized ovals. Capital X and Y are subject to scale effects and must therefore be treated with caution (Jones and Duncan 1996, Diez Roux 2004, Subramanian 2004).

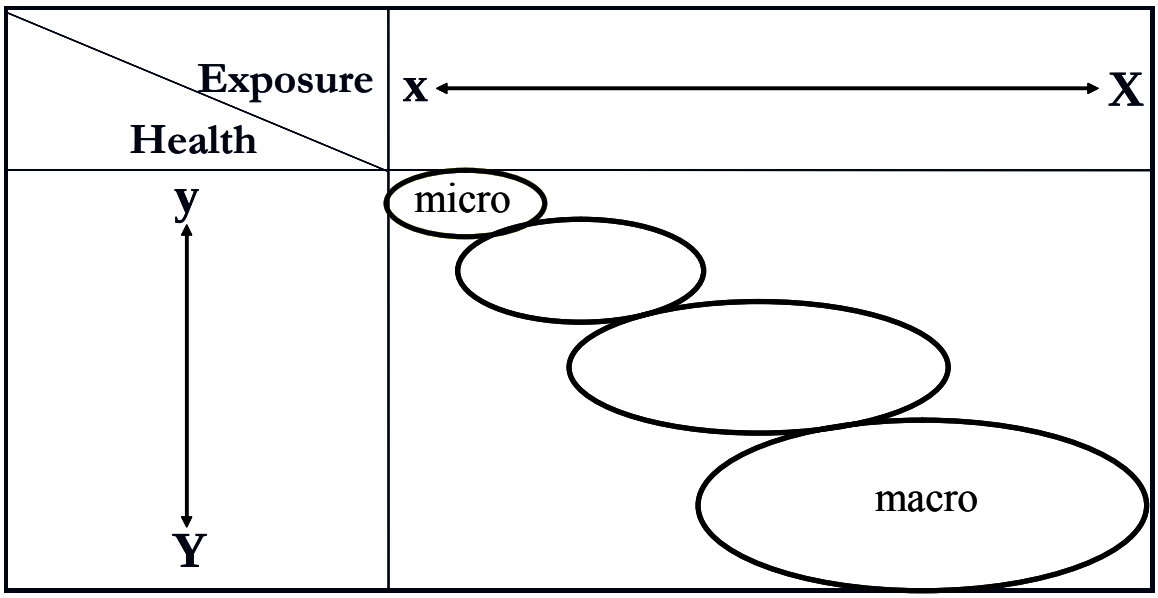


Figure 7: Conceptual Framework for Multilevel Models (adapted from Kawachi, 2004)

Aggregate-level and ecological studies in epidemiology involve "the investigation of group-level relationships between environment and health, by analysing spatial and/or temporal variations in exposure and health outcomes" (Nurminen and Briggs 1996:94).

(Nurminen and Briggs 1996) describe three ecological study approaches that can be characterised as: seeking associations between exposure and disease outcome in a population at a time; analysing relationships between exposure and health outcome compared in two or more populations with different exposure; and analysing time trends within an ecological design. The relationships between exposure and health outcome are assessed by following changes in exposure and rates of disease within a population over time. This study analysed multiple subpopulations that differed in their exposures to SES and AP at four consecutive time points.

The spatial units used for aggregation-level studies are often administrative (and sometimes have multiple distinct parts). The need to conduct aggregate-level studies stems from the difficulty of obtaining individual-level data on environmental exposures, the requirement to protect the confidentiality of participants; and to avoid stigmatising areas (Rushton 1998, Armstrong et al. 1999, Nurminen and Nurminen 2000). There is also a need to contrast with individual level results. Analysts may use different scales seeking gains and losses in the focus on phenomena when aggregated (Dungan et al. 2002).

Methodological issues that need to be considered in the design and analysis of an aggregate-level epidemiological study include selecting areas with populations that are: homogenously exposed; represent different extremes of exposure distribution; are comparable with respect to key population attributes (such as population size); and relate to the smallest units possible (Nurminen and Briggs 1996: 94-101, Nurminen and Nurminen 2000).

This last point about unit size is very important; indeed Brigg's (1997) suggests that:

> GIS data should only be used for applications for which their scale and resolution are appropriate. Too large a scale [small areas] is inconvenient and costly while too small a scale [large areas] (or at insufficient resolution) adds uncertainty and error to analyses, and may generate false conclusions (Briggs et al. 1997: 150).

The present study takes two scales: POA and SLA because of their availability for health research as well as the need to assess the issues of scale and zonation for aggregate-level analyses.

The Modifiable Areal Unit Problem

A description of the effect of scale on the statistical analysis of aggregate level health-environment relationships is the Modifiable Areal Unit Problem (MAUP). The problem was first identified by Geike and Behle in the 1930s (Marceau 1999) and defined by Openshaw and colleagues in the 1980s (Openshaw and Taylor 1981, Openshaw 1983). These problems mean that: when scales of observation or analysis change, such as the unit size, shape, spacing or extent, statistical results must also be expected to change (Dungan et al. 2002).

The Scale Issue

The term scale is used in ecology to refer to the levels in spacetime at which phenomena operate or are observed, and analysis of these (Dungan et al. 2002). There is a 'natural scale' for ecological phenomena. The scale issue occurs in ecological analyses performed at levels of aggregation that may not be at appropriate scale for that phenomenon. There may be different statistical results at different levels, and sometimes the opposite relationship may be found across scales. This effect is shown in figure 8 in a graph that compares the correlation coefficients for many variables computed using individual-level data with those found when the variables are correlated using aggregate-level data (Wrigley et al. 1996:29).

This S-shaped scatter graph shows a contrived comparison of correlation coefficients. This example draws on empirical evidence from various studies (Openshaw 1984, Steel and Holt 1996 and Holt 1996 cited in (Wrigley et al. 1996). In 1984 Openshaw was able to compute 780 pair-wise correlations between 40 variables at individual household level and at census district level for the city of Florence in Italy. These are compared in the scatter plot and the diagonal line represents the value if the analyses at the two levels agree. The results of this indicated that in general the aggregate-level correlations were larger in absolute magnitude (both positive and negative) than the individual-level correlations and that there is the potential for shift in sign between these correlations (Wrigley et al. 1996: 27-29).

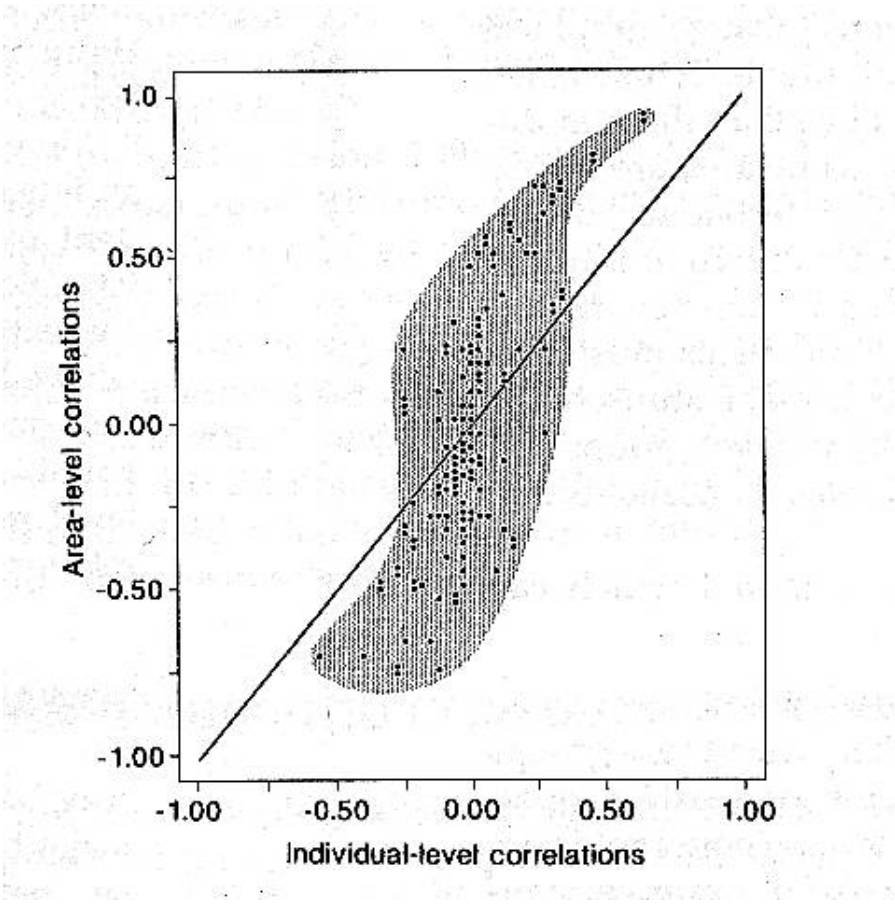


Figure 8: The Scale Issue (from Wrigley et al, 1996)

There are many ways to define the scale and zonation of the aggregate-level units and each of these would produce a different version of this scatter-plot. Because aggregation reduces variation in measures of the phenomenon there is a question about the meaning of scales of different size. Therefore this study examines this issue using the POA and SLA levels.

The Zonation Issue

The zonation issue refers to the "variation in results obtained from different ways of subdividing geographical space at the same scale" (Green and Flowerdew 1996: 41). An example of this is shown in figure 9 in which alternative aggregations of John Snow's point data of cholera cases in London produce different spatial patterns (Monmonier 1996). The point location data of cases when aggregated into different spatial units obscure and distort the understanding that Snow gained from plotting the residential locations of the cholera cases and correlating this with the water supply. Thus the zonation issue would have obscured the important spatial pattern that indicated a relationship between the cholera cases and the polluted water from the Broad Street pump.

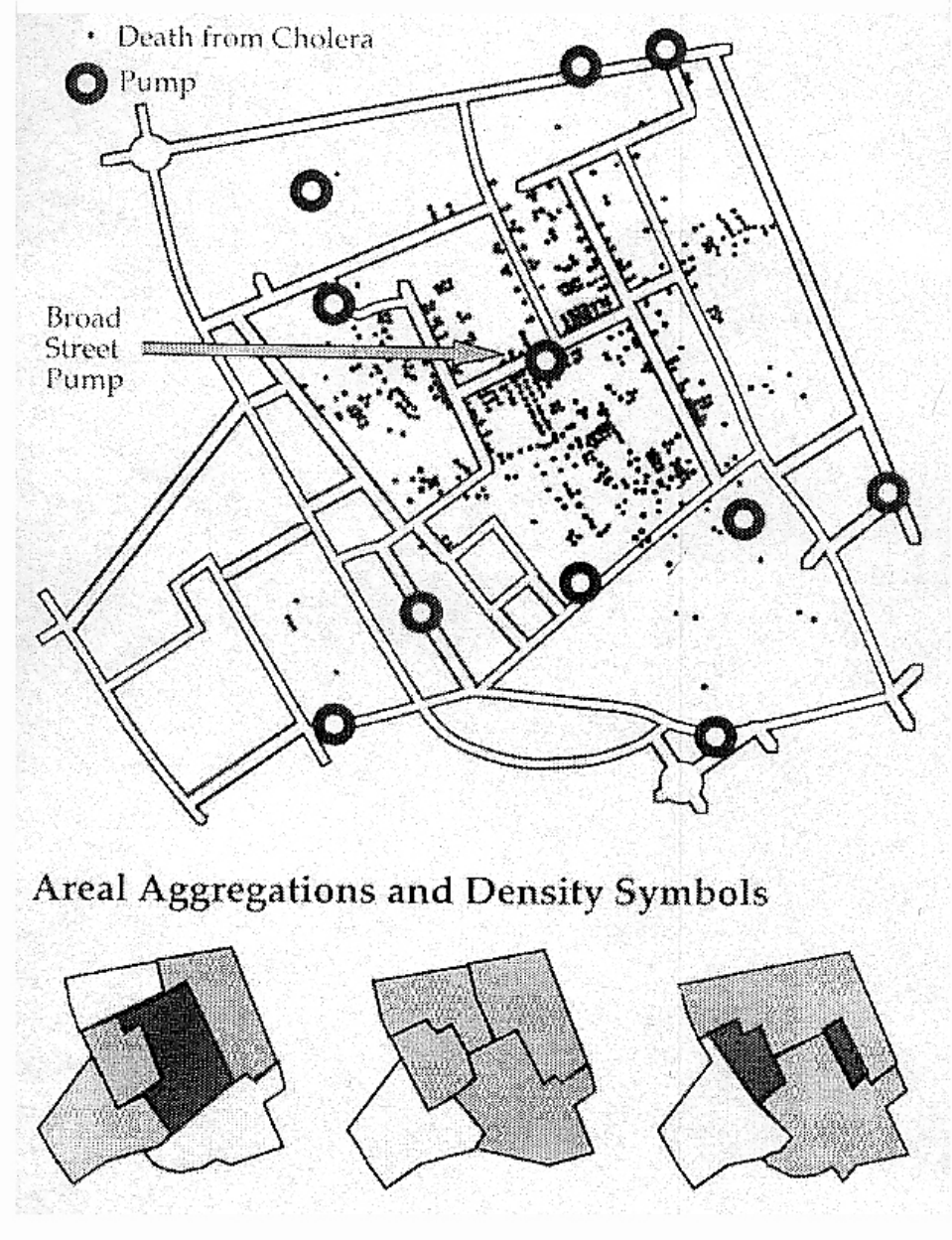


Figure 9: The Zonation Issue (from Monmonier, 1996)

The placement of boundaries is therefore a very important consideration in the construction of spatial units and the assessment of their usefulness for geographical analysis. Spatial analysts need to assess this issue in the observation and analysis stages of research (Dungan et al. 2002).

The Ecological Fallacy

Openshaw (1983) discusses the relationship between the MAUP and the ecological fallacy. As mentioned above an ecological fallacy occurs when it is assumed that results based on aggregate-level data can be applied to the individuals that comprise these (Openshaw 1983)

A more appropriate term might be the cross-level fallacy which would also include the opposite phenomenon: the atomistic fallacy, where incorrect inference is drawn about groups based on individual-level associations (Blakely and Woodward 2000). The concept can then apply to relationships observed at any level of the spacetime hierarchy compared to those at other levels when these are not the same (Nurminen 1997, Blakely and Woodward 2000, Pearce 2000, Lancaster and Green 2002).

In a geographical context the individuals can either be people who form the aggregate zones or groups, or the zones themselves prior to a subsequent aggregation. Whether the ecological fallacy is a problem depends on the nature of the aggregation being used. A completely homogenous zoning or grouping system would be free of this problem. However, as Openshaw (1983) points out:

> Most, if not all, zoning systems studied by geographers are internally heterogeneous so that the severity of any ecological fallacy problem depends largely on the nature of the aggregation units being studied (Openshaw 1983: 8).

For this study this means that POA and SLA may give contradictory views of the same space at the same time, thus obscuring understanding of the meaning of the results.

Consider the possibility of analysis of large areas with differing levels of air pollution. These areas may display a positive association between pollution level and CVD mortality rate. It is conceivable though that the individuals who die are those who - within their area - are exposed to the *lowest* level of ambient air pollution. They might be exposed to indoor chemical contaminants, or smoke more cigarettes. These are other factors in the ecology and must also be considered when defining important causal mechanisms amongst many factors.

An example of this phenomenon applied to spatial units is where large areas show a positive relationship between pollution and mortality rates while at a finer spatial scale it is those which have the lowest level of pollution that have the highest mortality rates. The inference made about the influence of pollution from the results at the larger level of aggregation is therefore fallacious when applied to the smaller level units that comprise them.

There is a body of geographical literature that identifies an effect of aggregation on the results of the correlation and regression method. Meade and Earickson (2000) discuss the example of a multi-scaled study of cancer and point out that:

> Correlations are especially affected by changes in aggregation. Data are often aggregated by an independent variable. This inflates correlation coefficients but usually does not affect regression coefficients….. [therefore] report regression coefficients and treat correlation coefficients with care (Cleek 1979 cited in Meade 2000: 440).

However regression results have also been found to be affected by spatial aggregation with "wide variations in goodness of fit… [and] major variability in the regression coefficients obtained" (Fotheringham and Wong 1991 cited in Green, 1996: 42). Therefore results from correlation and regression analyses must be interpreted with this issue in mind.

Methodology

Spatiotemporal Framework

This study uses a spatiotemporal framework. This can be described as a progression: starting with non-spatial exploration of the health and environmental data such as temporal trends and frequency distributions for example. Then assessing the spatial patterns at particular times (processing, linking and overlaying the different forms of data). Finally the analysis assesses the attributes of all the areas at multiple specific times.

Observations are made of processes in space and time together and the attributes of these observations are inextricably linked.

Non-spatial analysis of the frequency distribution of various cause-specific mortality data reveals a relatively heavy burden of CVD and some disparity in rates between areas of the city.

Temporal analysis shows patterns across time but the scale of temporal units is important. In this analysis assessment of daily rates indicates that six-month seasons representing summer and winter are appropriate. This is because the seasonal variation of CVD mortality rates shows a cycle operating at this frequency. The AP concentrations supplied by the NSW Department of Environment and Conservation are sampled by the minute at points in the monitoring network and these data were used to describe the spatial patterns during these six-month seasons.

Spatial analysis then identifies the form of spatial patterns (shape and location), SES and AP attributes. This stage must establish the scale of these features (i.e. clusters and hotspots of disease or regional trends in ambient pollution concentrations). Then there are issues of linkage of data between different mapping systems. In this analysis three-dimensional AP surfaces were converted to the spatial units for which the health and SES data are available.

The spatiotemporal modelling explores spatial patterns through time by linking spatially organised data at multiple consecutive temporal units. Then the analysis investigated the relationships between these variables.

Environmental epidemiology studies cannot expect high levels of explanation using only a few variables. Because simple models of complex systems only measure part of the important processes therefore they can only expect partial explanation of the patterns.

The age standardised CVD mortality rate for the areas at each time point are assumed to be a function of the AP variables and the SES variables at that time. In this multivariate data mining approach the two summers, two winters, AP variables and SES indexes are analysed together. Then the model is stratified for simple linear regression in which the standardised rate of the areas in winter was regressed against the disadvantage index, and then the rate of areas with low disadvantage index scores in winter were regressed against the PM_{10} concentrations.

Assumptions

It is assumed that the location of usual residence at time of death indicates the person's exposure to ambient AP levels or their experience of SES at temporal scales relevant to CVD ecology (for instance the intensity, duration and frequency of exposures at the area level are experienced by the people living there). This is not really true as people will move around (and out of) the city through the days, weeks

and years of their life, but the paucity of data is a reality of this situation. The older age of the individuals who are dying with CVD will mean their daily movements may be minimal, however their previous residential and occupational locations may have a large bearing on their health and this information is not available.

It is also assumed that the area of usual residence codes for each individual is correct. There were some coding errors identified and these were addressed in the construction of the dataset. Any remaining coding issues are assumed to introduce low-grade random error that will not bias the analysis.

Another assumption that needed to be made is that the relationships are the same in meaning at the two levels of aggregation. This was necessary for it to make any sense to test whether they are the same in magnitude. In the context of the scale and zonation issues discussed above this is an important assumption that is justified only if the two types of spatial units reflect similar levels in the spacetime hierarchy. Clearly the analysis would not be appropriate for individual-level and area-level relationships because of the difference in scale.

To briefly summarise the methodology section: this study is interested in the influence of SES and AP on CVD mortality and the geographical methods for understanding the spatial patterns of these relationships. The health effects are multifactorial and interactive and need to be studied at appropriate scales in time and space, with due consideration of the issues of scale and zonation for spatial analysis. The approach taken here was explorative with non-spatial analysis of temporal patterns, spatial analysis of the geographic locations of CVD, SES and AP attributes of areas and spatiotemporal analyses for all the areas at the different times.

Data

Health Data

The health data used in this study are age standardised mortality rates calculated using the direct standardisation method. The population data were obtained from the ABS for the residents of Sydney at the 1996 and 2001 census. The mortality data includes residential location at time of death aggregated to the spatial units described above.

The NMD is collated by the state Registrar of Births, Deaths and Marriages who send the files to the ABS. Any means of identifying individuals is removed and they are sold to research institutions such as the National Centre for Epidemiology and Population Health (NCEPH). The NCEPH copy was accessed for this study after gaining ethical approval from the ANU Human Research Ethics Committee. The ethics approval was granted provided that the data presented in maps is aggregated across time to avoid the identification of small areas with unusually high rates (National Health and Medical Research Council 1999). Therefore the maps use broad class breaks and averages across multiple seasons.

First the mortality records for deaths occurring between 1995 and 2002 were extracted as this is the period for which POA and SLA are both used in the dataset. The seasonal variation noted in figure 1 can be seen from an alternative perspective that gives the seasons an interannual context in the contour plot in figure 10. This shows the monthly average rates on the y-axis and the year on the x-axis, with interpolation between these data points and isopleths marking months with similar levels of CVD mortality across years. The seasonality was addressed by taking six-month seasons to classify this pattern (delineated by the horizontal dotted lines). During this period the monthly averages of CVD daily rates are generally higher in the cooler months from May to October and lower from November to April.

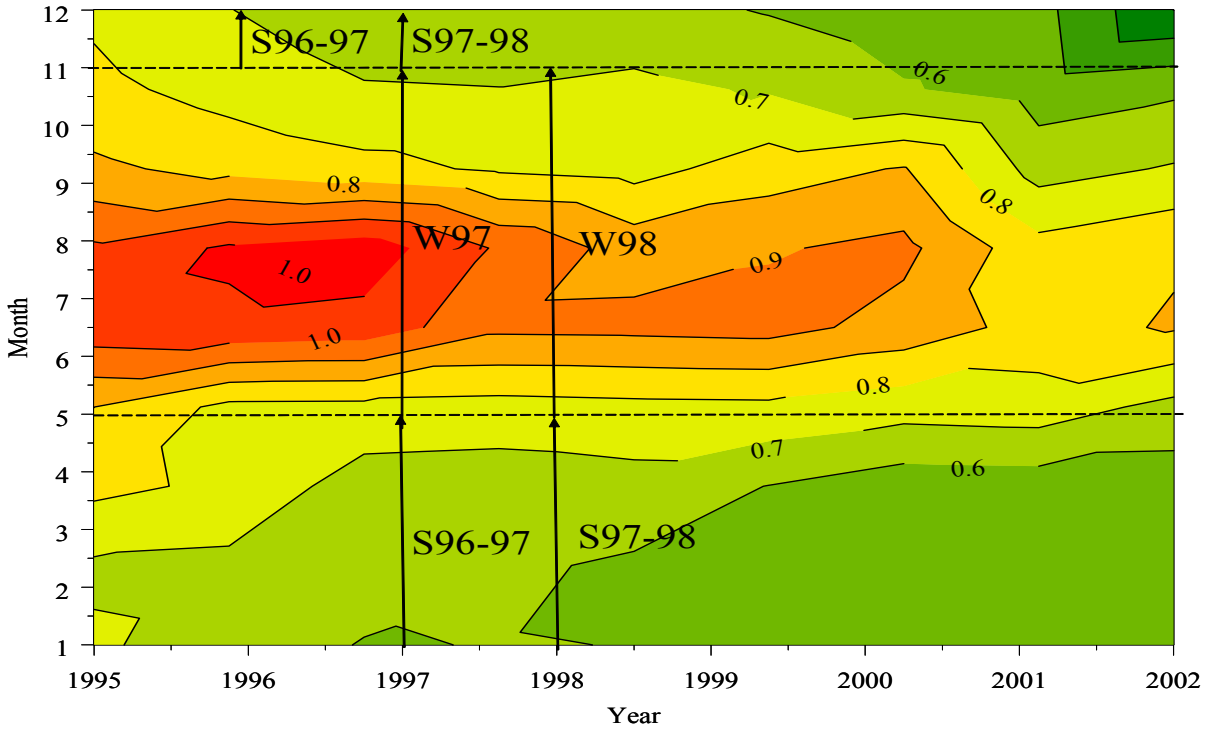


Figure 10: Monthly Averages of CVD Crude Rates per 100,000 from 1995-2002.

The winter peak can be seen in red and is consistent through the early period until about 2001, when the level drops and the interannular periodicity weakens. The four six-month seasons between November 1996 and October 1998 were selected and are shown by the vertical black arrows.

These seasons were defined after assessment of trends in the concentrations of the three pollutants (appendix 3), in particular PM_{10}. The study period was selected primarily because of the interesting interannual variations depicted in figure 11 (which covers 1994 as well as the 1995-2002 period covered by the mortality data as these extra periods were available and the extra processing time was minimal).

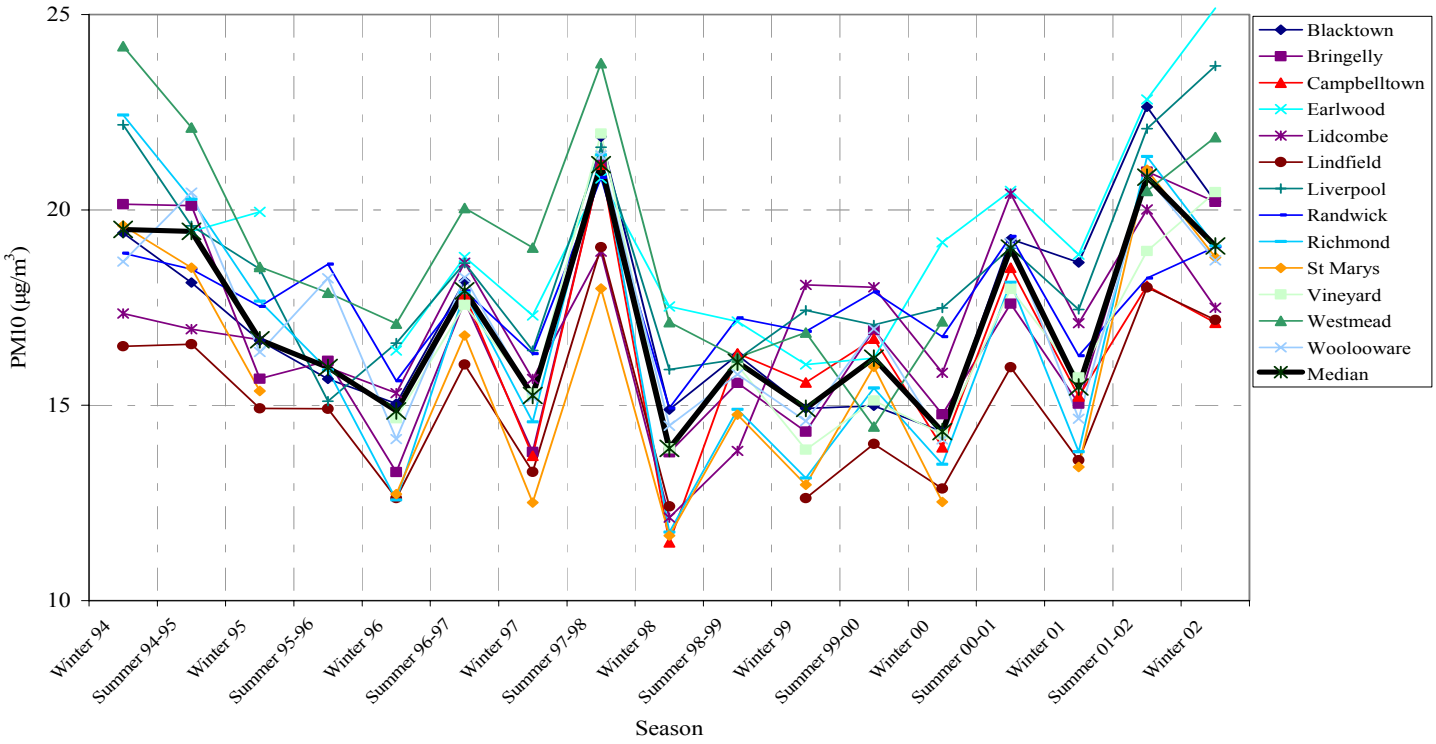


Figure 11: Six-Month Averages of Daily Average PM_{10} ($\mu g/m^3$) from 1994-2002

The seasonal difference is striking in the 1996 until 1998 period when compared with the uniformity across the seasons in the early and late parts of the series. This period is also considered suitable with respect to the SEIFA indexes, which should not have changed, greatly since quantified in the 1996 census.

Age Standardised Mortality Rates

Age standardisation is very important when comparing sub-populations that have different age structures. There are two common methods for age standardisation: direct and indirect

The direct standardisation method was chosen because it is considered more valid for inter-group comparisons whereas the indirect method is only valid for comparing the study population with the standard (Taylor 1998:275-276, Meade and Earickson 2000, Julious et al. 2001). The method is discussed more in appendix 2 and compared to the indirect method which is also used extensively in spatial epidemiological studies. It is noted that this method can produce extreme rates where populations are low. This is a phenomenon called rate instability and methods are available to adjust for this (Bailey and Gatrell 1995, Taylor 1998, Julious et al. 2001, Anselin 2003). However the CVD category has large numbers of deaths in each area and so should be less influenced by this issue.

The methods for adjusting rates require assumptions be made and so the exploratory analyses use unadjusted rates in order to be as assumption-free as possible. Analyses were then conducted that do adjust the rates using the variance of the estimates as a weight in the regression and excluding the upper outliers. These methods are discussed in the sequence in which they were conducted.

To compute the directly age standardised mortality rates, Age Specific Rates (ASR) for each age group in each area are first calculated. This is the number of deaths in an age group divided by the population in that age group. These rates are then applied to the population of this age group in the standard population (which was the total usual resident population of the Sydney SD at the 2001 census).

It should be noted that as the total population of the census year is used as the denominator for the six-month seasons, these rates are not comparable with other studies that calculate annual rates using the twelve-month period as the numerator. Another method available for these analyses is to calculate the rates using the person-years-lived, which in this case would mean dividing the population of Sydney in half (or even more precisely person-days-lived by dividing the population by 365 and then multiplying this by the days in each six-month period). This was not done here because of time constraints.

The direct method calculates the "expected" number of deaths that would have occurred if the standard population had the same mortality experience as the study area. Then these expected cases for each age group are summed and divided by the total population of the standard. This rate is multiplied by 1000 and this is the study area's age standardised rate. As an equation this is:

$$\text{Standardised rate}_j = \frac{\Sigma (ASR_{ij} * X_i)}{\Sigma X_i} * 1000$$

Where i is the age group, j is the area, and X_i is the population of age group i in the standard.

Study Region

The study region was defined to exclude areas outside the urban boundary of Sydney. The extent of the region selected can be seen in the maps in figures 3 and 4. There were 212 POAs selected from the 1996 ASGC (metadata in appendix 5). The POA were selected using criteria that controlled distance allowed between parts of the same code (see appendix 1), thus the exposure misclassification introduced by the multiple parts was minimised. The geographical centroids of each part and the distance between these were calculated. POAs with any parts greater than 5km distant were excluded from the study region. The POAs in the study region have a combined resident population in 1996 of 3,190,657.

The selected POAs where then overlayed with the SLAs boundaries (spatial metadata in appendix 5) and 40 areas were selected that had good agreement between these as can be seen in figure 4. There are 40 SLAs (from ASGC 1996) in the study area with a combined population of 3,258,351: a discrepancy of approximately 70,000 people or 2%.

These SLA codes were used to extract the data from the usual residence codes. These record's Postcode of usual residence were then assessed and it was found that some were coded to areas outside Sydney (as far removed as western NSW – see appendix 1).

There are some portions of SLAs not covered by POAs and vice versa. Thus some SLAs have more area than corresponding POA while a few POA are located outside the boundary of the SLA based study region. The study population was not adjusted for this study and there is some error introduced.

Summary of Health Data

The four seasons are delineated by the vertical dotted lines in figure 12 with more deaths in winter 1997 than 1998, while the two summers were very similar.

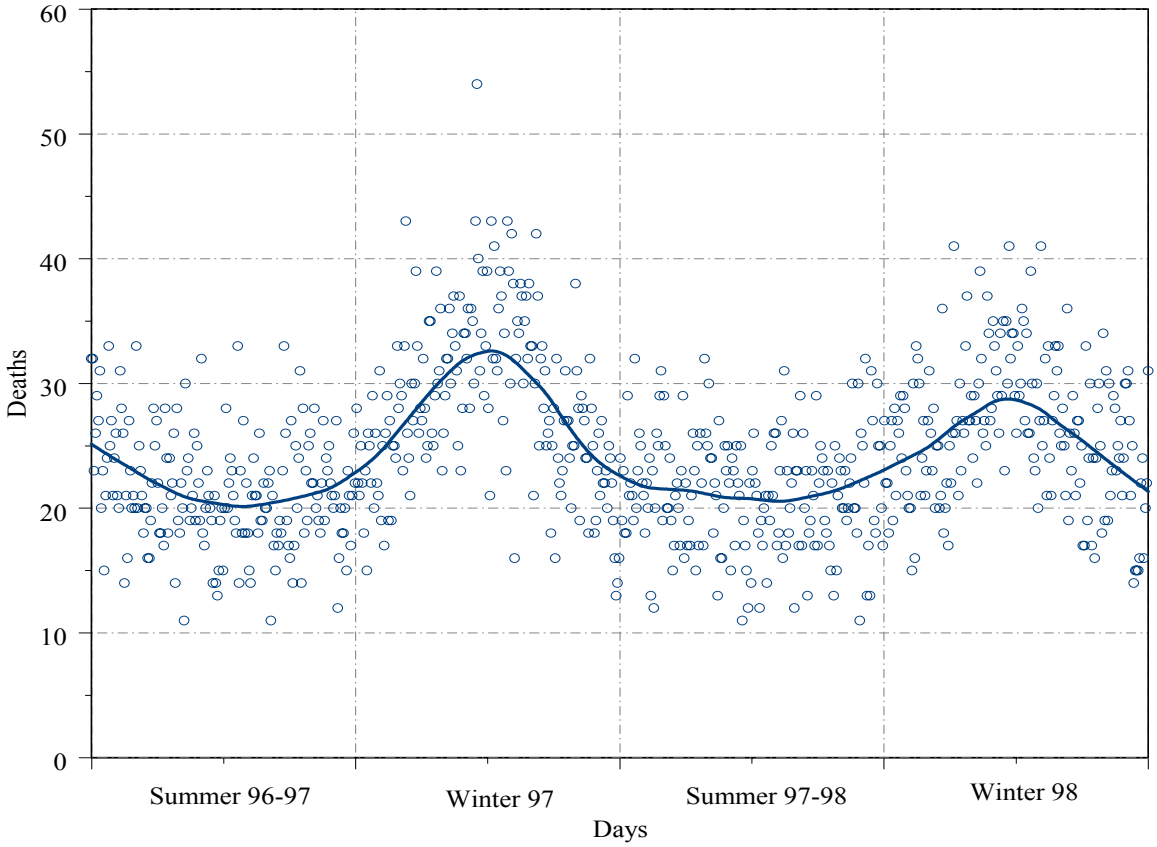


Figure 12: CVD Daily Deaths for the Study Region and Period

The six-month totals for the study region are shown in table 1.

Table 1: The CVD Mortality in the Region and Four Periods

Season	Number of deaths from CVD (n = 17,766)
Summer 1996-97	3,875
Winter 1997	5,225
Summer 1997-98	3,844
Winter 1998	4,818

Socio-Economic Status

SES in Sydney can be conceptualised as broad regional trend with some local clustering. The four urban indexes from the ABS Socio-Economic Indexes for Areas (SEIFA) are used for this study (see metadata in appendix 5). These are the Relative Disadvantage, Economic Resources, Education Occupation, and Urban Advantage. The indexes are all constructed so that the higher values indicate

higher advantage levels. This means that the disadvantage index gives the least disadvantaged areas the highest scores.

These indexes are derived from principal components analysis of census variables conducted for CCDs and then aggregated to other areas in the hierarchy. McCracken (2001) argues that it is too coarse an indicator for use in health studies and that other census data such as income may be more useful (McCracken 2001). However there are a number of other studies using the SEIFA indexes for geographical analyses of health inequalities (Hyndman et al. 1995, Yu et al. 2000, Glover et al. 2004).

The indexes are often defined into quantiles based on population. As the exploratory tree models are relatively assumption free and non-linear they should find adequate groupings of the index scores without imposing a classification rule such as this. The split points between SES groups identified in the tree models are then used to stratify these data.

Air Pollution

Hourly AP concentrations were obtained for each site in the Sydney monitoring network, which is shown in figure 9 (NSW Department of Environment and Conservation 2001).

The pollutants available include: coarse particulate matter (PM_{10}); fine particulate matter ($PM_{2.5}$); ozone (O_3); sulphur dioxide (SO_2); carbon monoxide (CO); nitrogen monoxide (NO); nitrogen dioxide (NO_2); and oxides of nitrogen (NOx).

Not all of the sites measure each of these pollutants and some have more complete records than others. The criteria that were decided to be used for selecting pollutants are that they: have relationships with CVD identified in the literature; measurement at a large number of sites and a relatively complete record for the period. The number of sites monitoring each pollutant is crucial to the spatial modelling and meant that some pollutants, although associated with CVD and with relatively complete records by site could not be used. For instance, $PM_{2.5}$ has been strongly associated with CVD (Morgan et al. 2003) however only a few sites measure this pollutant and so interpolated surfaces could not be computed. The pollutants selected were PM_{10}, NO_2, and O_3.

The Sydney airshed is connected to the Illawara airshed to the south and the Hunter in the north (McPhail 1996) and these exhibit complicated air movement patterns. Consequently there can be quite different patterns of pollution from day to day and from season to season (Shepherd 1996). The development of common seasonal conditions that give rise to different patterns of air pollution can be characterised in terms of seasonal types. In winter the airflow patterns around Sydney are predominantly local drainage flows down the valleys in the north and to the west of the city centre. There may also be a major flow off the Blue Mountains as well that overlays these local flows (Shepherd 1996). In summer the air movements are more dynamic, starting to move early in the day. Air moves down towards the bottom of the Sydney basin and flows eastwards out over the coast. The air will stay offshore and photochemical processes will produce smog, and with the advent of the sea breeze (usually a north-easterly) the polluted air starts to push back over the city and down towards Campbelltown in the south-west (Shepherd 1996). This air movement might be repeated over a few days and there may be a substantial build up of pollutants as a result. The general pattern of air pollution over Sydney therefore

shows higher concentrations of traffic related pollutants like NO_2 and PM_{10} in the lower parts of the basin while O_3 will by higher at the periphery of the city.

The PM_{10} levels seen in figure 11 are higher in summer than winter and this may be the influence of drought related dust or bushfire smoke from the surrounding landscape. Due to the meteorological conditions in the basin, winter levels of NO_2 are higher than in summer. O_3 is very reactive and therefore doesn't last long in the air before reacting with other chemicals. The accumulation of O_3 requires substantial photochemical reactions, which require solar energy. Therefore O_3 production is highest in the summer months (see temporal and spatial distributions of each pollutant in appendix 3).

Six-monthly averages were calculated from the hourly pollutant concentrations using the pollutant specific averaging methods from the National Environmental Protection Measures (NEPM) (National Environmental Protection Council 2001). These are shown in table 2.

Table 2: The National Environmental Protection Measures for Air Pollutants

Pollutant	Averaging period	Maximum concentration
Nitrogen dioxide	1 hour	12 pphm
	1 year	3 pphm
Photochemical oxidants (as ozone)	1 hour	10 pphm
	4 hours	8 pphm
Particles as PM_{10}	1 day	50 $\mu g/m^3$

Air pollution surfaces were constructed using tension splines in ArcGIS. Only sites with greater than 80% of observations in the period were used. No effort was made to adjust these surfaces for boundary problems that arise at the edge of three-dimensional modelled surfaces and this may cause exposure misclassification for the outer areas of the city (see appendix 3). As the same surfaces were used for both types of spatial units this should not affect the comparison between these.

Exposure Estimation

The AP surfaces were overlayed by the POA and SLA boundaries and exposure estimates were computed. The SLA estimates are population-weighted averages because of the large areas and thus the likelihood of exposure misclassification. To do this, multiply the areal average at CCD by the population in that area, then sum this within the SLA and then divide by the population of that SLA. The POA estimates were not weighted because they have much smaller areas and they are therefore less prone to this problem.

The PM_{10} monitoring sites use the Tapered Element Oscillating Microbalance (TEOM) instruments and this method has an issue relating to underestimation at low temperatures (Ayers et al. 2001). This study follows the advice in option four of the NEPC peer review committee that suggests using un-adjusted data as long as it is clearly noted that the instrument will underestimate PM_{10} concentrations at low temperature, particularly below 15 degrees Celsius (NEPC Peer Review Committee 2001). The effect on this study will be to underestimate the PM_{10} exposure in winter, making any relationships more difficult to observe (a type two error that protects against the false identification of a relationship). The effect of this error will vary between areas with different temperature profiles; it is expected to be slightly

worse away from the coast where it gets colder - so the south west will be under represented compared to inner suburbia.

In particular NO_2 is known to vary within hundreds of metres of roads (Briggs et al. 1997, Briggs et al. 2000, Zhu et al. 2002, Gilbert et al. 2003). A buffer of the roads network could be used as a proxy for fine scale variation in NO_2 but this would be difficult to link with the SLAs as each area would have many roads inside its borders whereas it would be more likely to find a POA without a road in it and compare these with roadside areas.

The process that produces O_3 means that peak concentrations are found in the periphery of the basin. This area has the sparsest coverage by monitors so there is high uncertainty regarding the exposure estimates based on the O_3 interpolations.

In summary of the data section, the dataset consists of eleven variables constructed in comparable ways for the POAs and the SLAs in the study region during the four consecutive seasons from summer 1996-97 through to winter 1998. The dependent variable is the directly age standardised mortality rate and other variables include Disadvantage, Economic Resources, Education Occupation, Urban Advantage, PM_{10}, NO_2, O_3, population density, type of season, and the position in seasonal sequence.

Analysis

The fact that interactions have been found in the relationships between SES and AP exposure and their health effects means that this analysis must use assumption free data mining techniques that are sensitive to the variation in the data. Also, an iterative statistical methodology for exploring the different aspects of these complex relationships is applied.

Tree-based models are used to first explore these data with as many variables as were available. The regression tree method is a flexible, non-parametric, and non-normal approach to multivariate models (Breiman 1984). Stratified Linear Regression were then used to quantify the relationships while controlling for variables that might also influence CVD mortality (Shannon et al. 2002).

Regression Trees

Regression tree modelling is a method for recursive partitioning of datasets (Breiman 1984). Tree models partition the dataset based on key variables that reduce the deviance in the subsets and places them in a hierarchy of importance. The data are successively broken into smaller, more homogenous groups. Splits are chosen that maximise the reduction in deviance of the "child" nodes compared to the "parent" node. The levels of each variable best defining the split in the data at each step in the partitioning procedure are identified and these splits form the nodes of the tree. The value of a split is measured as the reduction in the residual sum of squares and visualised in the graph by the length of the lines between the nodes. Splitting continues until the response values are all the same within an "endgroup" or data becomes too sparse for additional splitting.

This technique is able to explore complex relationships in a relatively assumption-free manner and are widely used in ecological modelling (Moisen and Frescino 2002, White and Sifneos 2002). The tree algorithm in the S-Plus software package was used for this analysis.

Stratified Linear Regression

Linear regression was used for analysing the relationship between variables and comparing these between the sets of spatial units. The linear regressions were stratified using the tree models to investigate the disadvantage and PM_{10} relationships in winter. The analysis was first conducted using unweighted standardised rates and this was followed by weighted regression because the linear regression model assumes constant variance for each area rate which is not true. The direct and indirect standardisation methods are compared in appendix 2. The least squares regression algorithm in S-plus was used for the analysis of the directly age standardised rates presented here.

The use of tree models to stratify linear regression enables control for covariates in the regression of an outcome onto a single predictor. It is assumed there are:

> Covariates defining subgroups that can be used to partition the dataset… for example, if $y_i = \beta_o + \beta_1 x_i + E_i$ for subjects less than a certain age, and $y_i = \beta_o^* + \beta_1^* x_i + E_i$ for subjects greater than that age, with the β's not necessarily equal, then partitioning a node based on an age cut point would allow one to fit the two regression models separately… these stratified linear regression models will have improved fit and interpretability over standard multivariate regression models (Shannon et al. 2002: 117).

The subsets of the data were used to fit linear regression models for:

a) Winter rates against the disadvantage index; and

b) Disadvantaged areas in winter for the PM_{10} concentrations.

Weighted Regression

It is widely accepted in environmental epidemiology that aggregate-level regression analyses should be weighted by the amount of information represented by each area estimate (Nurminen and Briggs 1996). However it is also true that:

> In environmental epidemiology, considerable natural variation in the data may be expected, even for data relating to short time periods and distances… such variation is an intrinsic part of the environment-health system and must be retained. Conversely, sampling and measurement errors must be identified, and either eliminated or assessed and controlled for…. [methods] should yield statistically valid and scientifically credible results… this means they should be unbiased and sensitive to the variations in the data at hand (Nurminen and Nurminen 2000: 68)

The results of the unweighted regression were compared with those weighted by the inverse of the variances of the standardised rates.

To calculate the variance for each age standardised rate at each time point, first the proportion of the age group in the standard population (that is the population in that age group divided by the total population in the standard multiplied by 1000) is applied to the counts observed in that age group in the study population (Clements 2005a) personal communication). Then this is divided by the square of the population in that age group in the study area. This is then summed for each area. Then the inverse is used as the weight in the regression model.

Expressed as an equation this is:

$$\text{Variance (age standardised rate}_j) = \frac{\Sigma W_i * y_{ij}}{\Sigma N_{ij}^2}$$

Where j = area, y = deaths, i = age, N = population, and

$$W_j = \frac{\text{Population in age group i}}{\text{Total population in standard}} * 1000$$

These weights are used by the least-squares algorithm in S-plus which minimises the sum of the squared residuals multiplied by the weights.

In summary of the analysis section, this study is focused on methodological development for understanding geographic patterns based on exploratory spatial data analysis of the available data.

Results

Standardisation

Age Specific Rates

The Age Specific Rates (ASRs) for CVD mortality in each area in Summer 1996-97 are shown in figures 13 and 14.

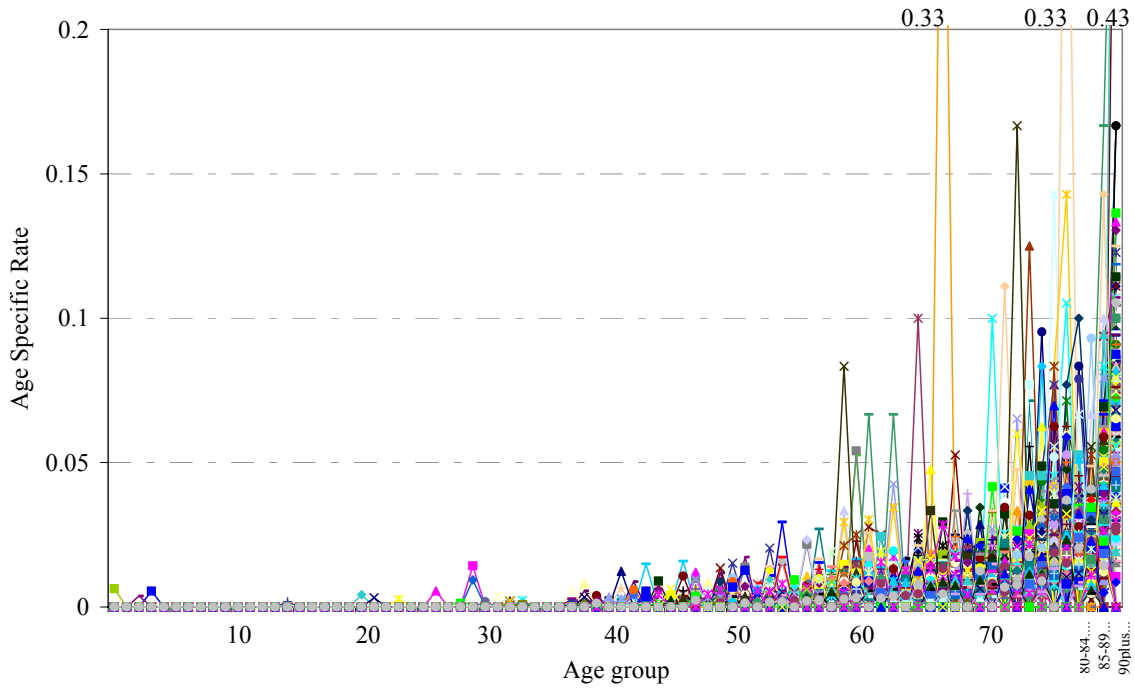


Figure 13: CVD Age Specific Rates for POA in Summer 1996-97

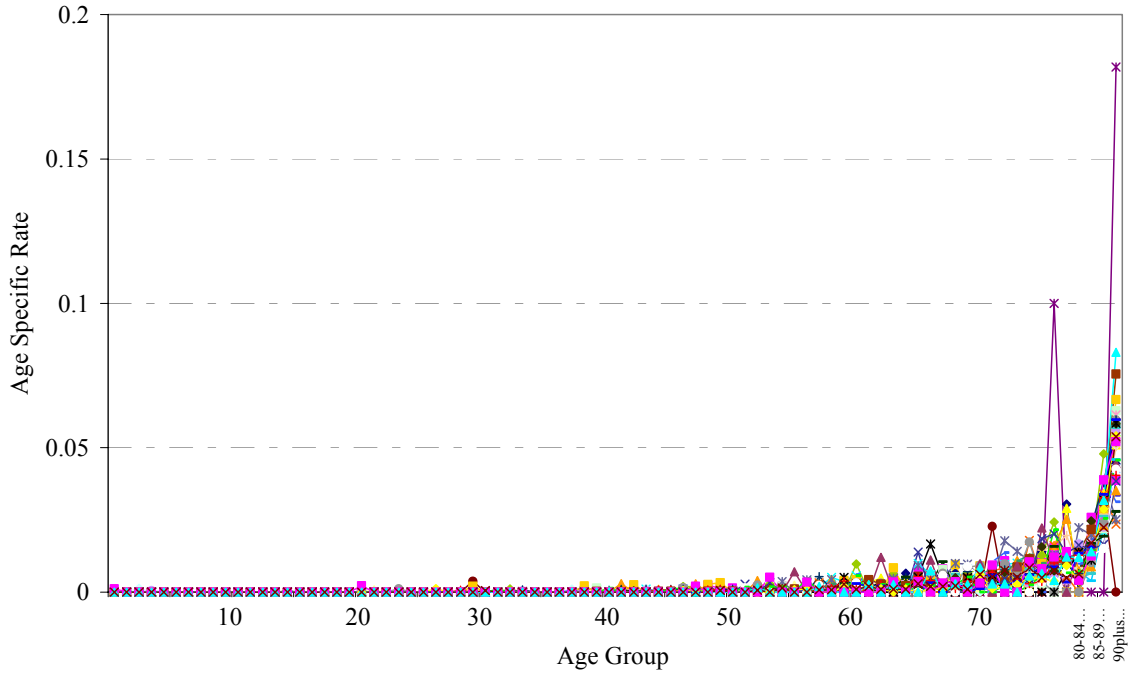


Figure 14: CVD Age Specific Rates for SLA in Summer 1996-97

There are some extremely high rates for a few age groups in certain POA and these are the source of the rate instability described above. There are only a few though and a lot of the areas appear to have ASRs within a certain range that is similar across all the areas which means they may not be affected by this instability.

The ASRs are more stable for the SLAs shown in figure 14. These show much more similarity, as well as lower rates overall.

Standardised Mortality Rates

The histograms in figure 15 and box plots in figure 16 show the variability between the CVD standardised mortality rates for the different types of spatial units. The variation apparent at different levels of aggregation is noteworthy.

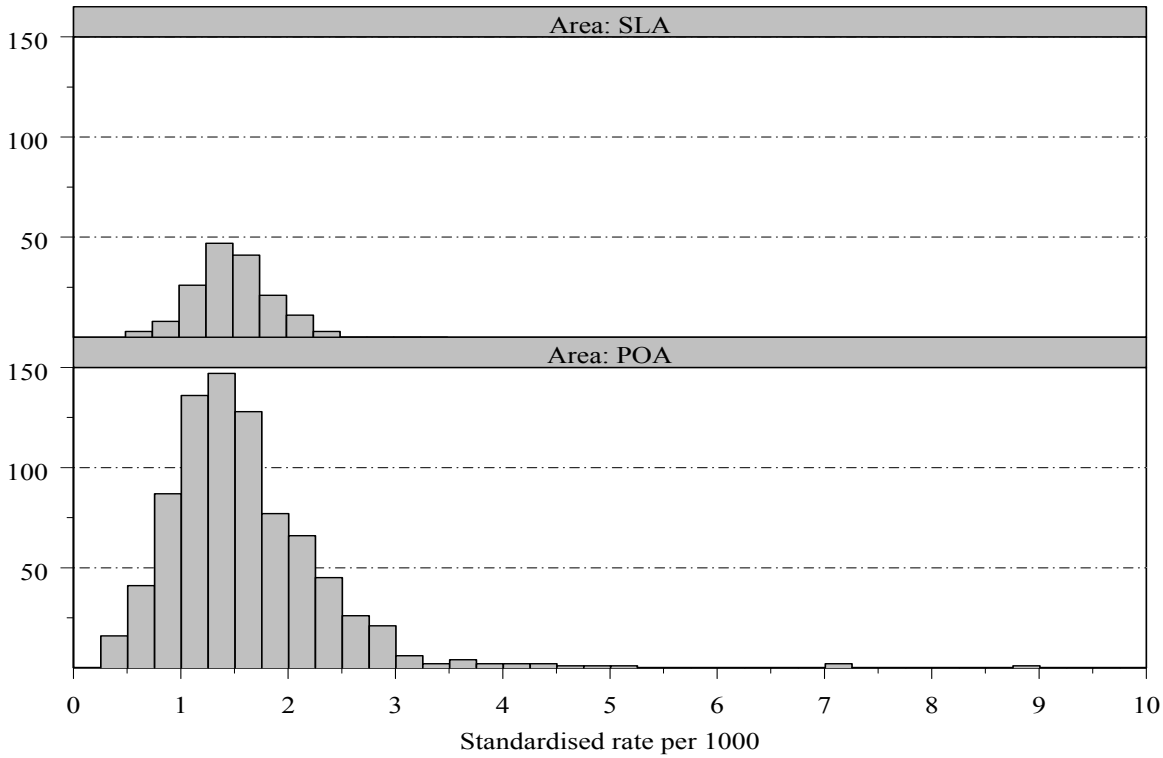


Figure 15: Frequency Histogram of POA and SLA Age Standardised Rates (all times)

There is clearly more variability in CVD rates observed at the postcode level than at the SLA level. The range of POA rates is wider than the SLA. The distribution of area rates in the histograms is essentially gaussian with the POA distribution skewed to the right.

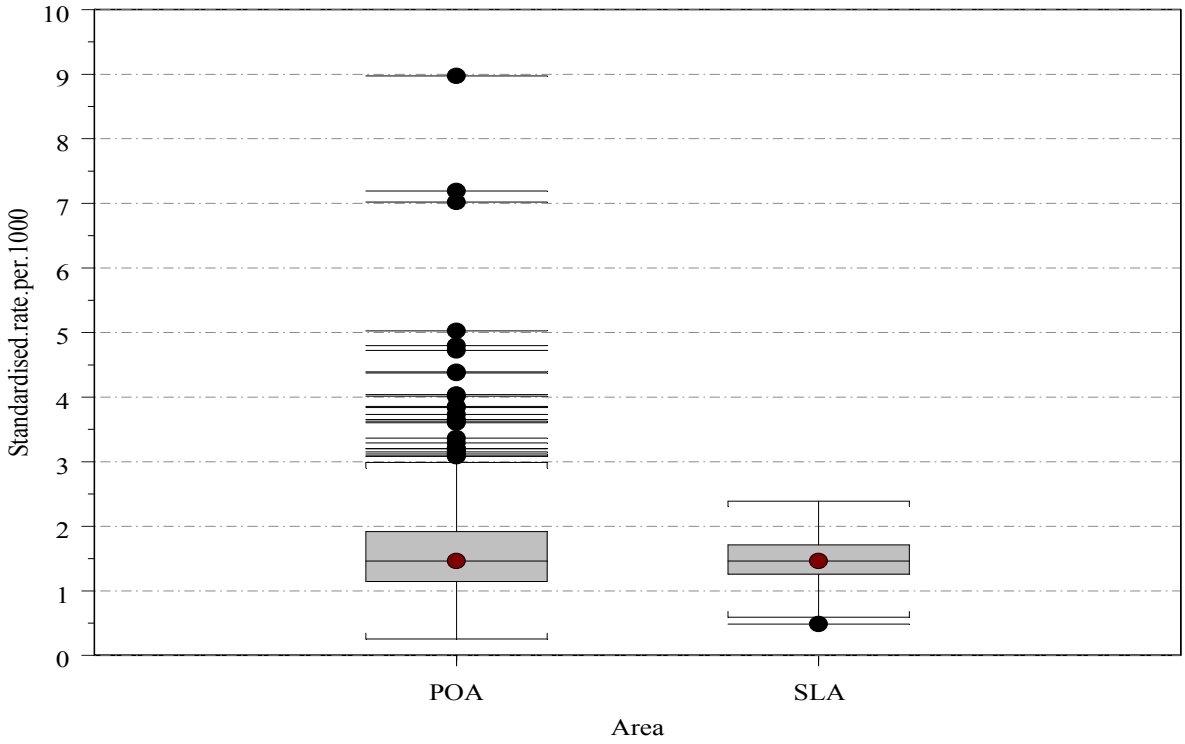


Figure 16: Box-plots of POA and SLA Age Standardised Rates (all times)

The median age standardised rate appears to be the same between the types of spatial units. However, the range of the two sets of spatial unit's rates differs substantially, showing the effect of aggregation on this measure. The three POA rates above 6 per 1000 are extremely high by comparison to the other POA rates and this is because of the use of direct age standardisation. This effect is created by the calculation of age specific rates for age groups that have very low populations, and therefore may fluctuate wildly based on very few deaths. When these are applied to the large populations in the standard, there may be many more deaths calculated than would really be the case.

The spatial distribution of the age standardised CVD mortality rates are shown in figures 17 through 20 on the following pages. These are split into winter and summer (six-month seasons) and each area shows the average of the rates calculated at each of the two seasons in the study period. The class breaks and colour ranges are standardised across the different maps to aid comparison. The areas with the highest rates are hidden by aggregating all areas above 2.25 per 1000 into the same class.

There are high average mortality rates in most parts of the city during winter, with concentration in the west and southwest regions. There are some areas that show clustering of high rates.

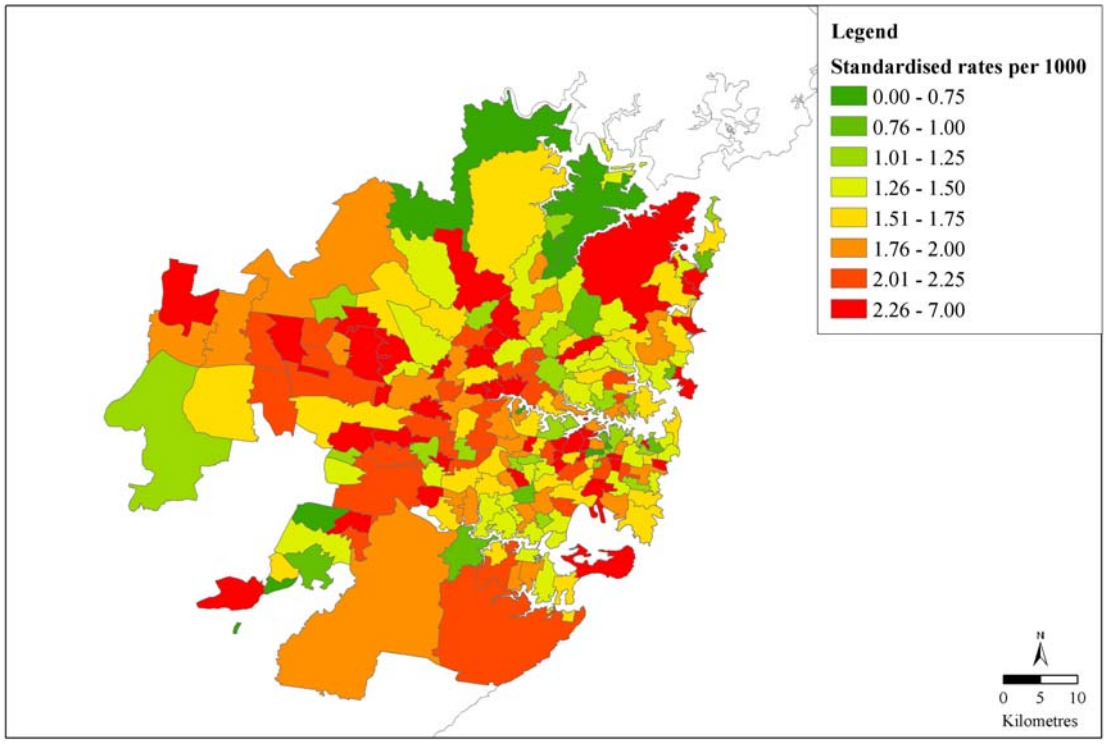


Figure 17: POA Standardised CVD Mortality Rates Averaged over the Winter Periods

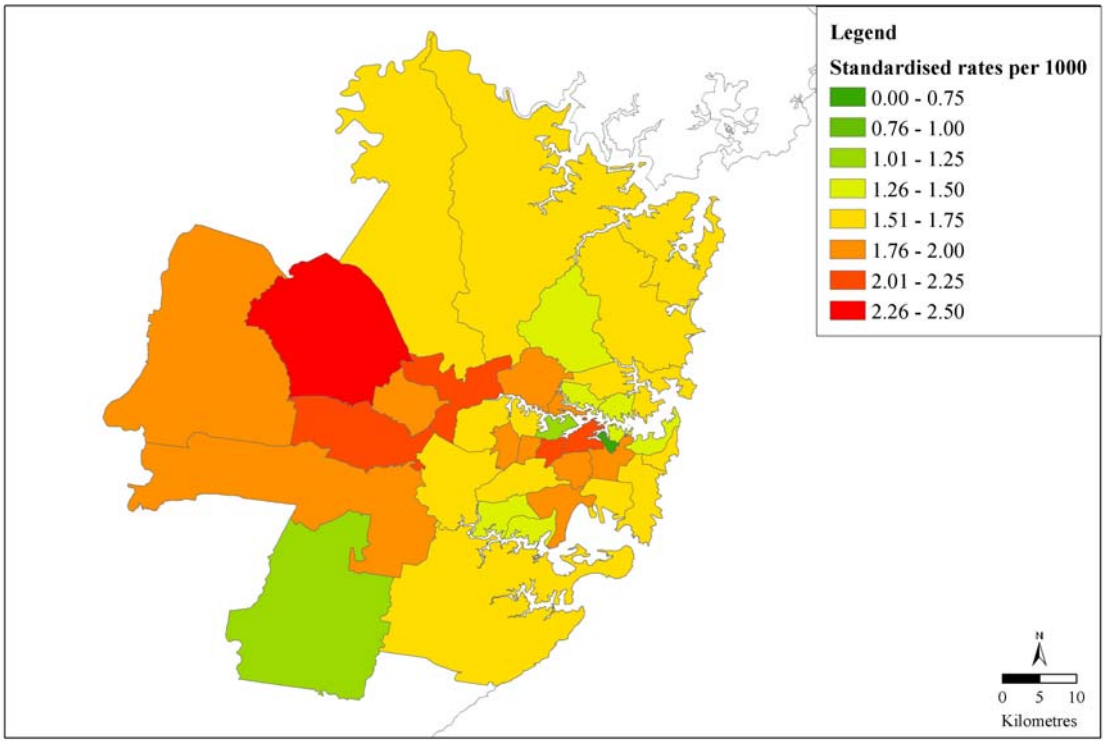


Figure 18: SLA Standardised CVD Mortality Rates Averaged over the Winter Periods

In summer the average rates are lower and there appears to be much more of a gradient between the eastern and western suburbs.

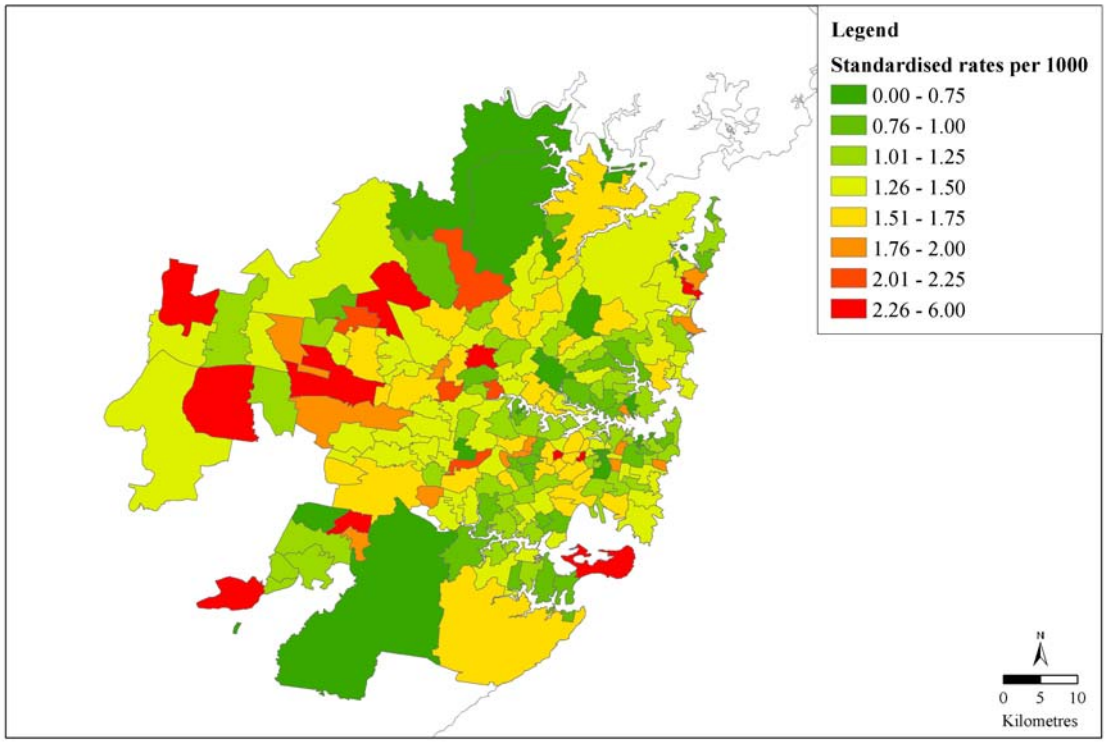


Figure 19: POA Standardised CVD Mortality Rates Averaged over Summer Periods

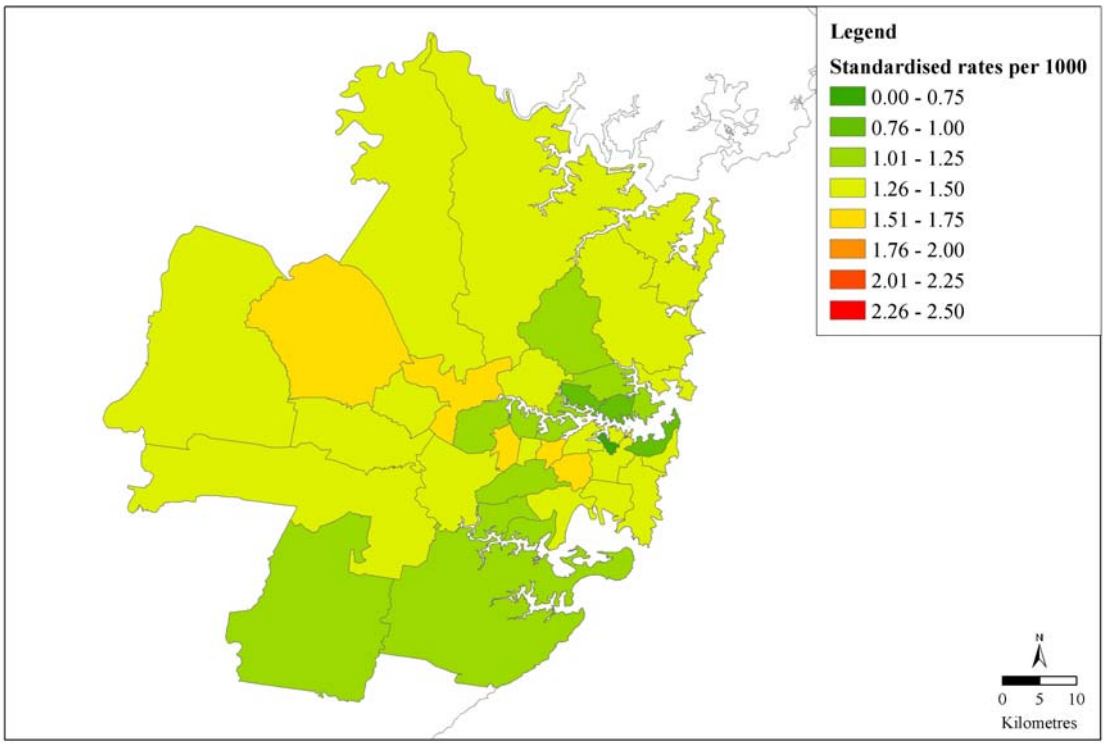


Figure 20: SLA Standardised CVD Mortality Rates Averaged over Summer Periods

Difference between Types of Spatial Units

The difference between the rates calculated at the two levels of aggregation is shown in figure 21. The rates for each area are used to make a gridded surface, and then each grid cell in the SLA rates surface is subtracted from the POA rates surface. This map shows that there are some SLAs that hide within their boundaries POAs with high and low rates that, when averaged across the SLA, become medium rates.

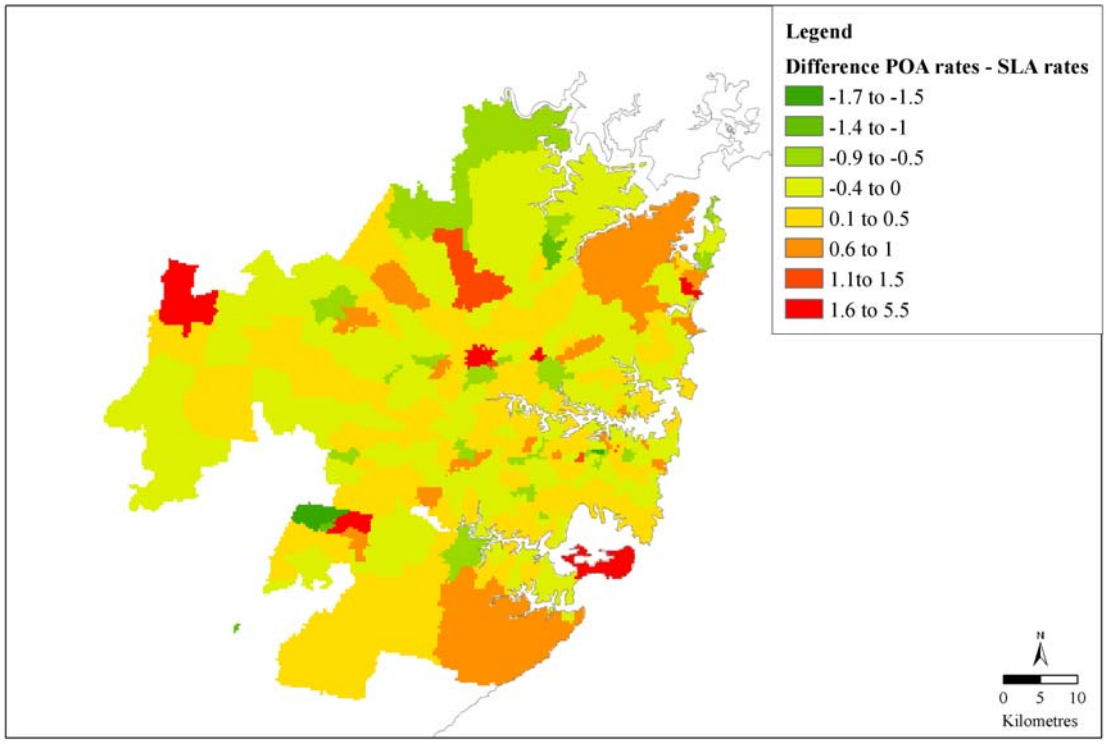


Figure 21: Difference between POA Average Rates and SLA Average Rates

Some of the extremely high rates may be due to the location of hospitals and nursing homes. This effect occurs regardless of the age standardisation because of the higher proportion of sick people who are likely to cluster together around these locations.

Areas where there are POA with higher rates than their SLA, next to POA with lower rates are an example of the zonation issue, where the SLA size and boundaries obscure fine spatial pattern by averaging out differences between areas.

Air Pollution

The time series in figure 11 showed that PM_{10} concentrations for the 1996 until 1998 period were higher in summer than winter. In figures 22 to 25 the spatial pattern of PM_{10} concentrations (averages of 24 hourly averages at each monitoring site) in each six-month season are displayed. These show a relatively consistent spatial distribution of this pollutant over time, with two distinct peaks of high concentrations over the inner west and inner south suburbs.

The exposure estimates for PM_{10} were derived from these spline interpolated surfaces and the process is depicted in figure 26. The SLA estimates are population weighted based on the CCD level areal averages whereas the POA estimates are not population weighted. This is because the POA are small enough that the areal averages will not cover the large range of concentration values that SLAs do.

The other pollutant's spatial and temporal patterns are displayed in appendix 3. The NO_2 spatial distribution is fairly consistent with PM_{10} however the peaks occur in winter rather than summer. The concentrations of O_3 are highest in summer and the spatial distribution is the opposite of the other pollutants. The peaks are found at the periphery of the city due to the photochemical processes that produce this chemical from the precursor compounds that are transported away from their source zones in the centre of the basin. The location of these peaks varies over time with the summer 96-97 peak in the southwest while the summer 97-98 peak is to the north.

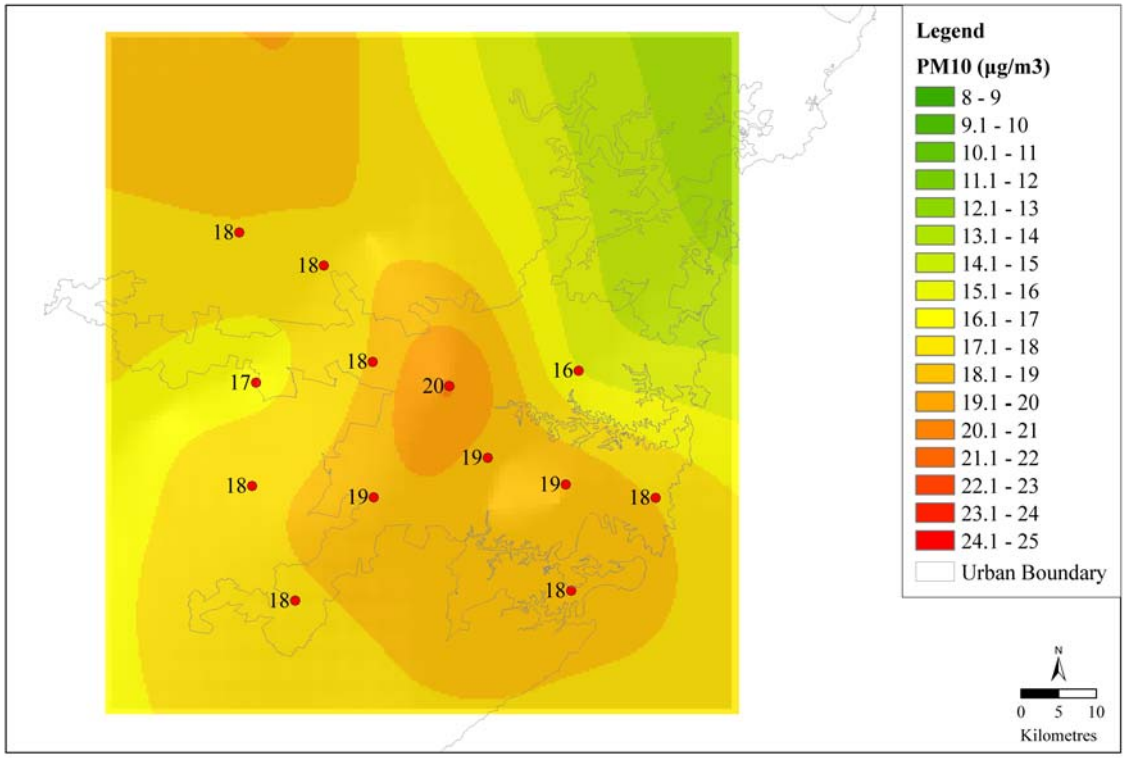


Figure 22: PM_{10} Concentrations in Summer 1996-97

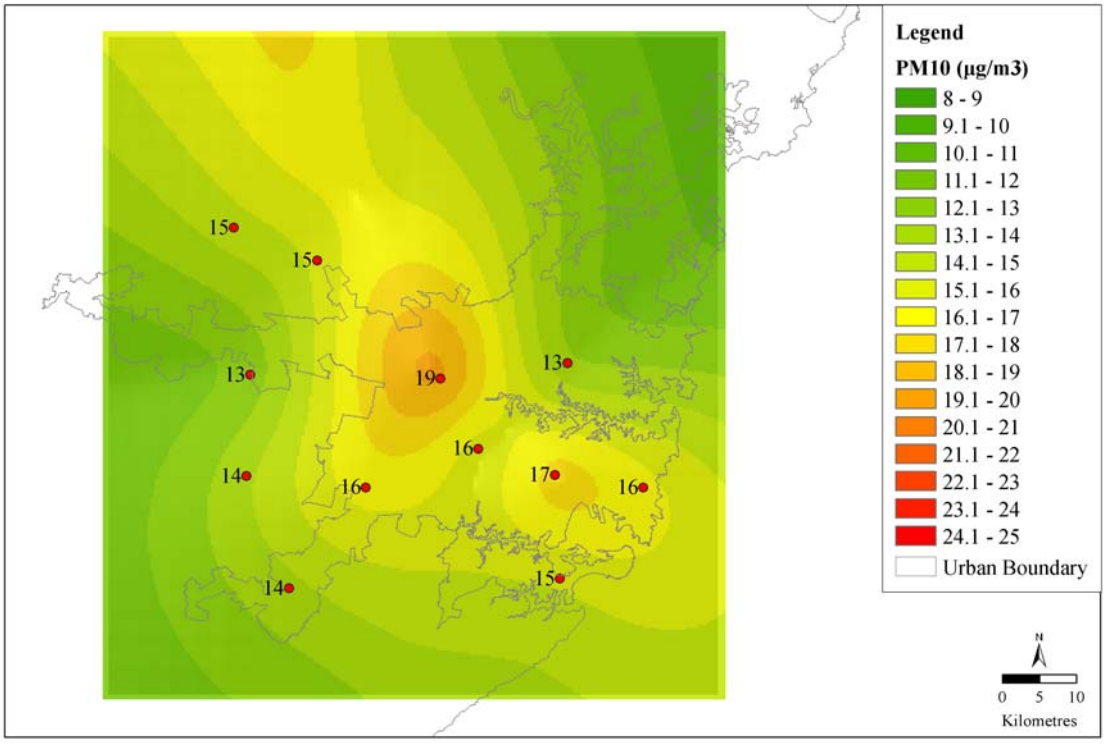


Figure 23: PM_{10} Concentrations in Winter 1997

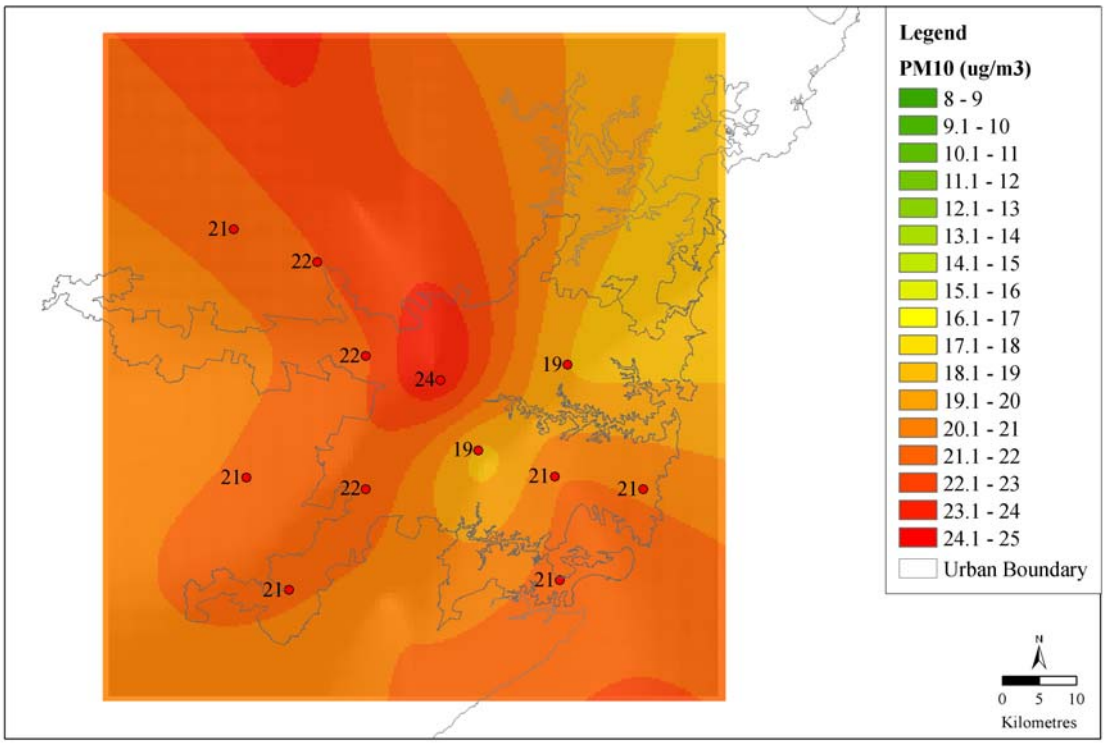


Figure 24: PM_{10} Concentrations in Summer 1997-98

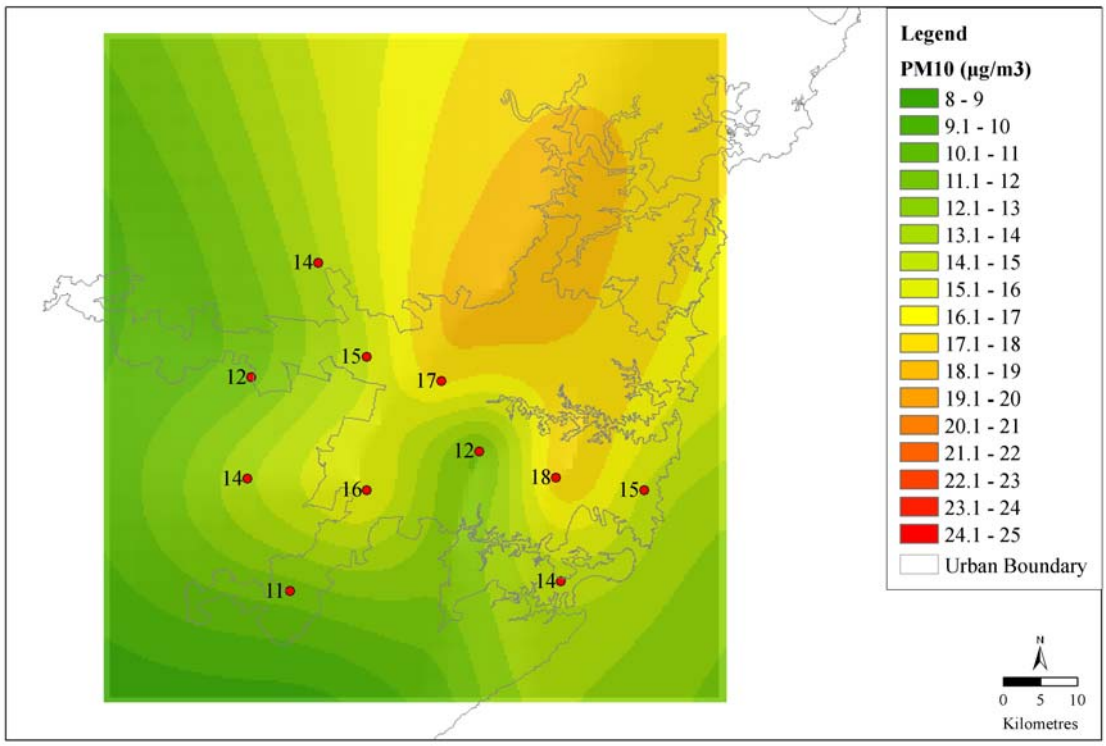


Figure 25: PM_{10} Concentrations in Winter 1998

The exposure estimates for PM_{10} were derived from the spline interpolated surfaces and the process is depicted in figure 26. The SLA estimates are population weighted based on the CCD level areal averages whereas the POA estimates are not population weighted. This is because the POA are small enough that the areal averages will not cover such a range of concentration values as SLA.

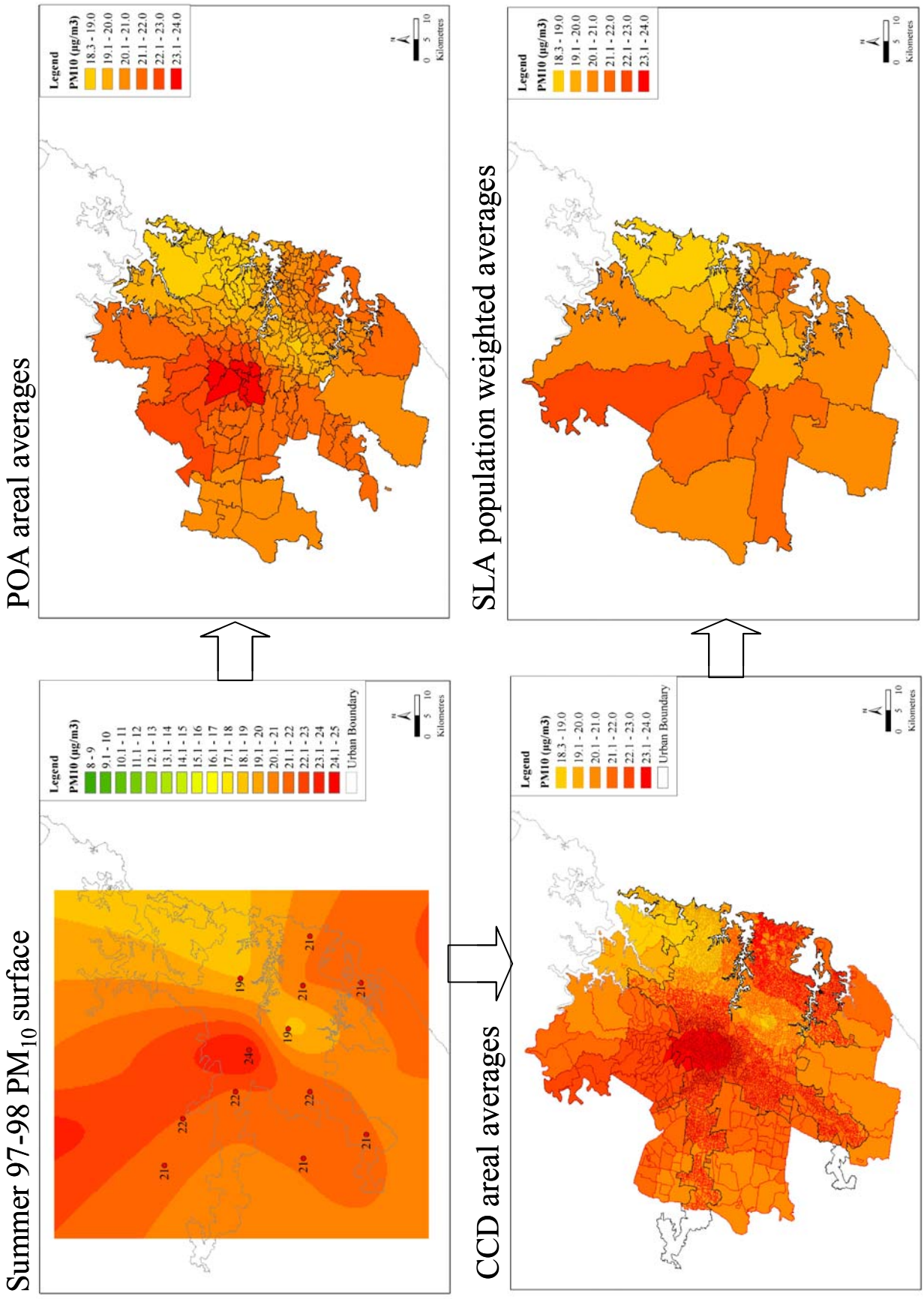


Figure 26: Exposure Estimates at CCD, SLA and POA Levels

Regression Tree Analysis

The tree models were used to explore the data and all the variables are included in this data mining stage. The POA tree model results will be presented first and then the SLA results.

Postal Areas

The results of the spatiotemporal modelling for POA rates in all four seasons together using regression tree analysis are presented in figure 27.

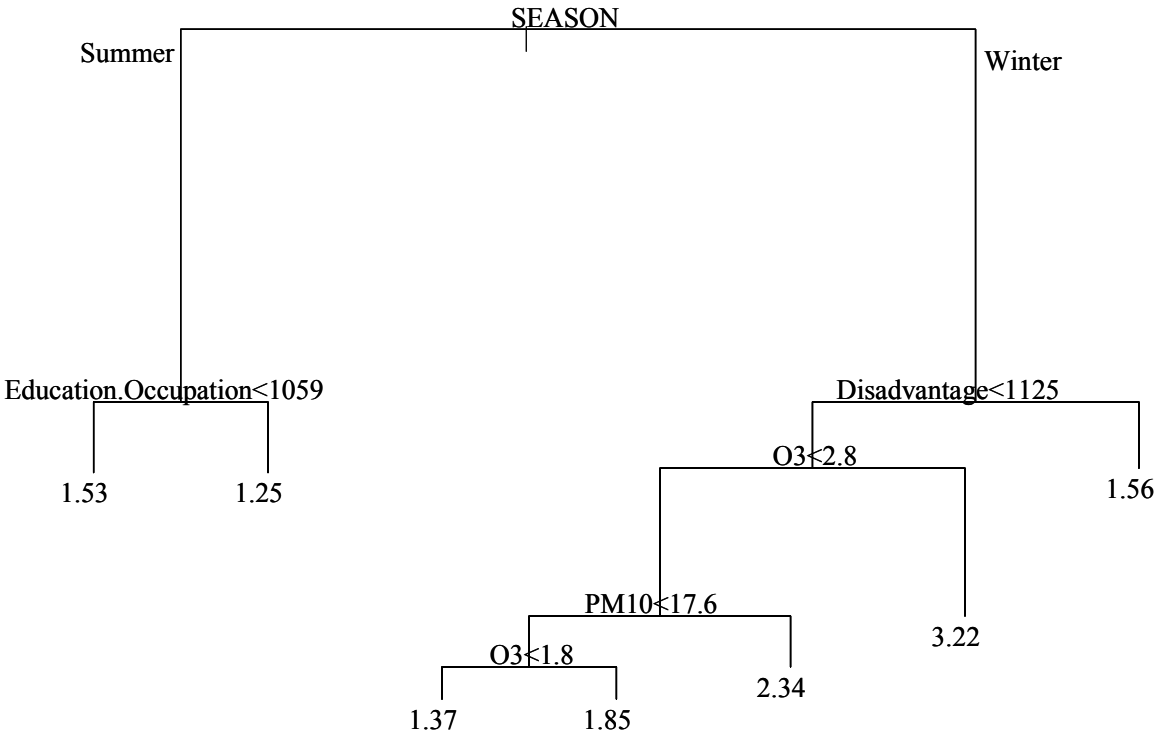


Figure 27: Tree Model of POA Standardised Rate (per 1000)

The tree model divides the POAs into seven non-overlapping endgroups at the bottom of the tree. Each of these groups has a similar standardised rate, is labelled by the mean of these and is defined by the rules given in the preceding levels of the tree.

The first level of the tree shows that season has the strongest effect. Note the longer lines between this level and the next; these are proportional to the reduction in deviance between the original group and the split groups (therefore this split is very effective at reducing the deviance of the group at the previous node).

SES dominates the next level of the tree for both summer and winter groups. In winter, the disadvantaged index was used to split with those below 1125 on the disadvantage index split into the higher group - therefore highly disadvantaged areas have higher rates. The highest endgroup mean standardised rate is found on this branch with low disadvantage scores. The POA with O_3 levels (above 2.8 pphm) group had a mean of 3.22 per 1000. Then the next highest group mean at 2.34 per 1000, had lower O_3 levels but high PM_{10} concentrations (above 17.6 $\mu g/m^3$).

The group with the lowest average rate (1.25 per 1000) were found in summer, in areas with high scores on the Education Occupation index, an indicator of high SES.

The relationship with O_3 in winter is counterintuitive as the summer concentrations of this pollutant are much higher than in winter. There may be differences in exposure (concentration pattern or pollution mix) and susceptibility that explain this result, or it may be a spurious association. This could occur due to the effect of the extremely high rates that may result from direct age standardisation in small populations. It has been noted that tree models are "adversely affected by outliers, which can cause very different tree results when they are included" (Miller and Franklin 2002: 244), just as other forms of regression analysis are also badly affected by these.

As discussed above in relation to figures 15 and 16, the rates above 6 per 1000 are extremely high. In fact the group defined by winter, low disadvantage scores and high O_3 includes two of these three extremely high rates and only contains nine POAs. Therefore the influence of this pollutant is unlikely and the three upper outliers (rates greater than 6 per 1000) were ignored. The resultant tree is shown in figure 28 which shows that O_3 became less important than PM_{10}. O_3 was also excluded from the model altogether however the other variable's order of entry and split level did not change so the tree model result shown here includes O_3 and excludes the three upper outliers.

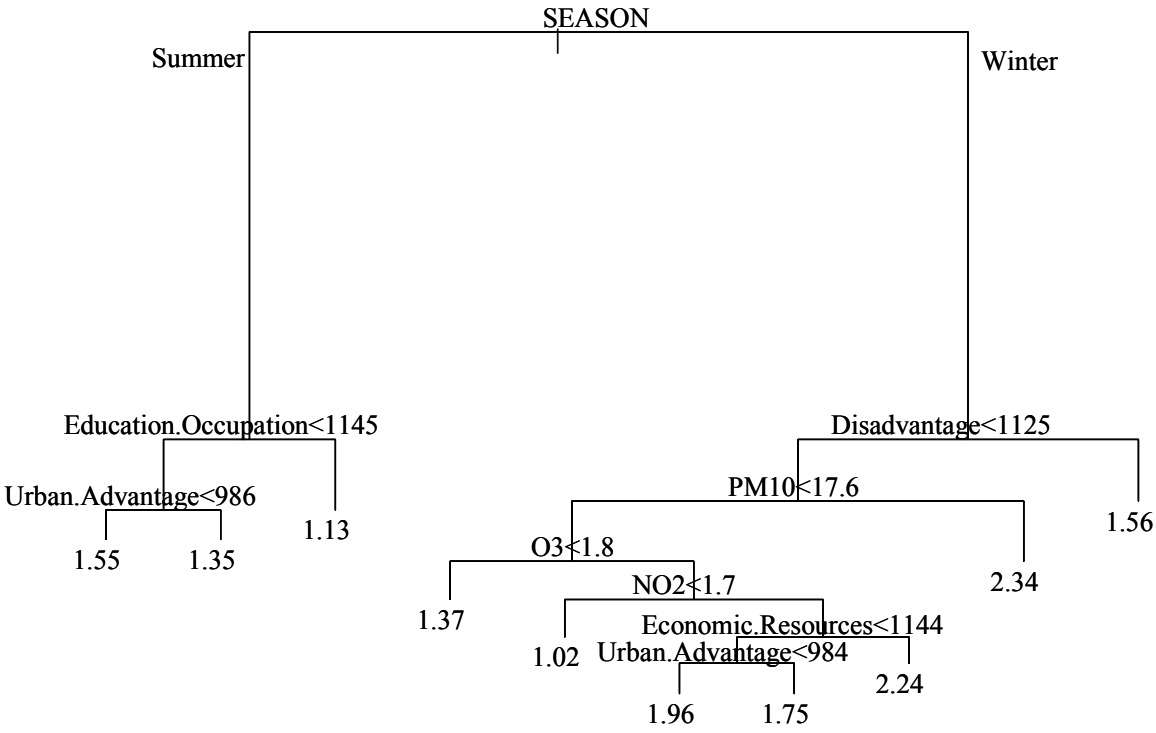


Figure 28: Tree Model of POA Standardised Rate (per 1000): Under 6

The POA tree model reduced the root deviance in the dataset from 355.0 to 283.9 (calculated by summing the deviance of each of the endgroups identified). This can be interpreted as the tree model reducing the deviance by 20%.

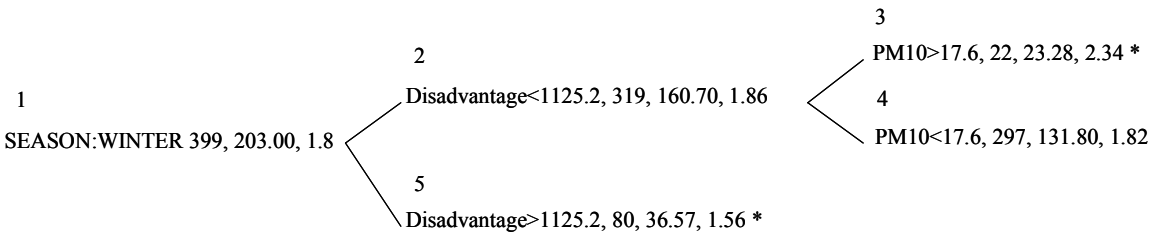


Figure 29: POA rates without upper outliers in winter tree diagram

In figure 29 the tree model results for the winter branch are shown horizontally for discussion. The subgroups are labelled by the variable that defined them, the number of data points, the sum of squares and the mean of the response variable. Thus the 399 POAs in winter (group 1) have a deviance of 203.00 and a mean standardised CVD rate of 1.8. These POA are then split by the disadvantage index at the level of 1125.2. There were 319 most disadvantaged areas (with low disadvantage scores) in group 2 and these were split by PM_{10} at 17.6 $\mu g/m^3$. The highest group of winter POA rates was the 22 areas that had low disadvantage scores and high PM_{10} levels with mean standardised rate of 2.34 per 1000 (group 3).

There were 297 most disadvantaged POA with low PM_{10} (group 4) which are split further by O_3, then NO_2 and then the Economic Resources and Urban Advantage indexes. These are not shown in figure 29 to avoid excessive detail. However it is noted that in figure 28 the NO_2 and O_3 splits both suggest that the relationship is positive for these pollutants also.

The group of areas with the lowest rates in the winter data were the least disadvantaged 80 POAs with high disadvantage scores (group 5) and an average rate of 1.56 per 1000.

An alternative to ignoring the upper outliers is to weight the regression tree model using the inverse of the variances for each rate estimate. The results of this method are shown in appendix 4. This found that the SES indexes were far more important than the AP variables, which either dropped down in the order of importance or where removed from the tree altogether. As the focus of this study is to explore understandings of the variations in the data, the results obtained by the unweighted regression trees are given attention here.

Statistical Local Areas

In the SLA regression tree model the influence of season was found to have the most effect on the differentiation of the data. This variable is followed by SES in both seasons and then PM_{10} enters the model in the third level, splitting the disadvantaged areas.

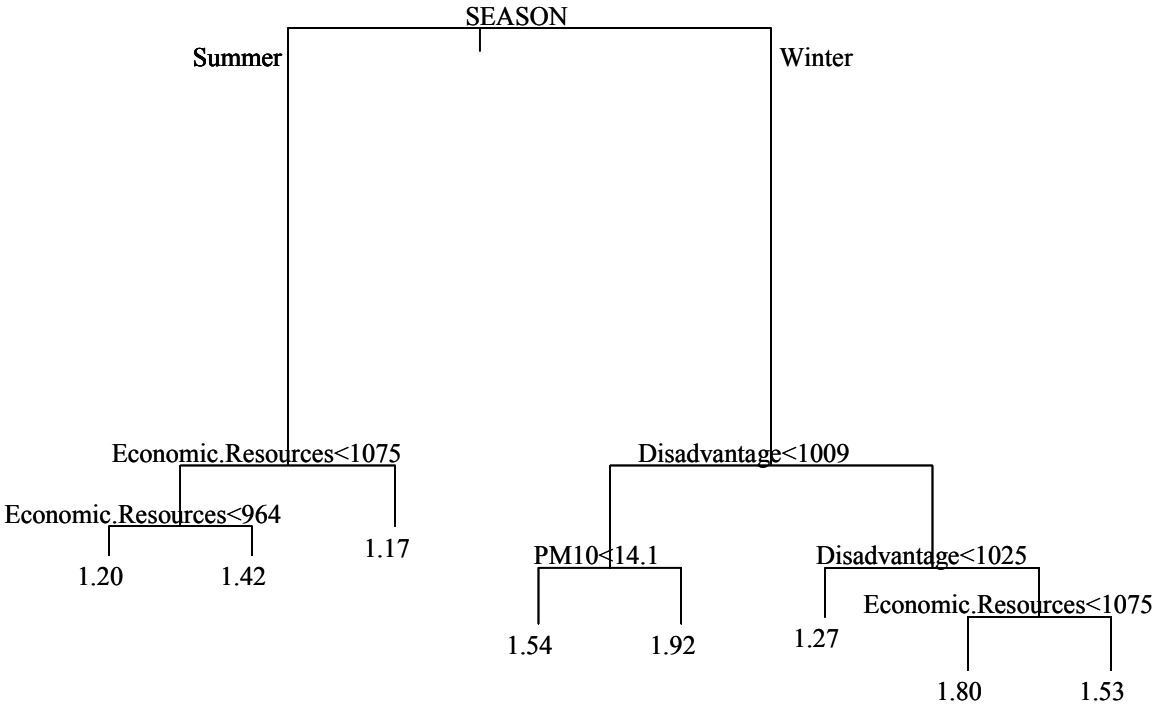


Figure 30: Tree model of SLA standardised rate (per 1000) for all variables at all times

The SLA tree model reduced the root deviance from 19.4 to a total end group deviance of 9.3. This means that the SLA tree model reduced the deviance by 52%.

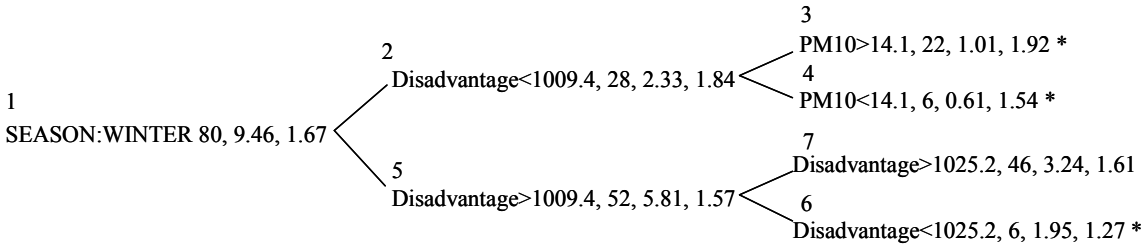


Figure 31: SLA Rates in Winter Tree Diagram

There are 80 SLAs in winter (group1) with a mean standardised CVD rate of 1.67 per 1000. The disadvantage index split these (as it did in the POA tree however the split level is lower at 1009). The most disadvantaged group with scores below this split level (group 2) had 28 areas and was subsequently split by PM_{10} at 14.1 $\mu g/m^3$. There are 22 areas with low disadvantage scores and high PM_{10} (group 3) that have the highest average rates of all winter SLA groups identified by the tree at 1.92 deaths per 1000. The low disadvantage and low PM_{10} areas (group 4) had 6 areas and average of 1.54 per 1000.

The 52 least disadvantaged areas (group 5) were split by disadvantage again (groups 6 and 7). This differentiated 6 middle class areas with medium disadvantage scores (>1009 and <1025) and the *lowest* of the winter group's rates at 1.27 per 1000.

The least disadvantaged areas (>1025 - group 7) are therefore interesting because they have relatively higher death rates than these middle class areas. These 46 SLAs are split by the Economic Resources variable to identify a group of 14 SLAs with low scores on this index (<1075) and a high average death rate at 1.80 per 1000.

The two regression tree results can be juxtaposed against each other and ranked in terms of the outcome variable as shown in figure 32. The tree on the left is the winter POA branch and the tree on the right is the winter SLA branch. These extend towards each other in the middle of the graph, and the levels of the tree partitioning process are labelled at the top. The standardised CVD mortality rate is shown by the horizontal lines. This graphic display of the tree models shows that the range of POA rates is greater than that in the SLAs.

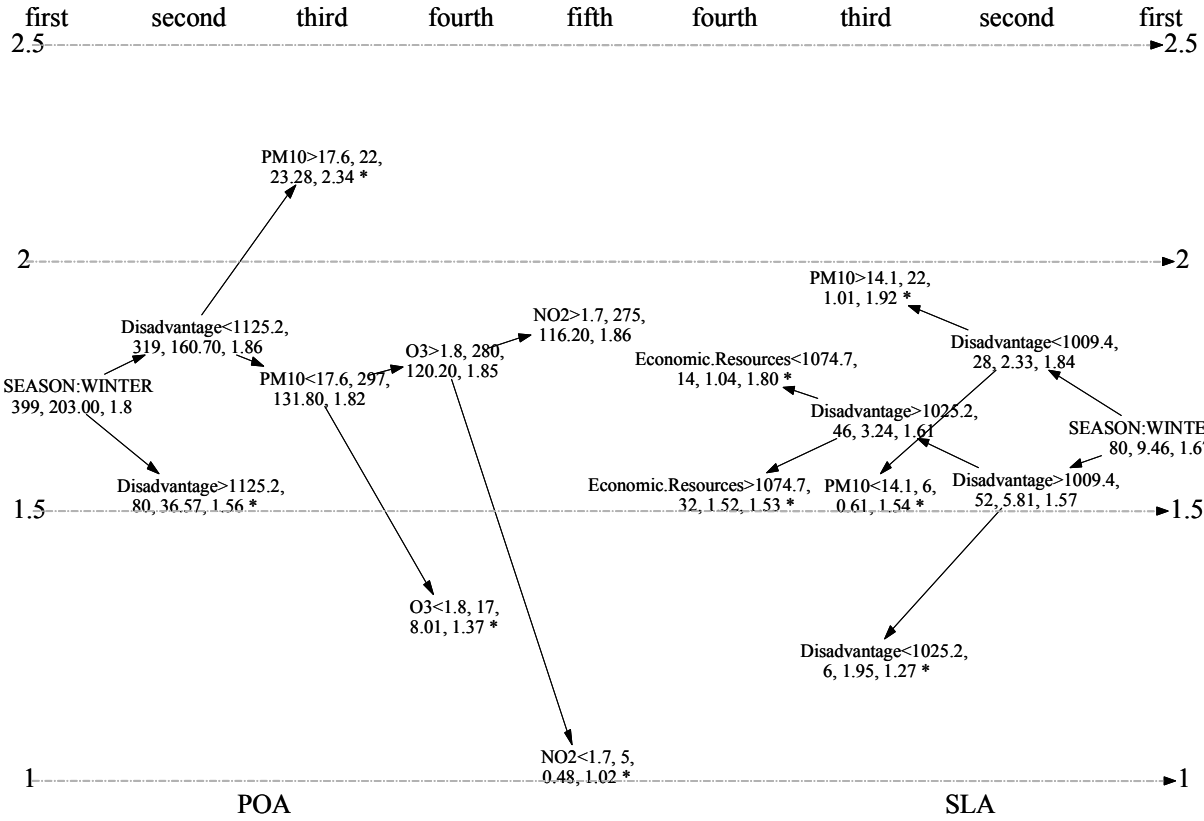


Figure 32: Graph of Winter Rules Juxtaposed Against the CVD Standardised Rate

The rules that define the groups with similar rates can be followed from the root to the branches. For instance the low disadvantage scores and high PM_{10} concentrations rule defines the highest rates in both types of spatial units. For the medium and low rate subgroups the POAs seemed to find more explanation with the O_3 and NO_2 variables than SLAs which tended to split on SES variables.

Visualisation of Relationships

The key relationships identified by the tree models were with the disadvantage index and PM_{10} concentrations in winter. These were explored first by mapping the regions defined by these rules, then using box plots and conditional scatter plots with locally weighted smoothing (loess) lines to understand how these variables influence one another.

The maps in figures 32, 33 and 34 show the areas that were most disadvantaged and had high PM_{10} concentrations in winter in 1997 and 1998. These were the data points in the endgroups that had the highest mean standardised rates and are shown overlaying their average winter standardised CVD rates.

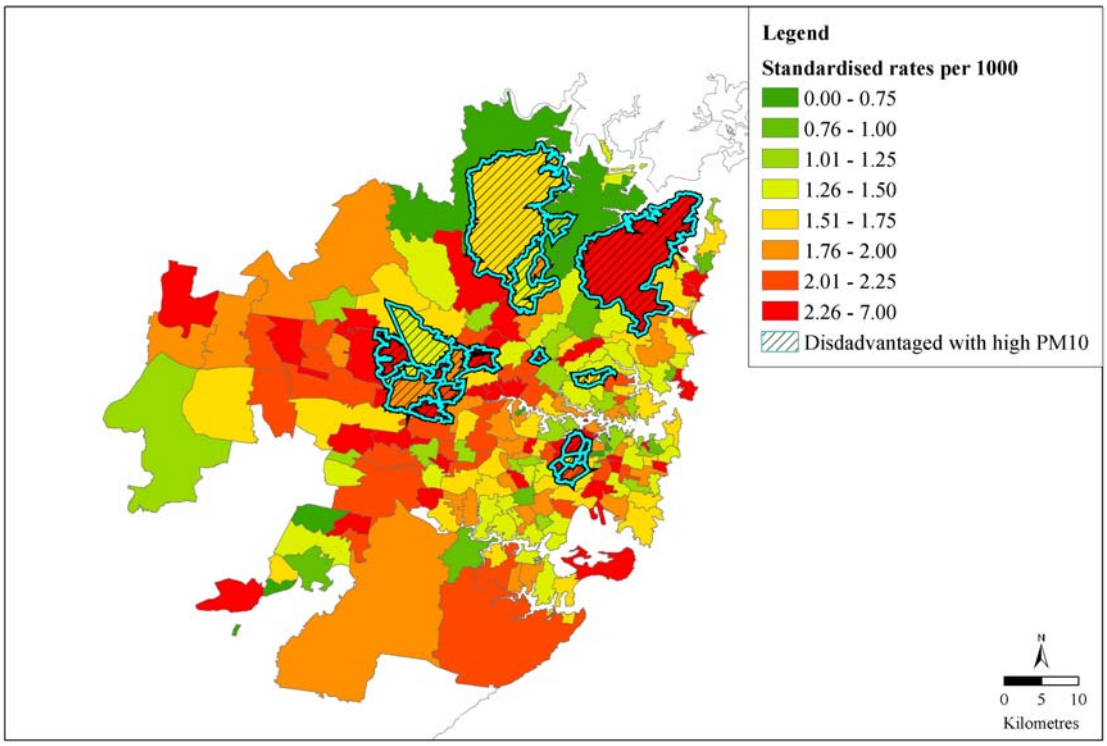


Figure 33: POA High Endgroup Areas

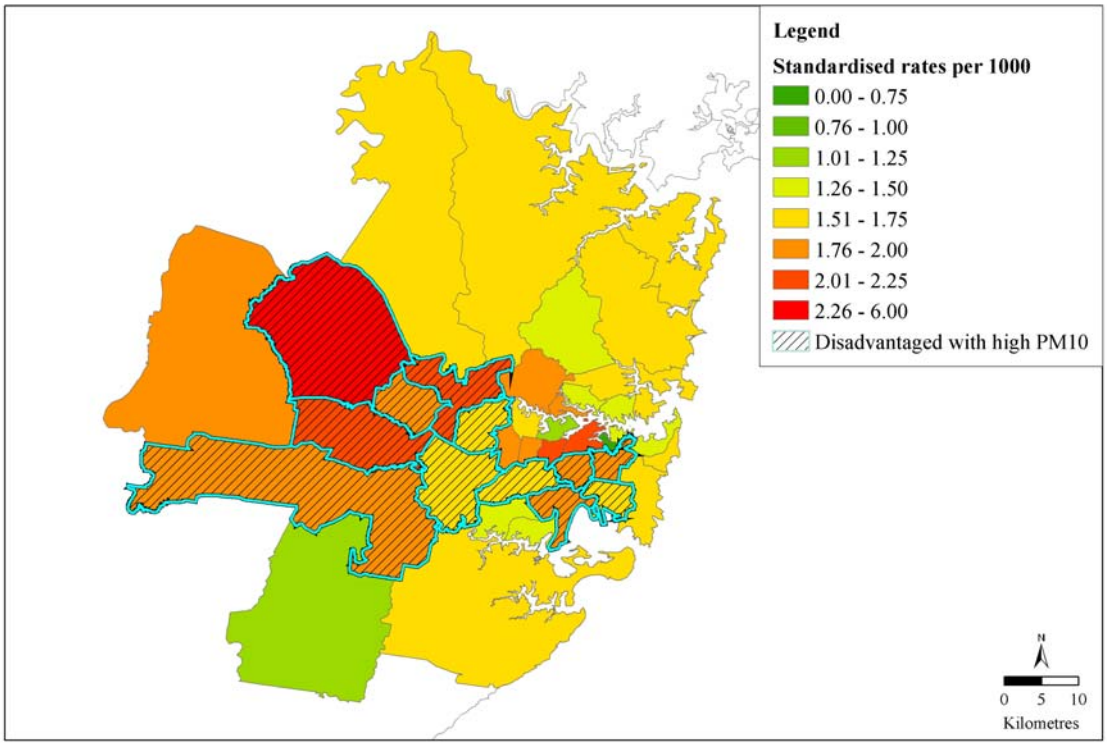


Figure 34: SLA High Endgroup Areas

This shows some agreement in the tree results with both spatial units highlighting the region around Marrickville and between Parramatta and Blacktown (see appendix 1 for map of suburb names). The areas that do not agree between the two types of spatial units indicate the uncertainty in the geographical patterns of the relationships identified. The SLA boundaries miss some small areas of high risk while showing the broad regional trend.

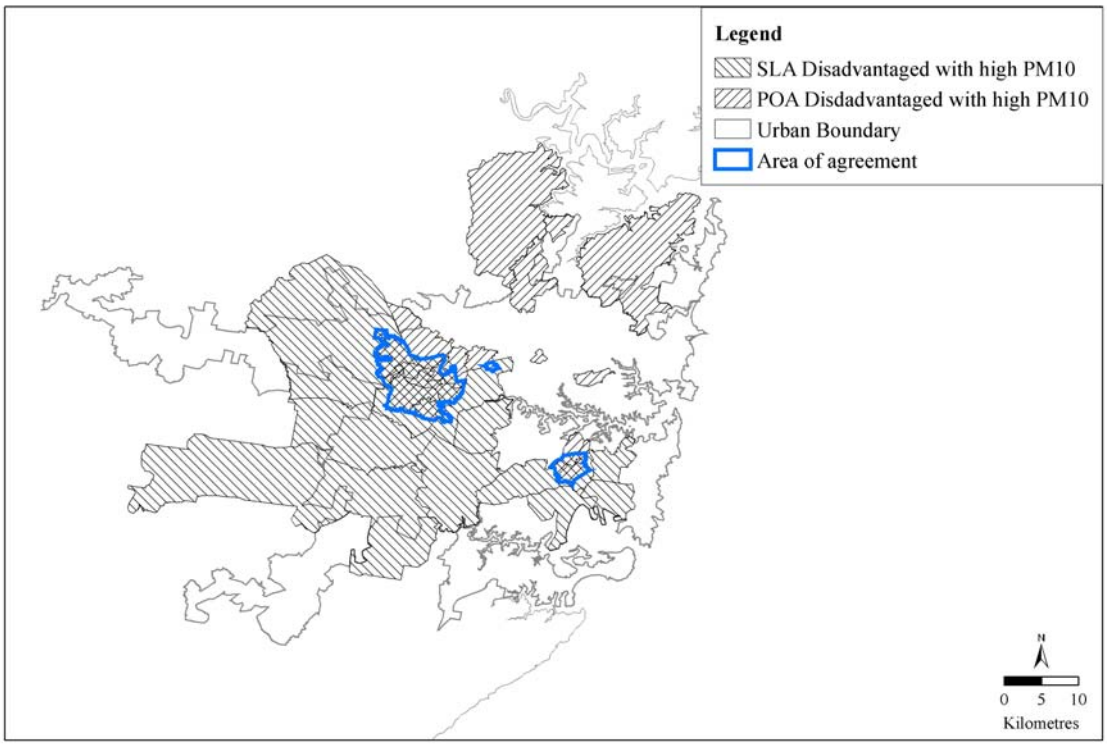


Figure 35: Area of Agreement between Tree Models

Season

The difference in standardised CVD mortality rates between the seasons is displayed by box-plots

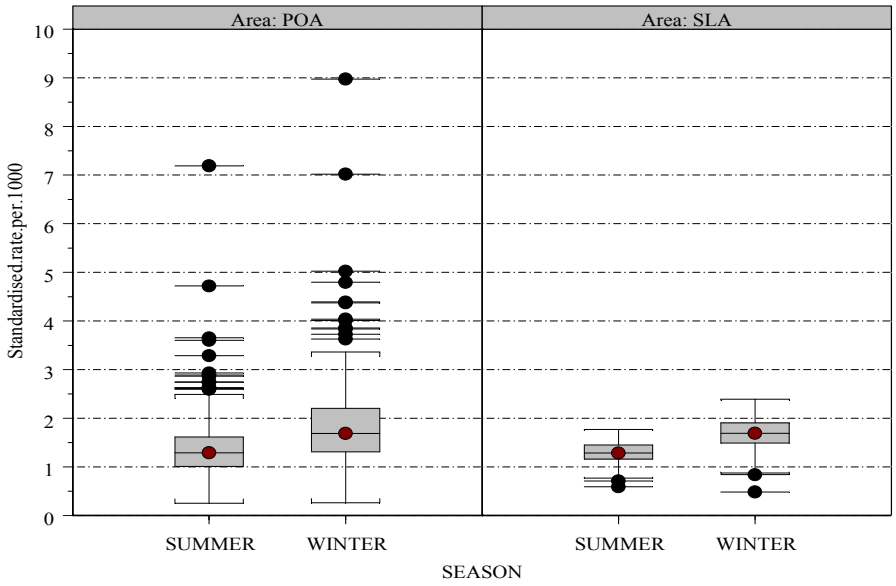


Figure 36: Standardised Rates by POA and SLA for the Two Types of Season

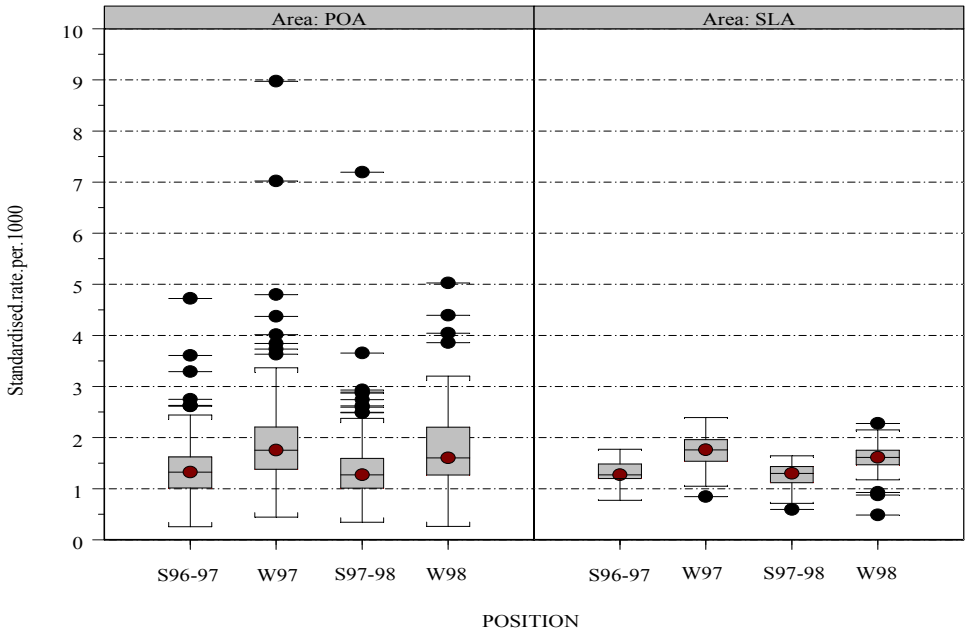


Figure 37: Standardised Rates by POA and SLA for the Four Seasons

These box plots show that the rates are higher in the winter seasons and that winter CVD rates were worse in 1997 (W97) than in 1998 (W98).

Disadvantage

The relationship between the disadvantage index and the standardised rates were visualised using scatter graphs with loess lines. These show that there is a negative relationship (as the disadvantage index scores increase SES decreases). This indicates that the less disadvantaged areas (with high scores) have lower CVD rates than the more disadvantaged areas. The relationship is essentially linear in the scatter and loess lines suggesting that linear regression modelling will be suitable.

In the figures 38 and 40 the relationship between CVD rates and disadvantage index are conditioned by season. In the left hand conditional plot the data are broken into winter and summer subsets while the right hand plot is broken into the four separate seasons. Figures 39 and 41 focus on the relationship between the disadvantage index and CVD rates in winter only. The ranges of the axes are the same for both graphs to enable comparison of the relationship despite the different range in this variable across the spatial units.

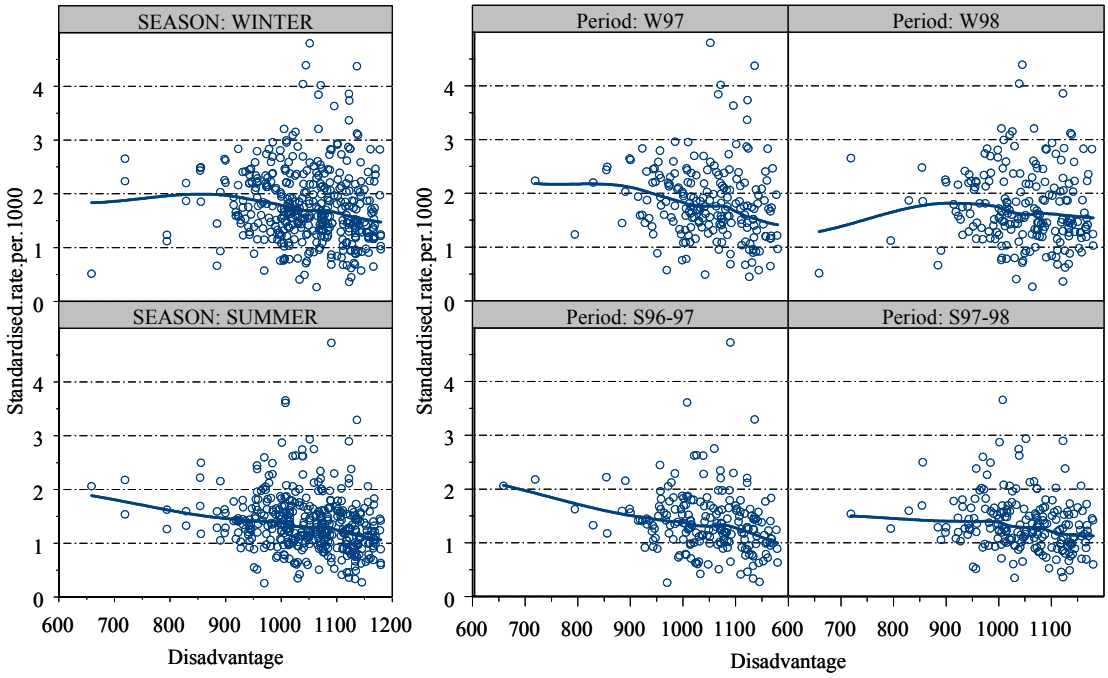


Figure 38: POA Rates against Disadvantage by Season

The relationship between CVD and the disadvantage index shows a negative trend in each of the separate seasons perhaps with the exception of Winter 1998 (W98).

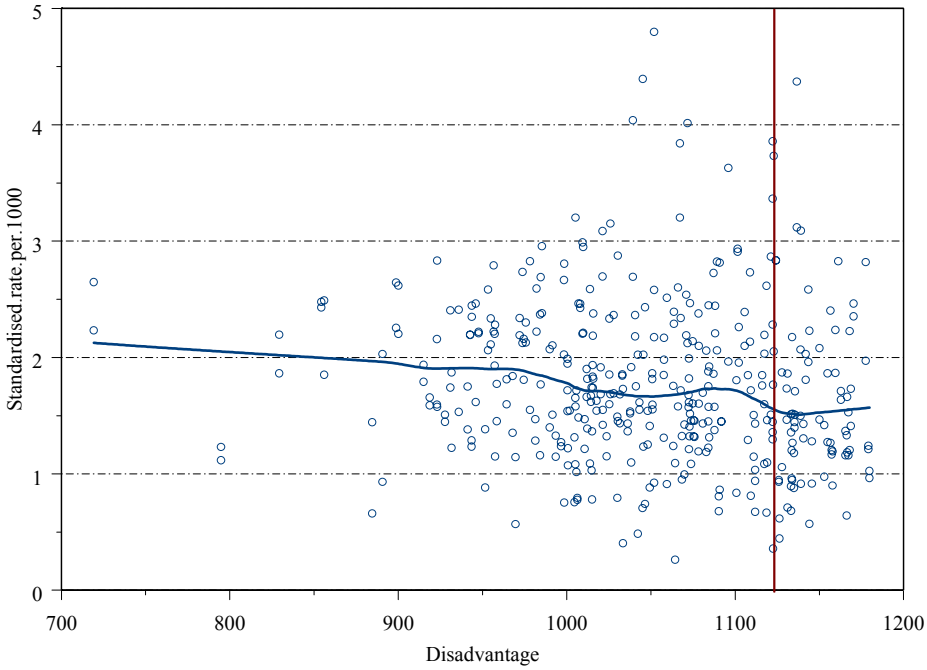


Figure 39: Winter POA Rates against the Disadvantage Index

The loess line fitted to the winter POA rates scatter graph against disadvantage is shown in figure 39. The disadvantage index threshold defining the POAs in the tree model was 1125 and is marked by the vertical bar.

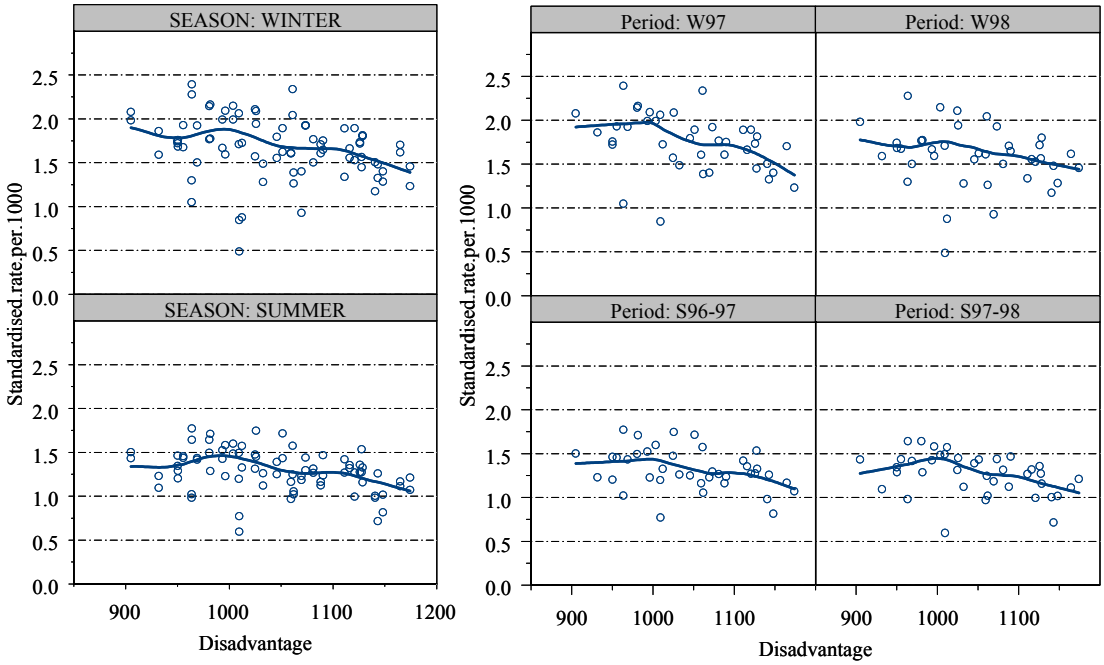


Figure 40: SLA Rates against Disadvantage by Season

The same as negative relationship is evident for SLA rates against disadvantage shown in figure 40. In figure 41 the winter relationship is shown with the split level 1009 identified by the tree.

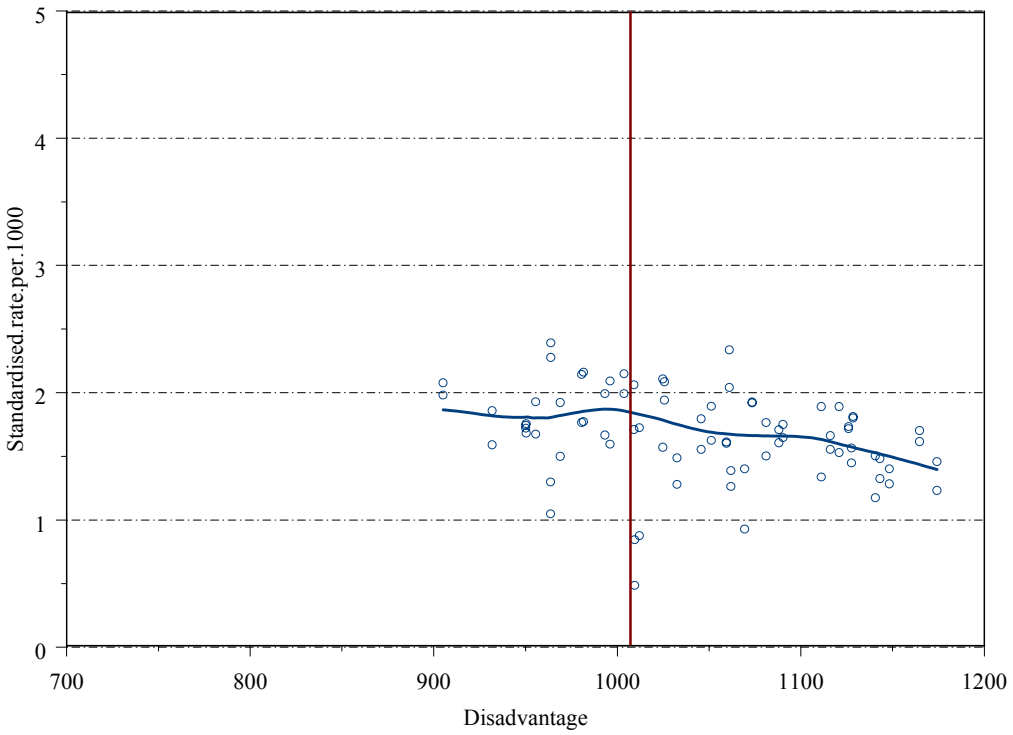


Figure 41: Winter SLA Rates against the Disadvantage Index

In the tree model different SEIFA indexes entered the model at similar positions and with similar meanings. This is because they are highly correlated. The close similarity is visualised in the scatter matrix shown in figures 42 and 43. The loess lines indicate that despite some potential to differ, on the whole the indexes have very similar values.

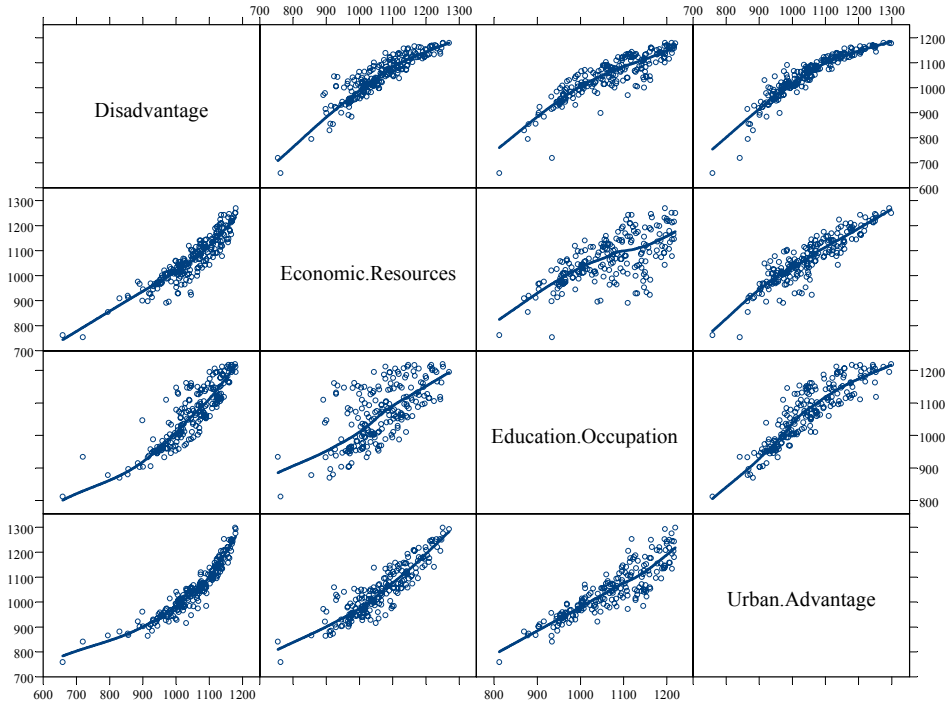


Figure 42: POA SEIFA Scores

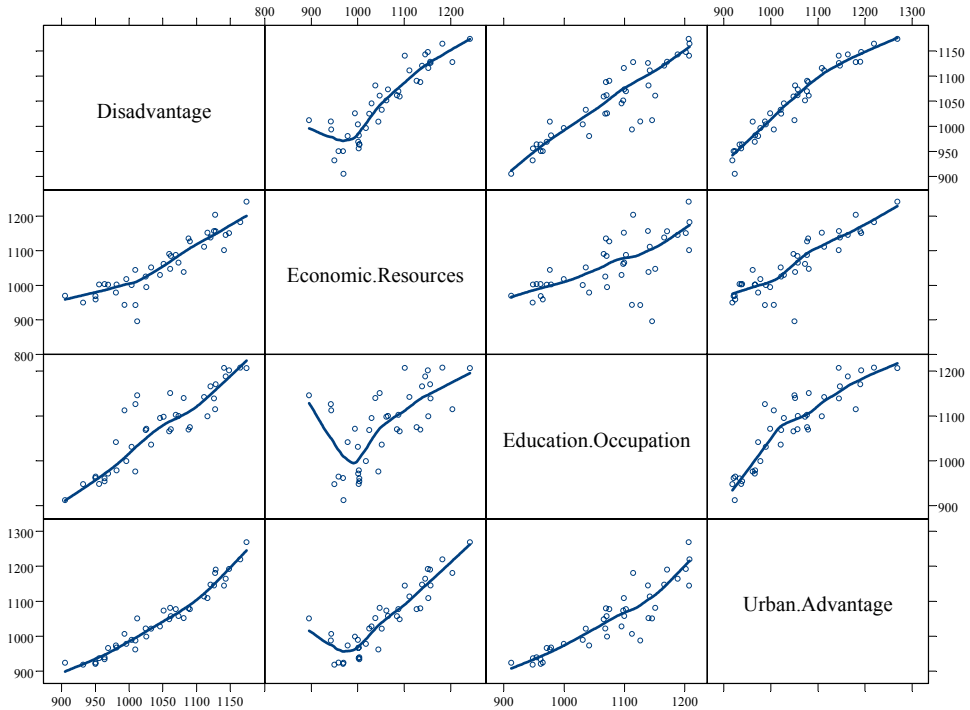


Figure 43: SLA SEIFA Scores

PM_{10}

The PM_{10} relationship with the winter CVD rates were then conditioned by the disadvantage index. In the left hand conditional plot in figure 44 the two subsets are defined by the split level identified in the tree model whereas in the right hand plot the data are broken into quartiles, showing a decreasing influence as disadvantage scores increase, suggesting an interaction such that the most disadvantaged areas are affected by PM_{10}.concentrations more than the least disadvantaged areas.

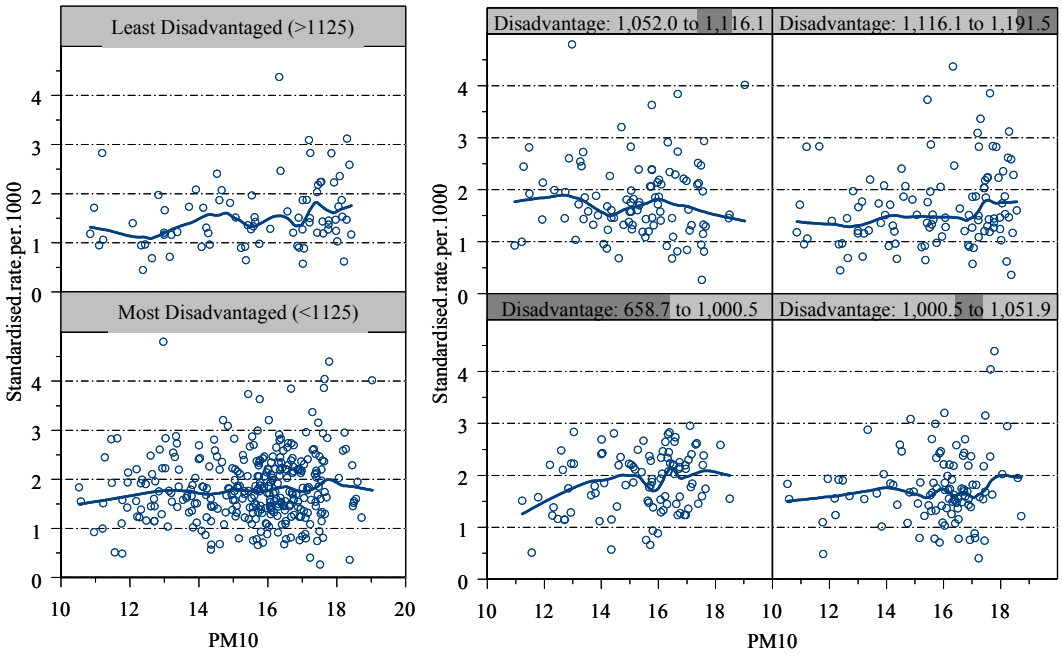


Figure 44: Winter POA Rates against PM_{10} by Disadvantage Groups

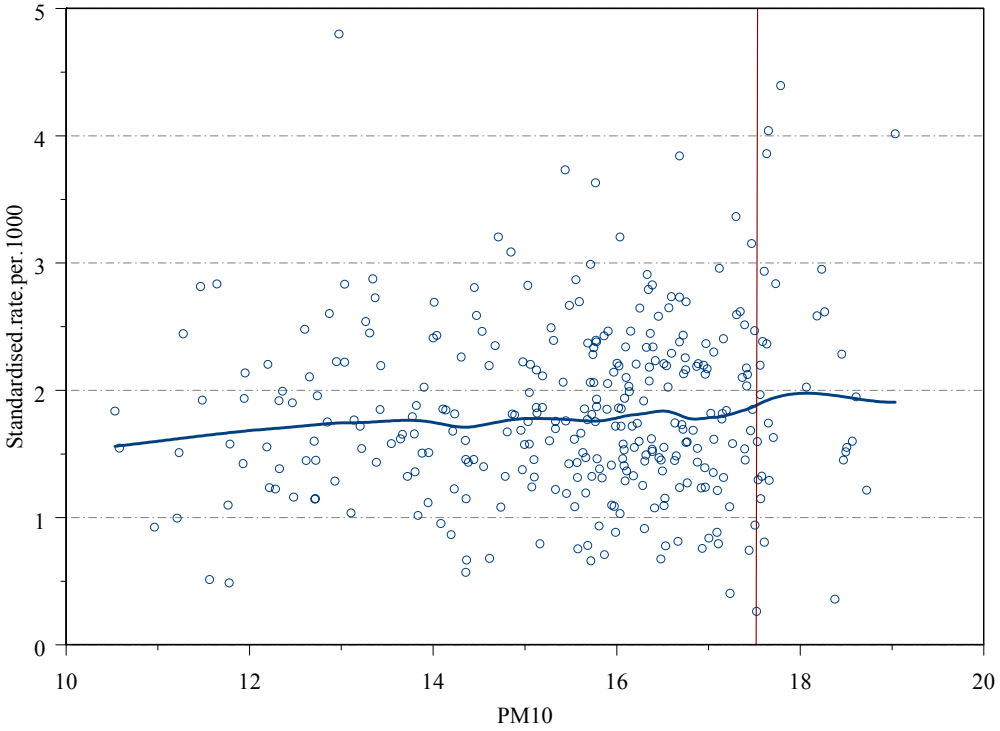


Figure 45: Winter Disadvantaged POA Rates against PM_{10}

The greater influence of PM_{10} on disadvantaged areas is more obvious in the SLA conditional plot in figure 46, possibly accentuated by the smaller range on the y axis.

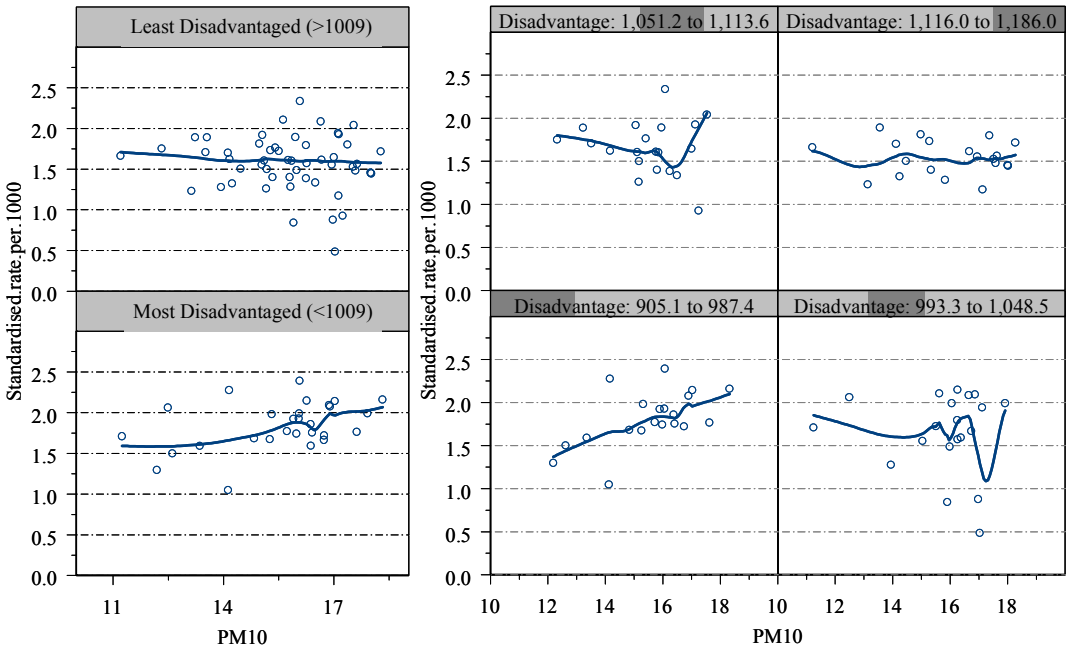


Figure 46: winter SLA rates against PM_{10} by disadvantage groups

The disadvantaged SLAs in winter with PM_{10} greater than 14.1 $\mu g/m^3$ were identified by the tree model. The relationship in figure 47 shows a positive gradient that is essentially linear.

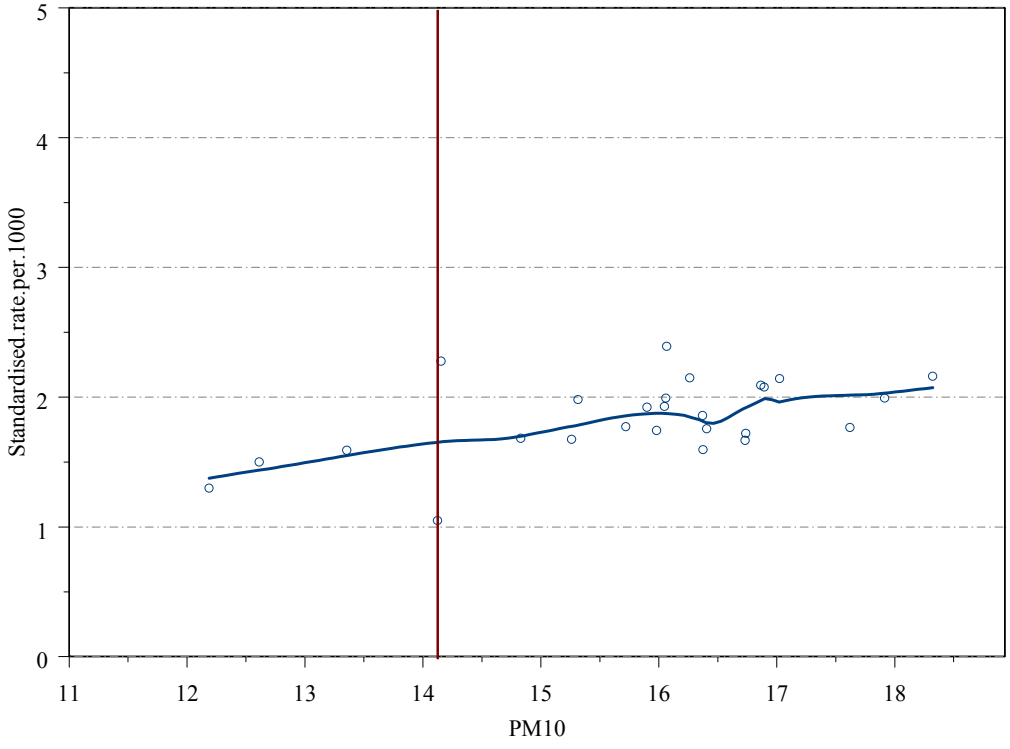


Figure 47: winter disadvantaged SLA rates against PM_{10}

Linear Regression

Socio-Economic Disadvantage in Winter

In table 5 the results of regression (without POA upper outliers) are displayed. These show that both spatial units found similar regression coefficients but different p-values. The SLA p-value would satisfy the 95% confidence level whereas the POA result would not. The SLA R-squared is a factor of ten higher than that for the POAs. But at 0.079 this result is not a high level of explanation from this variable.

Table 3: Regression Results of Standardised Rates on Disadvantage

	POA				SLA			
	Value	Std. Error	t value	Pr(>|t|)	Value	Std. Error	t value	Pr(>|t|)
Intercept	2.574	0.451	5.712	0.000	3.095	0.554	5.591	0.000
Disadvantage	-0.001	0.0004	-1.755	0.080	-0.001	0.0005	-2.587	0.012
R-Squared	0.008				0.079			

The regression coefficients are shown in figure 48 with 95% confidence intervals (calculated by adding and subtracting 1.96 times the standard error). These overlap and the two results are therefore not statistically different. As the p-value shows the SLA result is significantly different from zero, whereas the POA result is not, indicated by the 95% confidence interval being so close to zero.

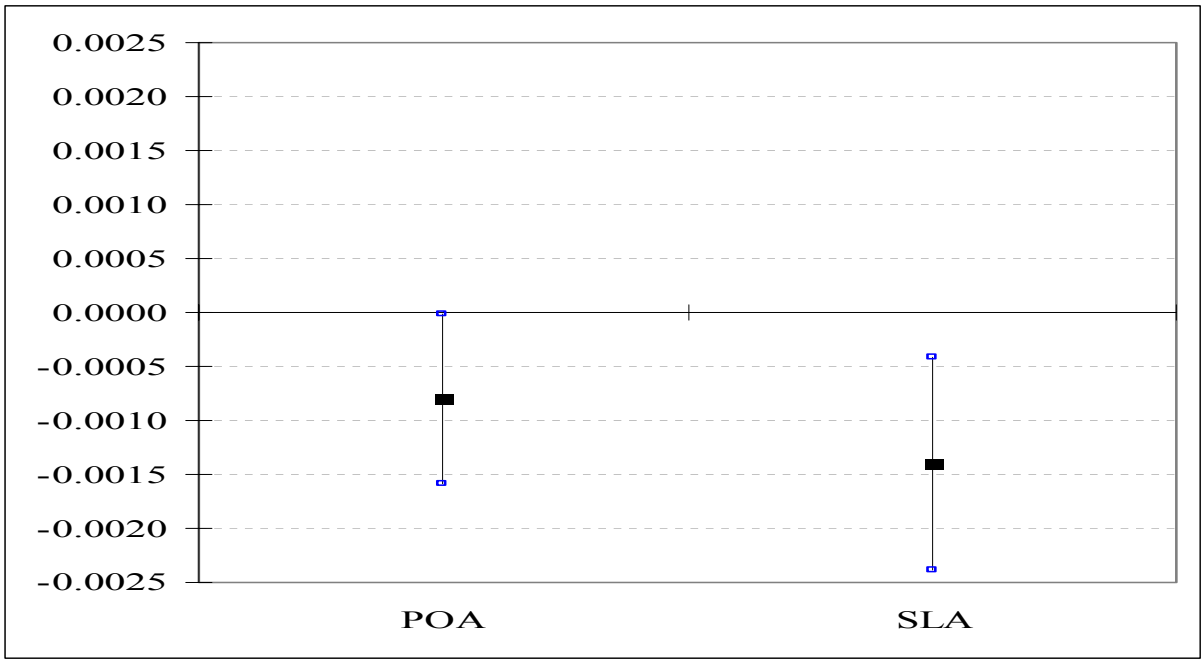


Figure 48: Comparison of Disadvantage Regression Coefficients and 95% Confidence Intervals

PM_{10} in Most Disadvantaged Areas in Winter

Table 4 presents the results of the linear regression of PM_{10} in the most disadvantaged areas in winter. These results show that the regression coefficients are also strikingly similar between the two types of spatial units. The R-squared value of the SLA analysis is 0.18, much larger than that for the POA of 0.009. This implies that the SLA regression explains the variation better than the POA result.

Table 4: Regression Results of Standardised Rates on PM_{10}

	POA				SLA			
	Value	Std. Error	t value	Pr(>|t|)	Value	Std. Error	t value	Pr(>|t|)
Intercept	1.233	0.353	3.494	0.001	0.756	0.454	1.665	0.108
PM_{10}	0.039	0.023	1.739	0.083	0.070	0.029	2.406	0.024
R-Squared	0.009				0.182			

In figure 49 the regression coefficient is significantly positive for the SLA result but the POA result is not and the 5% confidence level is negative.

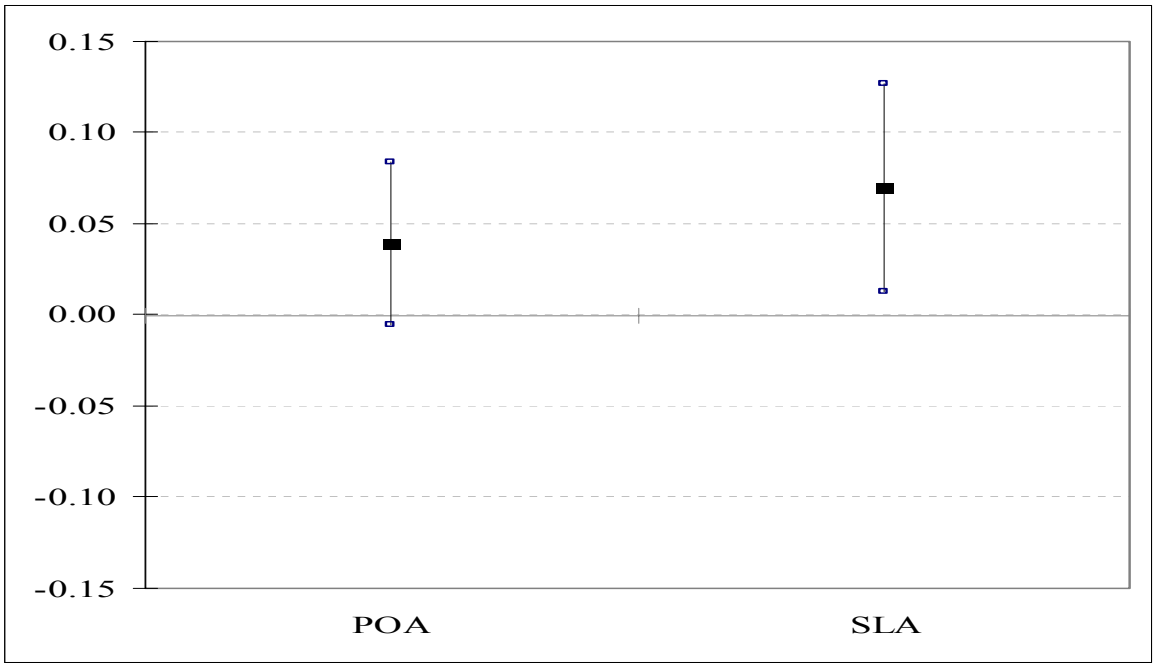


Figure 49: Comparison of PM_{10} Regression Coefficients and 95% Confidence Intervals

In summary these regression results show that the regression coefficients are not significantly different depending on the type of spatial units used. The disadvantage index is negatively associated with winter CVD rates whereas the PM_{10} relationship is positive and both these results suggest that the 95% confidence limit is achieved only for the SLA-level of aggregation.

Weighted Linear Regression

As noted previously the regression of directly age standardised rates should be weighted. The inverse of the variance was used for these weighted regressions.

Socio-Economic Disadvantage in Winter

Table 5: Weighted Regression Results of Standardised Rates on Disadvantage

	POA				SLA			
	Value	Std. Error	t value	Pr(>|t|)	Value	Std. Error	t value	Pr(>|t|)
Intercept	2.954	0.331	8.933	<0.001	3.409	0.458	7.446	<0.001
Disadvantage	-0.001	0.0003	-4.243	<0.0001	-0.002	0.0004	-3.819	0.0003
R-Squared	0.043				0.158			

The results of the weighted regression of winter rates against disadvantage scores show that the regression coefficients and the p-values are similar for both units. This time the POA had a much higher R-squared value (0.043) than the unweighted regression result however it is still quite small if compared with the SLA regression R-squared of 0.158. This level of explanation is not high, but this is to be expected for these relationships that are confounded by many extraneous variables.

As in the unweighted regression the coefficients are not statistically different between the two types of spatial units, however they are now both more strongly negative, with the POA 95% confidence level further away from zero and the p-value now much lower at <0.0001 while the SLA p-value is 0.0003.

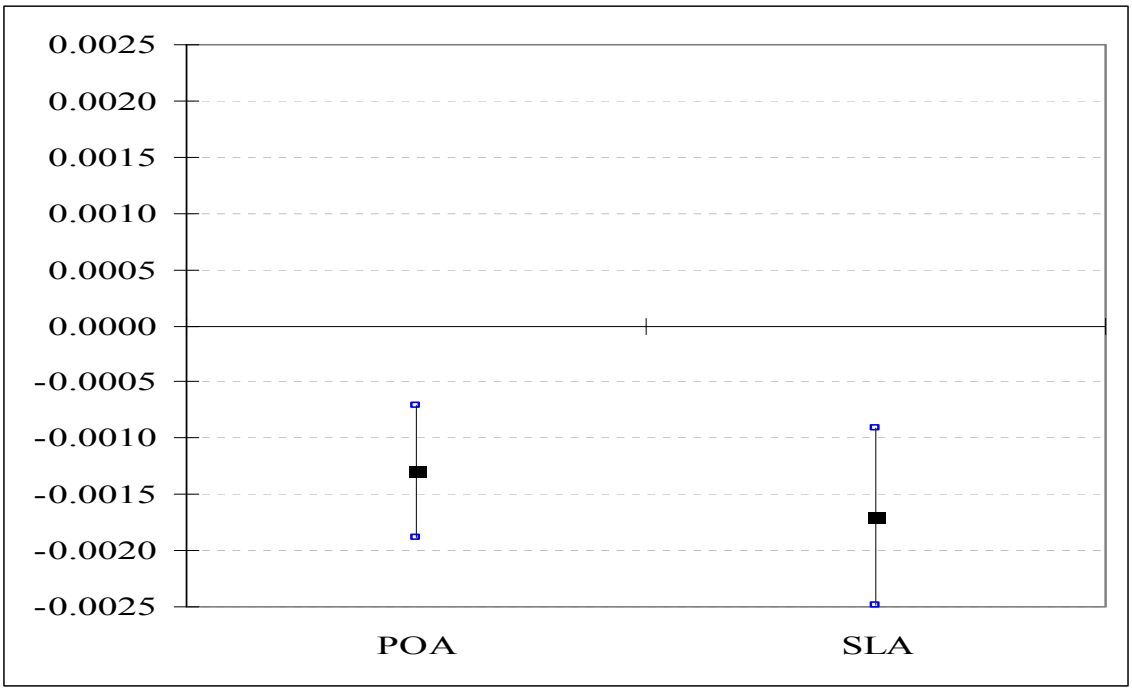


Figure 50: Comparison of Disadvantage Regression Coefficients and 95% Confidence Intervals

PM_{10} in Most Disadvantaged Areas in Winter

The weighted regression results for PM_{10} in the most disadvantaged areas in winter are shown in table 6. The regression coefficient for POA, whilst still positive, is now much smaller than it is for SLA and the p-value is very large at 0.63. The R-squared values are also appreciably different between the two units with POA achieving very low explanation (0.001) while SLAs have 0.23.

Table 6: Weighted regression results of standardised rates on PM_{10}

	POA				SLA			
	Value	Std. Error	t value	Pr(>|t|)	Value	Std. Error	t value	Pr(>|t|)
Intercept	1.477	0.279	5.290	0.000	0.574	0.451	1.273	0.214
PM_{10}	0.009	0.018	0.486	0.627	0.081	0.029	2.788	0.010
R-Squared	0.001				0.230			

In figure 51 the comparison between the coefficients shows that the POA level of aggregation finds no significant positive relationship whereas the SLA level does. This implies that the SLA result should be treated with caution when making inferences about the influence of this pollutant on CVD mortality rates.

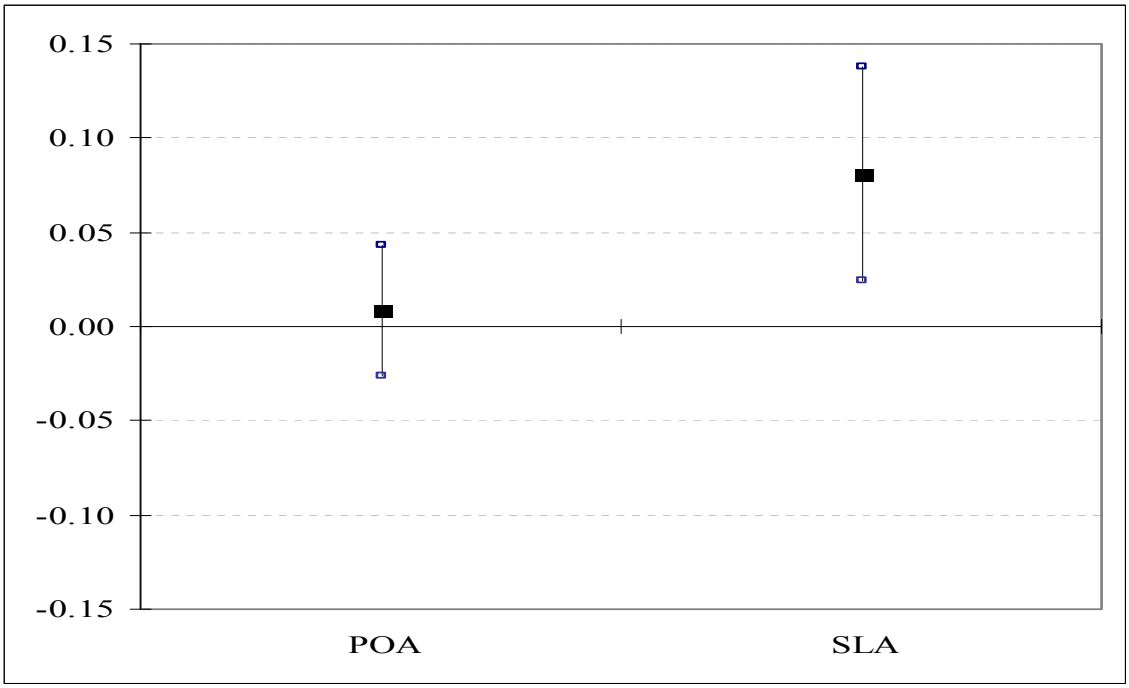


Figure 51: Comparison of PM_{10} Regression Coefficients and 95% Confidence Intervals

In summary, the data visualisation and regressions show that the disadvantage index has a negative relationship with CVD mortality whilst the PM_{10} concentrations have a positive relationship. The level of variation explained by the geographical overlay of SES and AP on CVD mortality rate is quite small, due to the large number of other influences unaccounted for (such as education, smoking or diet). They also show similarities in terms of direction of relationship with some key differences in the magnitude of the regression coefficient (the POA result is consistently lower than the SLA result) and the statistical confidence in the result (POA results are not significant for PM_{10} whereas SLA results are).

Discussion

In this discussion the focus is primarily on the differences found between the levels of aggregation in the different stages of the spatiotemporal modelling approach. As outlined in the methodology section the modelling progressed through quantification of SES, AP and CVD for the two types of spatial units at four consecutive time points. Then regression tree models established key relationships that were then investigated using stratified linear models. These were conducted for the directly age standardised rates with and without weighting to account for the different variances of the differently sized areas. The understanding of these relationships gained by these models are explored.

The Relationships

The relationships identified using the regression tree models that defined the highest death rates were found to be in winter between PM_{10} levels and CVD rates in the most disadvantaged areas.

The stratified linear regression models found that the relationships between the disadvantage index, PM_{10} and winter CVD mortality rates were slightly different between the two types of spatial units.

Regression Tree Models

The first split the tree models made was on season as summer rates are lower than winter rates (figures 28 and 30). There are various reasons why summer rates might be low that include; higher levels or different spatial patterns of pollution; higher levels of infectious diseases in winter that interact with CVD; and the possible protective effect of vitamin D from sunlight on this class of diseases (Scragg 1981).

The variables used in winter are the same between type of units and therefore the rules defining high CVD mortality rates can be characterised in the statement: "relatively disadvantaged areas with high PM_{10} concentrations are subject to higher rates". This shows that the direction of the relationship is consistent with those identified from the literature. The lowest death rates for POA were defined by scores on the disadvantage index greater than 1125 but in SLAs the areas with disadvantage scores above 1009 *but below 1025* had the lowest rates.

The PM_{10} relationship had split levels well below the NEPM level of 50μg/m3 (National Environmental Protection Council 2001) and this is consistent with recent studies that suggest effects of PM_{10} on CVD mortality at the internationally low levels experienced in Sydney (Morgan et al. 2003).

Table 7 summarises the use of the different variables in each of the two tree models. For simplicity, only the variable's first entry to the model is described. The POA tree used all the variables and achieved a reduction in deviance of 20% whereas the SLA tree only used some of the variables and explained 52% of the variation.

Table 7: Summary of Tree Model Results for the Two Types of Spatial Units

Units	Theme	Variable	Yes/No	Level / branch	Split	Direction
SLA	SEASON	Winter/Summer	Yes	First	Winter	High
	SES	Disadvantage	Yes	Second / winter	1009	Negative
		Urban Advantage	No			
		Education-Occupation	No			
		Economic Resources	Yes	Second / summer	1075	Negative
	AP	NO_2	No			
		PM_{10}	Yes	Third / winter with low disadvantage	14.1 µg/m^3	Positive
		O_3	No			
POA	SEASON	Winter/Summer	Yes	First	Winter	High
	SES	Disadvantage	Yes	Second / winter	1125	Negative
		Urban Advantage	Yes	Third / summer with low Edu-Occu	986	Negative
		Education-Occupation	Yes	Second / summer	1145	Negative
		Economic Resources	Yes	Sixth / winter, low dis, low PM, high O3 and high NO2	1144	Positive
	AP	NO_2	Yes	Fifth/ winter, low dis, low PM10, high O3	1.7 pphm	Positive
		PM_{10}	Yes	Third/winter with low disadvantage	17.6 µg/m^3	Positive
		O_3	Yes	Fourth/winter, low dis and low PM10	1.8 pphm	Positive

Stratified Linear Regression Models

The results of the stratified linear regression for the rules defining highest mortality rate endgroups were comparable. The sizes of the regression coefficients are similar between the spatial units.

Disadvantage

The disadvantage index displayed a negative relationship for both types of spatial units in winter. This means that more disadvantaged areas (with low disadvantage scores) have higher CVD rates than less disadvantaged areas. The POA results have a shallower slope than the SLAs results.

The difference between the p-values of the POA and SLA regression coefficients is interesting for the interpretation of the winter disadvantage relationship. The unweighted POA regression coefficient is less significant than that found for the SLA-level regression. But the weighted regression coefficient for POA was more significant ($p = <0.0001$) than the SLAs ($p = 0.0003$) implying that the POA-level view of this relationship shows a stronger influence than the SLA-level view.

PM_{10}

PM_{10} had a positive relationship with CVD rates in both types of spatial units (for disadvantaged areas in winter). However the p-value was not significant for POA whereas it was significant for SLA.

This implies that the SLA-level view identified an influence of PM_{10} on CVD mortality whereas the POA-level did not. Therefore it might be assumed that SLAs pick up the PM_{10} relationship better than POA, or that this is a spurious association found from analysis at an inappropriate scale or using spatial units with inappropriate boundaries.

The Effect of Scale on Understanding

Spatial pattern can be thought of as broad regional trend overlayed by highly variable local patterns:

> Different spatial patterns of disease are revealed by analyses that use large spatial filters with analysis that use smaller filters. Different diseases have patterns that are interesting at different spatial scales (Rushton 1998: 67).

This means that understanding the geographical patterns of diseases and their causes may depend on the scale of observation and analysis of these.

The comparisons of the spatial units investigated in this study shows slight difference in regression coefficients for disadvantage but substantial difference in the statistical confidence of the results for PM_{10} depending on the scale. This may be because the disadvantage phenomenon operates at finer resolution than the ambient conditions of PM_{10} and that the influence of the large-scale phenomenon is lost when observed with fine scale units. Another reason this might occur is due to poor statistical power due to low numbers. It is possible that observation over more time periods (either using longer duration in total or smaller temporal units) may increase the ability to observe the PM_{10} relationships.

In addition to these issues, the CVD category used here is very broad and includes many sub-categories of diseases that may not be influenced by particulate matter. These would dilute the noticeable effect of any association that may exist between specific CVD types and PM_{10}.

Understanding of spatial patterns is further complicated because environmental epidemiology studies cannot expect high levels of explanation and definitive outcomes using simple models of the complex system. The level of explanation achieved by the tree models and stratified linear regression models is relatively high considering this limitation.

Ultimately the differences found in the relationships at the two sets of spatial units are minimal. This result is consistent with a recent study that assessed difference in relationships identified between hospitalisation that found a strong association between SES and hospitalisation at the CCD level in Perth, and weaker associations when the data are aggregated to POA and SLA areas (Glover et al. 2004). This suggests that similar relationships are discerned but that the strength of these varies depending on the scale of the spatial units.

Management Implications

The finding that the relationships at different scales in Sydney vary depending on scale of spatial units has strong implications for the management of health surveillance and support systems. The POA and SLA spatial units have different scale qualities that should be taken into account in observation and analysis. This result may aid researchers to choose appropriate spatial units from those available for investigating relationships.

Increasing the range of choice of spatial units is also recommended. The criteria for defining aggregation units might be population size or area (Armstrong et al. 1999, Glover et al. 2004). However the restricted access to small area data is a reality likely to continue and there must be some compromise between scale of phenomenon under investigation and scale of observation available for release. When analysts do not have input to the aggregation they should access as many different levels as possible to explore differences between results at each level of aggregation and type of spatial units.

The debate about access to small area data is ongoing. The laws have been the subject of some controversy. It has been suggested that the laws are ineffective in stopping unscrupulous intrusion of direct marketing and profit driven data mining (Needham 2005). In fact a British researcher has recently claimed the laws are damaging population health and deaths resulting from data protection can be equated to the damage done to society by child murders (Peto et al. 2004). The solution might come from geographical masking techniques (Armstrong et al. 1999), but for the time being, the Australian National Mortality database remains aggregated to SLAs and POAs. The effect of these aggregations should be taken into account when analysing these data.

The NEPM air pollution standards may need to be lowered in response to these (and other) findings that PM_{10} levels less than half the recommended level for concern are influencing the rates of CVD mortality. On the other hand the rates are only associated in the most disadvantaged group and so if these people were less disadvantaged there would not be the same problem. One management approach that addresses this would be to distribute the profits that the car manufacturers and factory owners receive from the source of the pollution to the people who are affected.

Conclusion

During the period November 1996 until October 1998 there were geographical patterns in Sydney that indicate relationships between the SEIFA disadvantage index, ambient PM_{10} concentrations and CVD mortality rates. These were analysed at two scales of aggregation; small POA and larger SLA. This study used exploratory data analysis and modelling techniques to reduce the complexity of these relationships, focusing attention on the effect of scale on the understanding of the geographical patterns. The results for each type of spatial units were compared and found to be similar yet slightly different. This implies that both sets of spatial units are useful for gaining understanding.

The results show that when health and environmental data are aggregated to larger spatial units variation is reduced; therefore smaller areas allow observation of fine resolution heterogeneity between parts of the city. The relationships identified, whilst similar between the spatial units, were different and this can be seen not only in: the order of importance and splitting level for each variable in the regression tree models (pages 47 to 50), but also in the spatial location of the groups defined by the tree decision rules (pages 51 and 52); the breadth of the scatter and shape of loess lines in the visualisation analysis (pages 54 to 58, also see indirectly standardised mortality ratios in appendix 2); and the magnitude of linear regression coefficients (pages 59 to 62).

Studies that use just one set of spatial units for analysing health-environment relationships might yield less (or different) understanding from those using multiple levels. Therefore to understand geographical patterns the effect of scale is very important to assess. It is necessary to explore the effect this has on analysis by conducting exploratory studies at different spatial grain size and configuration. For this the CCD level would be ideal because they are the smallest areas available that still protect confidentiality. However the "view" of geographical patterns might break down at this scale. The availability of data aggregated to these small areas would allow analysts to construct many different sized spatial units and in this manner it would be possible to assess the influence of both scale and zonation issues and to devise an optimal zoning system for the phenomena of interest.

The future of geographical health studies will include aggregate-level health and environment observations such as these in multi-level modelling (Jones and Duncan 1996). In these nascent statistical methods the information from different levels of aggregation (down to individuals) may be integrated and there are some examples where this strategy can bring extra predictive power, description and precision to efforts to understand patterns of health and environmental variation (Subramanian 2004).

In synthesis, this research found that as data are aggregated to different spatial units the relationships observed may change. In the case investigated here the significant negative relationship shown between the disadvantage index and winter CVD mortality rates by the different types of spatial units is quite similar, reflecting the strength of the underlying association. However the significant positive relationship with PM_{10} at the SLA level was called into question by the POA results that suggest the influence of this pollutant is minor, if indeed one exists. This disparity is due to the different spatial scales and shapes of the two types of spatial units.

Bibliography

Allen, T. F. H., and T. W. Hoekstra. 1992. Toward a Unified Ecology. Columbia University Press, New York.

Anselin, L. 2003. ACE 492 Spatial Analysis Course on-line Class Notes 6. Rate Maps and Smoothing at http://sal.agecon.uiuc.edu/courses/sa/index.html. Website accessed on 12/10/2004.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. Statistics in Medicine **18**:497-525.

Australian Bureau of Statistics. 2002. Australian Social Trends 2002: Health - Mortality and Morbidity: Cardiovascular disease: 20th century trends. http://www.abs.gov.au/Ausstats/abs@.nsf/0/4c9f3487e20b75caca256bcd008272f5?OpenDocument#Links. Website accessed on. 18/2/05.

Australian Bureau of Statistics, and Space-Time Research Pty. Ltd. 2002a. Population Estimates by Age and Sex,state and territory-specific datasets, 2001 (ABS Cat. No. 3235.0-8.55.001). *in*, Canberra, ACT.

Australian Bureau of Statistics, and Space-Time Research Pty. Ltd. 2002b. Population Estimates by Age and Sex,state and territory-specific datasets, 2002 (ABS Cat. No. 3235.0-8.55.001). *in*, Canberra, ACT.

Australian Bureau of Statistics, and Space-Time Research Pty. Ltd. 2002c. Population Estimates by Age by Sex, state and territory-specific datasets, 1991 and 1996 (ABS Cat. No. 3235.0-8.55.001). *in*, Canberra, ACT.

Ayers, G. P., M. Edwards, and J. L. Gras. 2001. Fine Particle Measurement Study Phase B ~ Data Analysis. Prepared for Environment Australia by CSIRO Atmospheric Research.

Bailey, T. C., and A. C. Gatrell. 1995. Interactive spatial data analysis. Longman Scientific & Technical, New York ; Burnt Mill, Harlow.

Bennett, S. 1996. Socioeconomic inequalities in coronary heart disease and stroke mortality among Australian men, 1979-1993. Int J Epidemiol **25**:266-275.

Berkman, L. F., and I. Kawachi. 2003. Neighborhoods and health. Oxford University Press, Oxford ; New York.

Blakely, T. A., and A. J. Woodward. 2000. Ecological effects in multi-level studies. J Epidemiol Community Health **54**:367-374.

Breiman, L. 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, Calif.

Briggs, D. J., S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. Vanreeuwijk, K. Smallbone, and A. Vanderveen. 1997. Mapping Urban Air Pollution Using GIS - a Regression-Based Approach. International Journal of Geographical Information Science **11**:699-718.

Briggs, D. J., C. de Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone. 2000. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Sci Total Environ **253**:151-167.

Briggs, D. J., and K. Field. 2000. Using Geographic Information Systems to Link Environment and Health Data. *in* C. Corvalan, D. J. Briggs, and G. Zielhuis, editors. Decision Making in Environmental Health. E & FN Spon, London and New York.

Burnett, R., R. Ma, M. Jerrett, M. S. Goldberg, S. Cakmak, C. A. I. Pope, and K. D. 2001. The spatial association between community air pollution and mortality: a new method of analyzing correlated geographic cohort data. Environ Health Perspect **109**:375-380.

Cleek, R. K. 1979. Cancer and the Environment: the Effect of Scale. Social Science and Medicine, 13D, 241-247. ix, 340. *Cited in* M. S. Meade and R. J. Earickson, editors. Medical Geography, Second Edition (2000). Guilford Press, New York.

Clements, M. 2005a. Personal Communication: Calculating the Variance for Areal Directly Age Standardised Mortality Rates. *in*, NCEPH, Canberra, ACT.

Clements, M. 2005b. Personal Communication: Generalised Linear Modelling of Indirectly Age Standardised Mortality Rates. *in*, NCEPH, Canberra, ACT.

Diez Roux, A. V. 2004. Estimating neighborhood health effects: the challenges of causal inference in a complex world. Soc Sci Med **58**:1953-1960.

D'Souza, R. 2005. NZHIS ICD9-10 mapping file errors. *in*, NCEPH, Canberra.

Dungan, J. L., J. N. Perry, M. R. T. Dale, P. Legendre, S. Citron-Pousty, M. J. Fortin, A. Jakomulska, M. Miriti, and M. S. Rosenberg. 2002. A balanced view of scale in spatial statistical analysis. Ecography **25**:626-640.

Elliott, P., and D. Wartenberg. 2004. Spatial epidemiology: Current approaches and future challenges [Review]. Environmental Health Perspectives **112**:998-1006.

Fotheringham, A. S., and D. W. S. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis, Environment and Planning A, cited in Green, M., Flowerdew, R. (1996) New evidence on the modifiable areal unit problem. *Cited in* P. Longley and M. Batty, editors. Spatial analysis: Modelling in a GIS environment. GeoInformation International, Cambridge.

Gilbert, N. L., S. Woodhouse, D. M. Stieb, and J. R. Brook. 2003. Ambient nitrogen dioxide and distance from a major highway. The Science of The Total Environment **312**:43-46.

Glover, J., D. Rosman, and S. Tennant. 2004. Unpacking analyses relying on area-based data: are the assumptions supportable? Int J Health Geogr **3**:30.

Gould-Ellen, I., T. Mijanovich, and K. N. Dillman. 2001. Neighborhood effects on health: Exploring the links and assessing the evidence. Journal of Urban Affairs **23**:391-408.

Green, M., and R. Flowerdew. 1996. New evidence on the modifiable areal unit problem. *in* P. Longley and M. Batty, editors. Spatial analysis: Modelling in a GIS environment. GeoInformation International, Cambridge.

Gregory, I. N., D. Dorling, and H. R. Southall. 2001. A century of inequality in England and Wales using standardized geographical units. Area **33**:297-311.

Gunderson, L. H., and C. S. Holling. 2002. Panarchy; Understanding Transformations in Human and Natural Systems. Island Press, Washington, DC.

Hyndman, J. C., C. D. Holman, R. L. Hockey, R. J. Donovan, B. Corti, and J. Rivera. 1995. Misclassification of social disadvantage based on geographical areas: comparison of postcode and collector's district analyses. Int J Epidemiol **24**:165-176.

Jones, K., and C. Duncan. 1996. People and places: The multilevel model as a general framework for the quantitative analysis of geographical data. Pages 79–104 *in* P. Longley and M. Batty, editors. Spatial analysis: Modelling in a GIS environment. Geoinformation Group, Cambridge:.

Julious, S. A., J. Nicholl, and S. George. 2001. Why do we continue to use standardized mortality ratios for small area comparisons? Journal of Public Health Medicine **23**:40-46.

Kawachi, I. 2004. Neighborhoods and Health. *in* Methodology and Training Seminar Series, National Centre for Epidemiology and Population Health, Canberra, ACT.

Kawachi, I., and M. S. O'Neill. 2005. Exploration of health disparities. Environmental Health Perspectives : 100-107.

Kay, J. J., H. A. Regier, M. Boyle, and G. Francis. 1999. An ecosystem approach for sustainability: addressing the challenge of complexity. Futures **31**:721-742.

Krieger, N. 1994. Epidemiology and the Web of Causation - Has Anyone Seen the Spider? Social Science & Medicine **39**:887-903.

Krieger, N., P. Waterman, D, , J. Chen, T, , D. Rehkopf, H, , and S. Subramanian, V,. 2004. Geocoding and monitoring US socioeconomic inequalities in health: an introduction to using area-based socioeconomic measures. Harvard School of Public Health. Boston, MA: Available at: http://www.hsph.harvard.edu/thegeocodingproject/. Website accessed on 11/9/2004..

Lancaster, G., and M. Green. 2002. Deprivation, ill-health and the ecological fallacy. Journal of the Royal Statistical Society Series A-Statistics in Society **165**:263-278.

Marceau, D. J. 1999. The scale issue in social and natural sciences. The Canadian Journal of Remote Sensing **25**:347-356.

McCracken, K. 2001. Into a SEIFA SES cul-de-sac? Aust N Z J Public Health **25**:305-306.

McPhail, S. 1996. The Metropolitan air quality study Implications for policy development. Pages 91-95 *in* Health and Urban Air Quality in NSW Conference. The NSW Health Department, Sydney.

Meade, M. S., and R. J. Earickson. 2000. Medical Geography, Second Edition. Guilford Press, New York.

Miller, J., and J. Franklin. 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling **157**:227-247.

Moisen, G. G., and T. S. Frescino. 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling **157**:209-225.

Monmonier, M. S. 1996. How to lie with maps, 2nd edition. University of Chicago Press, Chicago.

Morgan, G., S. Corbett, J. Wlodarczyk, and P. Lewis. 1998. Air pollution and daily mortality in Sydney, Australia, 1989 through 1993. Am J Public Health **88**:759-764.

Morgan, G., D. Lincoln, V. Sheppeard, B. Jalaludn, J. F. Beard, R. W. Simpson, A. Petroeschevsky, T. O'Farrell, and S. Corbett. 2003. The Effects Of Low Level Air Pollution On Daily Mortality And Hospital Admissions In Sydney, Australia, 1994 To 2000. Epidemiology. **14**.

National Environmental Protection Council. 2001. Ambient Air Quality NEPM. http://www.ephc.gov.au/nepms/air/air_nepm.html. Website accessed on 16/5/2004

National Health and Medical Research Council. 1999. National Statement on Ethical Conduct in Research Involving Humans. Canberra, ACT. http://www7.health.gov.au/nhmrc/publications/pdf/e35.pdf. Website accessed on 2/7/2004

Needham, K. 2005. Privacy law fears hampering cancer research. *in* The Sydney Morning Herald, January 8, 2005, Sydney, NSW, Australia.

NEPC Peer Review Committee. 2001. Collection and Reporting of TEOM PM10 Data. Technical Paper No. 10.

New Zealand Health Information Service. 2004. Third mapping fix in ICD-10-AM to ICD-9-CM-A mapping. *in* http://www.nzhis.govt.nz/documentation/mapping/mappingfiles.html. Website accessed on 11/3/2004.

NSW Department of Environment and Conservation. 2001. National Environment Protection Measure for Ambient Air Quality: Air Monitoring Plan for NSW - June 2001.

Nurminen, M. 1997. The Ecological Method: Linkage Failures and Bias Corrections. Pages 21-72 *in* C. Corvalan, M. Nurminen, H. Pastides, editors. Linkage methods for environment and health analysis : technical guidelines. A report of the health and environment analysis for decision-making (HEADLAMP)

project. UNEP, USEPA, and Office of Global and Integrated Environmental Health, World Health Organization, Geneva.

Nurminen, M., and D. J. Briggs. 1996. Approaches to Linkage Analysis: Overview. Pages 93-119 *in* D. J. Briggs. Linkage methods for environment and health analysis: general guidelines. A report of the health and environment analysis for decision-making (HEADLAMP) project. UNEP, USEPA, and Office of Global and Integrated Environmental Health World Health Organization, Geneva.

Nurminen, M., and T. Nurminen. 2000. Methodologic issues in linking aggregated environmental and health data. Environmetrics **11**:63-73.

O'Neill, M. S., M. Jerrett, L. Kawachi, J. L. Levy, A. J. Cohen, N. Gouveia, P. Wilkinson, T. Fletcher, L. Cifuentes, and J. Schwartz. 2003. Health, wealth, and air pollution: Advancing theory and methods [Review]. Environmental Health Perspectives **111**:1861-1870.

Openshaw, S. 1983. The modifiable areal unit problem. Geo Books, Norwick.

Openshaw, S., and P. J. Taylor. 1981. The modifiable areal unit problem. Pages 60-69 *in* N. Wrigley and R. J. Bennett, editors. Quantitative geography : a British view. Routledge & Kegan Paul, London ; Boston.

Pearce, N. 2000. The ecological fallacy strikes back. Journal of Epidemiology & Community Health **54**:326-327.

Peto, J., O. Fletcher, and C. Gilham. 2004. Data protection, informed consent, and research. BMJ **328**:1029-1030.

Pikhart, H., M. Bobak, P. Gorynski, B. Wojtyniak, J. Danova, M. A. Celko, B. Kriz, D. Briggs, and P. Elliott. 2001. Outdoor sulphur dioxide and respiratory symptoms in Czech and Polish school children: a small-area study (SAVIAH). Small-Area Variation in Air Pollution and Health. Int Arch Occup Environ Health **74**:574-578.

Rixom, A. 2002. Performance league tables BMJ **325**:177-178.

Rushton, G. 1998. Improving the Geographic Basis of Health Surveillance using GIS. Pages xvii, 212 *in* A. C. Gatrell, M. Lèoytèonen, and European Science Foundation., editors. GIS and Health. Taylor & Francis, London ; Philadelphia, PA.

Scoggins, A., T. Kjellstrom, G. Fisher, J. Connor, and N. Gimson. 2004. Spatial analysis of annual air pollution exposure and mortality. Science of the Total Environment **321**:71-85.

Scragg, R. 1981. Seasonality of cardiovascular disease mortality and the possible protective effect of ultra-violet radiation. Int J Epidemiol **10**:337-341.

Shannon, W. D., M. Faifer, M. A. Province, and D. C. Rao. 2002. Tree-based models for fitting stratified linear regression models. Journal of Classification **19**:113-130.

Shepherd, N. 1996. The Metropolitan air quality study Implications for policy development. Pages 88-90 *in* Health and Urban Air Quality in NSW Conference. The NSW Health Department, Sydney.

Stieb, D. M., S. Judek, and R. T. Burnett. 2002. Meta-analysis of time-series studies of air pollution and mortality: effects of gases and particles and the influence of cause of death, age, and season. J Air Waste Manag Assoc **52**:470-484.

Stocks, N. P., P. Ryan, H. McElroy, and J. Allan. 2004. Statin prescribing in Australia: socioeconomic and sex differences. A cross-sectional study. Med J Aust **180**:229-231.

Subramanian, S. V. 2004. The relevance of multilevel statistical methods for identifying causal neighborhood effects. Soc Sci Med **58**:1961-1967.

Taylor, R. 1998. Standardisation. *in* C. B. Kerr, R. Taylor, and G. Heard, editors. Handbook of public health methods. McGraw Hill, Sydney.

Taylor, R., T. Chey, A. Bauman, and I. Webster. 1999. Socio-economic, migrant and geographic differentials in coronary heart disease occurrence in New South Wales. Aust N Z J Public Health **23**:20-26.

Turrell, G., and C. Mathers. 2001. Socioeconomic inequalities in all-cause and specific-cause mortality in Australia: 1985-1987 and 1995-1997. Int J Epidemiol **30**:231-239.

Turrell, G., and C. D. Mathers. 2000. Socioeconomic status and health in Australia. Med J Aust **172**:434-438.

White, D., and J. C. Sifneos. 2002. Regression tree cartography. Journal of Computational and Graphical Statistics **11**:600-614.

World Health Organization. 2000. Air Quality Guidelines for Europe, Second Edition. European Series, No. 91 World Health Organization, Regional Office for Europe, Copenhagen.

Wrigley, N., T. Holt, D. Steel, and M. Tranmer. 1996. Analysing, modelling and resolving the ecological fallacy. *in* P. Longley and M. Batty, editors. Spatial analysis: Modelling in a GIS environment. GeoInformation International, Cambridge.

Yassi, A., Kjellstrom T, de Kok T, and T. Guidotti. 1998. Basic environmental health. World Health Organisation WHO, Geneva.

Yu, X. Q., C. Robertson, and I. Brett. 2000. Socioeconomic correlates of mortality differentials by local government area in rural northern New South Wales, 1981-1995. Aust N Z J Public Health **24**:365-369.

Zhu, Y. F., W. C. Hinds, S. Kim, and C. Sioutas. 2002. Concentration and size distribution of ultrafine particles near a major highway. Journal of the Air & Waste Management Association **52**:1032-1042.

Appendix 1: Health and Population Data

Ethics

Approval was obtained from the NCEPH ethics committee and the ANU human research ethics committee to access the National Mortality Database. This is a requirement in the NCEPH's agreement with the ABS who sell the database. This study was approved as protocol 2004/272, signed by the ANU committee chairman Dr Peter Hiscock on the 24/9/2004.

Melissa Goodwin, the NCEPH data manager, extracted the data for all causes of death for the Sydney SD for the period 1995-2002. The mortality data for 1995 until 2002 used the International Classification of Disease (ICD) system to define cause of death. ICD9 is used till 1998 and ICD10 from 1999 till the present. The CVD category was defined here by the codes 390.0 - 459.9 in ICD9 (Morgan et al. 1998). In ICD 10 (1999 till present) codes: G45.0 - G46.8, I00.0 - I99.9, M30.0 - M31.9, N28.0, R00.0 - R03.1, R58.0 were identified (New Zealand Health Information Service 2004). Not all of these are actually CVD conditions and some of the deaths registered in later years may not be CVD (some registrations may be delayed for years after the event) (D'Souza 2005) personal communication). This issue of excess cases was assessed and it was found that only minor errors are introduced and primarily for the later part of the period (1998-2002). As the main focus is on 1996-98 and the records predominantly hold the ICD 9 codes the adjustment needed would therefore not have a significant effect

Concordance

Areal interpolation is used for the integration of data geocoded to different spatial units (Briggs and Field 2000, Gregory et al. 2001). Both the populations and health records coded to spatial units may need to be recoded to match between different editions of spatial units when these change periodically as they do in the NMD. SLA concordances for coding recent registrations back to old SLAs were required for this study in Sydney for the 96-98 period. This was merely a matter of merging three Blacktown SLAs into one, and changing a few codes. A table that has the old codes next to their new ones is made (in splits the old codes are duplicated in the first column whilst a merge would duplicate new codes next to old ones). In a relational table, the new codes are joined to the data from the recent period and then summing across the merged parts.

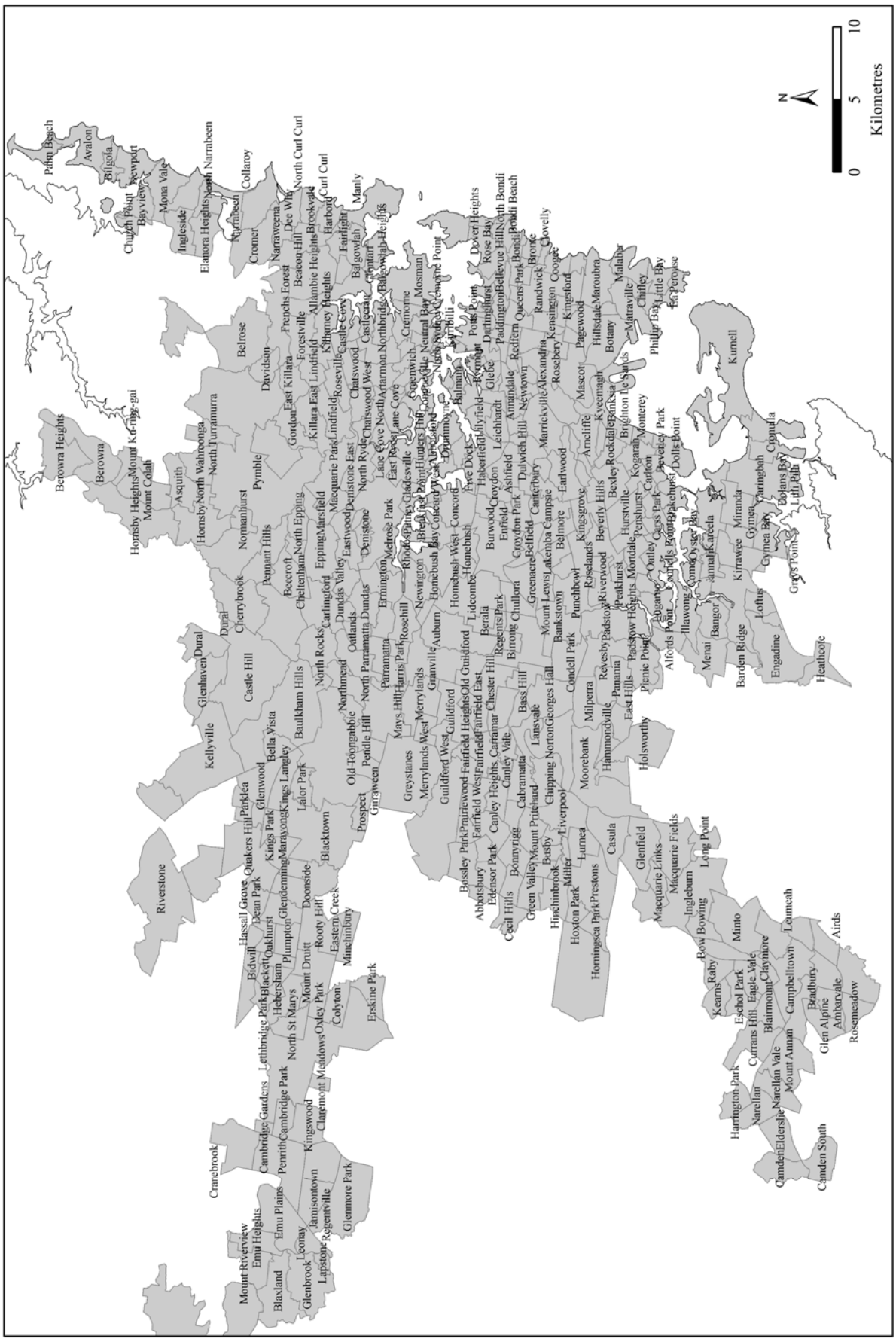
Postal area selection methods:

Criteria were:

1. Population density and urban boundary, and
2. Multi-part polygons with parts greater than 5km apart excluded.

The POA spatial boundaries were converted from AGD 66 to AMG zone 55 so that the coordinates are in metres. Using Pythagoras' theorem (the hypotenuse is the straight-line between any two centroids and the sides are x1 – x2 and y1 – y2) distance between the centroids of each part were calculated and any parts over 5km distant excluded.

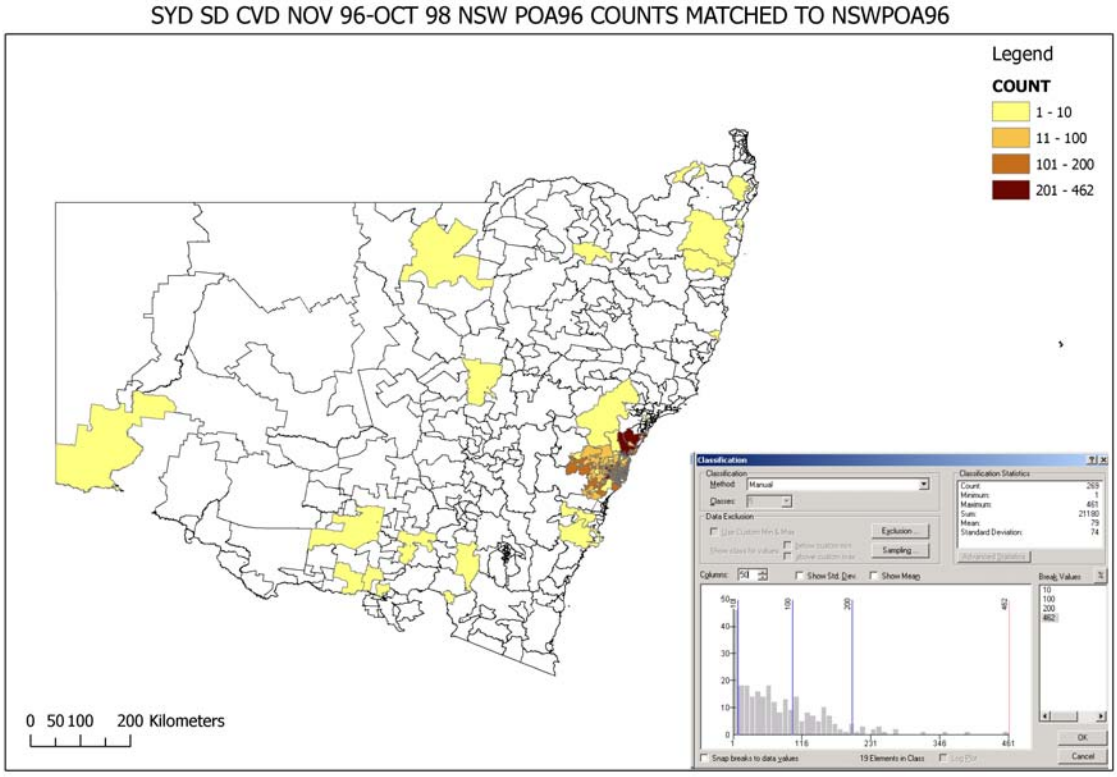
The population data used to calculate these daily rates were interpolated from the resident population in 1991, 1996, 2001 and 2002 (Australian Bureau of Statistics and Space-Time Research Pty. Ltd. 2002c,, 2002a,, 2002b). SEIFA access was by arrangement with Melissa Goodwin (NCEPH data manager) and Prof. Jim Butler (NCEPH deputy director).



Map of Sydney Suburbs

There were some records coded to postcodes that did not exist in the POA boundaries file. These may be postcodes that are not included due to being too small for any CCD to be designated to it or for some other reason (such as the postcode 2253 that is used for wharves at Fisherman's Point, Collingridge Point, and Little Wobby Beach along the delivery route of a boat, whereas the areas in between are given postcodes 2159, 2250 and 2083). These postcodes are locatable using the Australia Post database and the online 'whitepages' directory, however as no population data are available there is no meaningful information to be gained apart from their spatial location. It is possible to merge these cases into the POA in which they are located, however this is laborious and was not prepared for the exploratory study presented here

Another issue in the POA and SLA coding is mismatch between location of these for the same death. It was found that the NMD had records identified in SLAs in Sydney but by POA in other regions, as far removed as western NSW. The extent of this mismatch for the 1996-98 period is shown in figure?



These records were excluded from the analysis.

It was then discovered that the set of records that had both SLA and POA in Sydney had some who were recorded in postcodes not located in the SLA but as far away as the other side of the city. The extent of this problem may be identified using a concordance that is available from the ABS, however it was found that the 1996 SLA to POA concordance table also had POA coded to the incorrect SLA. This issue was not addressed after exploration showed that the majority of POA codes are overlaying the correct SLA.

Appendix 2: Standardised Mortality Rates

Standardisation Explained

Standardisation is a set of methods for adjustment of data for the influence of confounders often employed to compare populations or the same population across time. (Taylor 1998:275). The problem stems from the different proportion of each population that is at risk of a health event occurring to them. As an example, if a health outcome is most likely to occur at a certain age, then the population at risk is those people of that age in an area. The difference in age structure between study populations makes the comparison of their health status problematic. This common problem may be dealt with by age standardisation. There are two distinct methods for this called the direct and indirect methods and these are summarised below.

Age-specific rates

The age specific rate (ASR) is the number of cases of the disease in question observed in a specified age group of the study population divided by the number of people of that age in the same place and time (Meade and Earickson 2000:414). This can be expressed as the equation:

$$ASR = \frac{\text{Observed number of events in a specified population of people aged x}}{\text{Total population aged x during the specified time in that place}}$$

The most precise rate would be calculated by using the number of people at risk of the same health event happening to them (for example using the number of women of reproductive age in a population to calculate fertility rate from the number of births in that population (Meade and Earickson 2000:404). However in many cases the number of persons at risk of a health event in a given place and time is not available (for example exposure to air pollutants with high spatial variation). The total resident population is often used in these situations.

If a period of time is being analysed then the number of years each person lived at risk in the specified place and time needs to be included in the measure. To do this one must calculate the amount of person time lived, often expressed in Person-Years Lived (PYL). This can either be estimated by multiplying the mid-period number of people (as from a periodic census) by the number of years the cohort was followed, or by interpolating the annual populations between censuses and summing the numbers for each year. If the period is less than one year then the person-days lived may be calculated by multiplying the census count of population by the proportion of the year that has elapsed.

Direct Age Standardisation

Direct age standardisation is a method used to calculate a standard rate (and rate ratio) that is often used to compare and map health data (Meade and Earickson 2000:416). This method first calculates an ASR for each age group in the study population(s). Then these are multiplied by the populations of each age group in some standard population (often a state or country total) to calculate the expected number of cases. The sum of the expected numbers in each age group is then divided by the total standard population to give the standard rate. The standard rate can be understood as the number of deaths per person that would have been expected had the "study populations, given their separate [health] experiences, had the same age structure" (Meade and Earickson 2000:414).

The directly age adjusted standard rate ratio is calculated by dividing the standard rate by the crude rate of the standard population (calculated by dividing the total number of cases in the standard population by the total number of people in that population).

Indirect Age Standardisation

To indirectly age standardise, the ASRs of the standard population are calculated first. Then the population of each age group in the study population is multiplied by these rates. This gives the expected number of cases if the study population had the same health experience of the standard, given its age structure. Dividing the number of observed by this number of expected gives the standard ratio. If this standard ratio is multiplied by the crude rate of the standard population this gives the indirectly age standardised rate (Meade and Earickson 2000:415).

To calculate this measure for a period one must first calculate an expected number of cases which can be done indirectly by multiplying the PYL in each age cohort for the period by the ASR of the standard population in that age group. The observed number of events for the period may then be divided by the expected number to give an age standardised ratio for the period (Yassi et al. 1998:91)

The indirect method is not appropriate if the age structure of the study populations vary from that of the standard population.

> "The main advantage of the indirect method is that it is not necessary to know the age distribution of observed cases or events in the study populations..... other advantages include the ease of calculation of confidence intervals using the Poisson method and the production of a ratio statistic in a one step procedure. A disadvantage of the indirect method is that it is generally considered to be less precise in adjusting for age than the direct method, particularly when the age structure of the study population is radically different from that of the standard…. Another disadvantage is that it is usually considered that SMRs or SIRs from study populations can only be legitimately compared with the standard and not with each other; however Armitage and Berry (1994) provide a method of calculation of confidence intervals that would permit such intergroup comparisons". (Taylor 1998:277-278).

There is still debate:

> "some authors advocate direct standardisation as it involves adjustment to a common standard (Julious et al., 2001) – in practice in our own experience, the two methods nearly always give near-identical results." (Elliott and Wartenberg 2004:11).

The claim by Elliott that this nearly always doesn't matter is not convincing. The paper by Julious (2001) has a really good discussion and strongly argues that indirect method can introduce bias, and is backed up by Rixom (2002). If we are to use it the standard rates should come from a population close to the make up of the study populations.

Comparison of the direct and indirect methods

There is some debate between medical geographers and epidemiologists about which method is the best to use when age standardising. Meade and Earickson (2000)suggest using direct methods for mapping incidence rates/ratios as comparisons made of these are more valid than those based on the indirect method:

"Biostatisticians find indirect standardisation so clearly invalid that the equation is not presented here. The problem is that the SMR is actually a weighted average in which the weights are derived from the study population, and so the standard is always the experience structure of the study population. The idea that indirect standardisation is based on the non-study population as the common standard is a misconceptionwhich results in a common methodological error: mapping SMRs [calculated by the indirect method] for comparison, although they do not share a standard population. Whenever age specific rates are available, which is almost always the case today, a mappable ratio can be easily developed: the standard rate ratio [calculated by the direct method]" (Meade and Earickson 2000:415-416).

The opposite position is taken in Yassi et al (1998) who do not describe the direct approach in their book "Basic Environmental Health". They present the calculation of a standardised mortality ratio as the observed number of deaths in the study population divided by the expected number of deaths (calculated by multiplying the PYL by the rate for the disease(s) in the age cohort being considered in some standard population) (Yassi et al. 1998:91). The authors do not mention that this is an indirect method, nor do they mention the fact that there is an alternative direct method.

A good discussion of the differences between these methods is given in (Taylor 1998:275). Describe the strengths and weaknesses of each method and provided the following table.

Type (standard)	Advantages	Problems
Direct (population)	Precise Valid for intergroup comparisons Useful for international comparisons if an agreed standard Can be used on a linear scale The same standard can be used for multiple different diseases	Unstable with small numbers Poisson confidence intervals not straight forward
Indirect (rate)	Can be used in absence of age specific data Produces a ratio to the standard in one step Poisson confidence limits straightforward	Less precise with large differences in age structure Valid for study population versus standard only unless special statistical techniques employed Ratios should be used only on a log scale Multiple standards are required for different diseases

It can be argued that both methods may be employed in appropriate situations. The choice of which standardisation method to use depends on:

a) availability of age specific data for the study population and

b) the size of the groups (or areas) to be standardised.

Where the study populations are small, the indirect method is considered more stable. This is because the age specific rates of the standard population are based on larger numbers than those of the study population, which may be small and fluctuate widely (Anselin 2003).

Another approach would be aggregating small study populations (by space, time, disease class, age group or some other group) to provide larger numbers. It seems counterintuitive to apply the national rates of diseases to local populations that obviously will have different social, environmental and health experiences than some national average experience (what ever this might mean).

Standardisation method conclusion

The proposed study will make use of both the indirect and direct methods of age standardisation and assess the usefulness and appropriateness of each to the various types of health data available.

Calculating Mortality Rates in Small Areas in Sydney

The Australian National Mortality Database

The data is collected by State registrars of births, deaths and marriages, they are the ones who code the data to SLA/Postcode of usual residence. SLA has been used for a long time but recently NSW, VIC and QLD have included postcode in an un-used field called Xregyy (this had been intended for extra information about the registration year but never had anything in it). They send it to the ABS who then sell it on to AIHW and the NCEPH. NCEPH researchers are required to submit application to the ANU ethics committee before using the data. The ethics documents will be attached at the end of this appendix.

Summary of Method:

Extract and reformat deaths in Sydney and analyse all causes, non-external, cardiovascular and respiratory. Choose disease.

1. Describe chosen disease through the years 95-02
2. Choose years/seasons
3. Stratify by area and age
4. Join with Usual Resident Population by area
5. Calculate age specific rates by area
6. Join to standard population
7. Calculate the expected cases by area
8. Sum exp cases and divide by total persons in standard population and multiply by 1000 = standard rate per 1000

S-plus syntax for linear regression models

UNWEIGHTED

```
lm(Standardised.rate ~ Disadvantage, data = “POAorSLA”, subset =
SEASON == WINTER, na.action = na.exclude)

lm(Standardised.rate ~ PM10, data = “POAorSLA”, subset = SEASON ==
WINTER & Disadvantage < “threshold”, na.action = na.exclude)
```

WEIGHTED

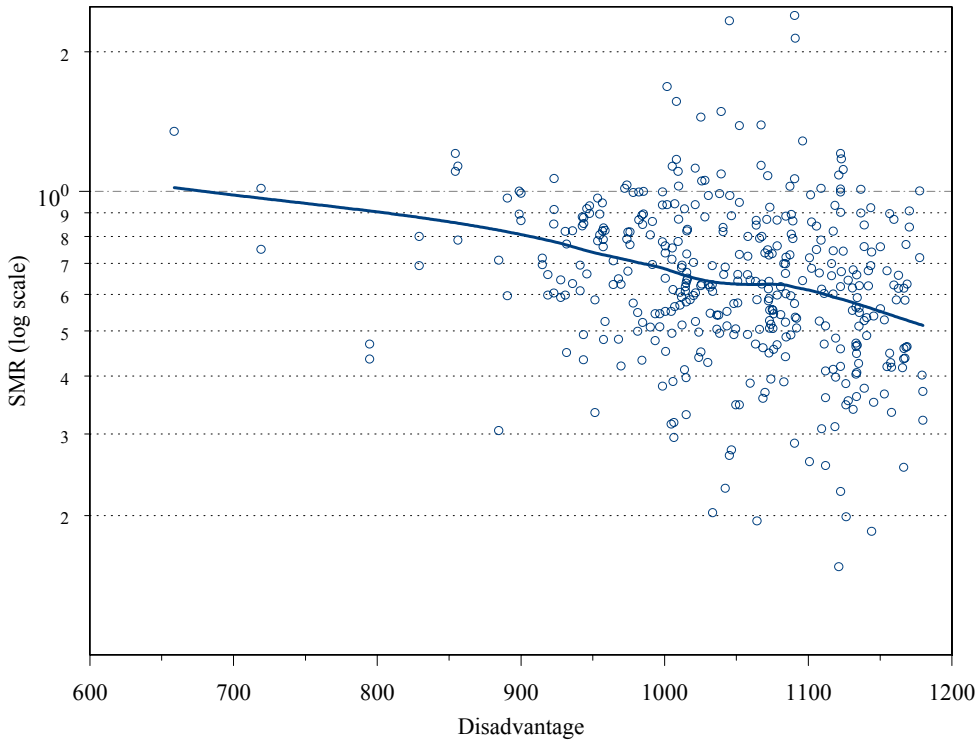
```
lm(Standardised.rate.per.1000 ~ Disadvantage, data = ”POAorSLA”,
weights = Weight, subset = SEASON == "WINTER", na.action = na.exclude)
```

```
lm(Standardised.rate.per.1000 ~ PM10, data = ”POAorSLA”, weights =
Weight, subset = SEASON == "WINTER" & Disadvantage < “threshold”,
na.action = na.exclude)
```

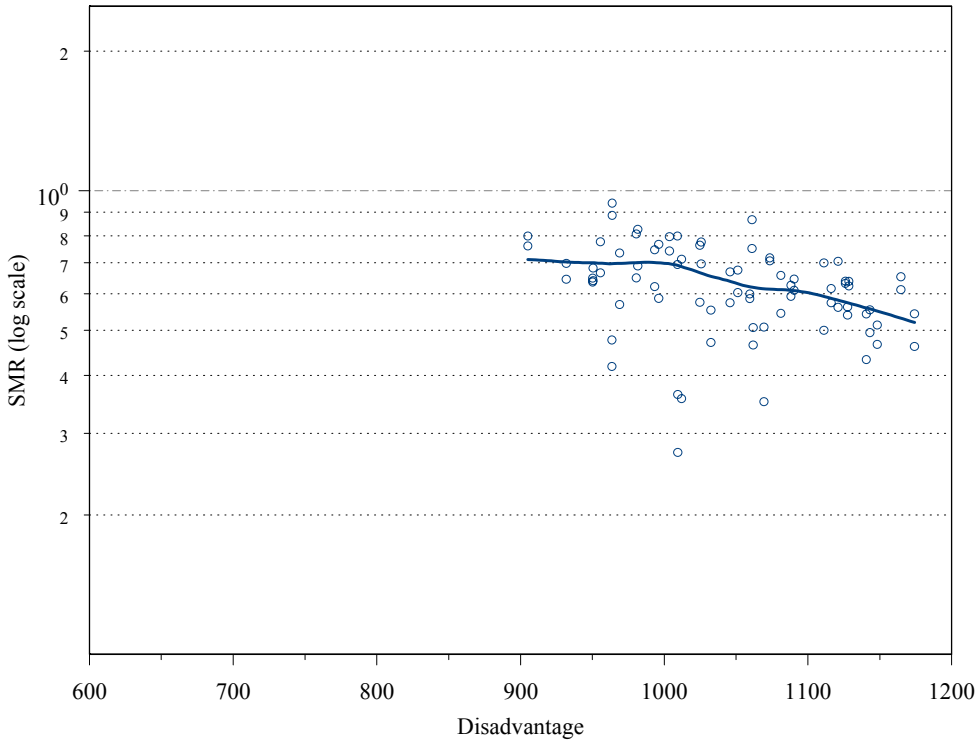
Generalised Linear Models of Indirect Rates

Indirectly age standardised ratios were calculated for the two winters to check and see if the results are the same as those for the direct method. The standard age specific rates were calculated internally using the study region deaths for the two winters 1997 and 1998 over the 1996 population of the study region. Then these were applied to the POA and SLA person years lived (calculated by dividing the population in 1996 in half). Below the SMR is visualised and the relationships are the same as the directly standardised rates.

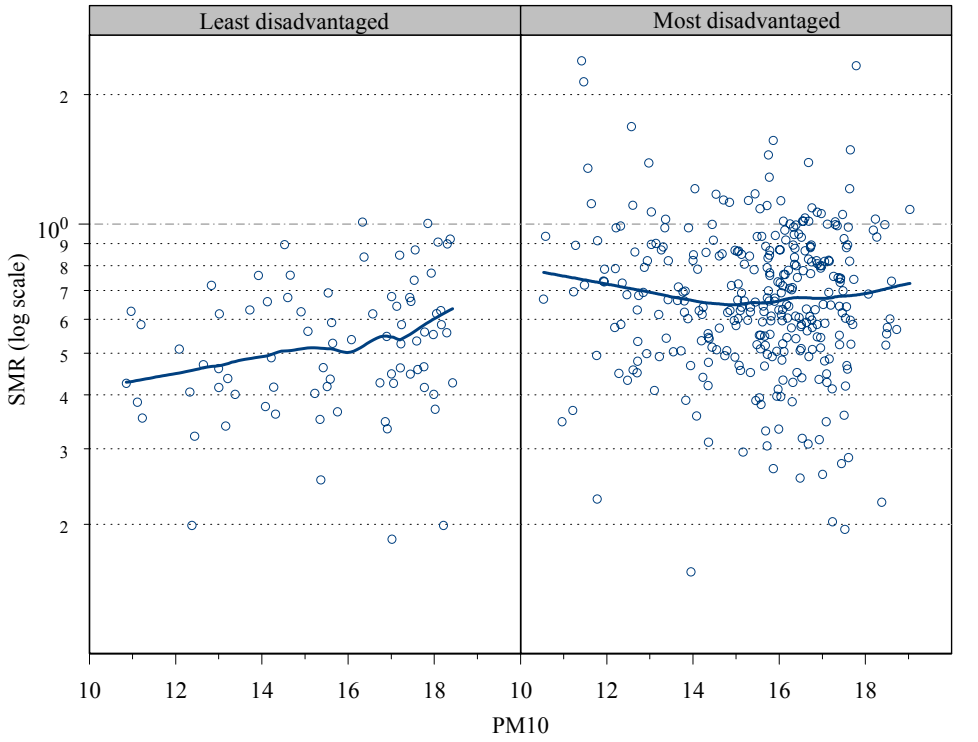
Indirect method



POA winter SMR against disadvantage

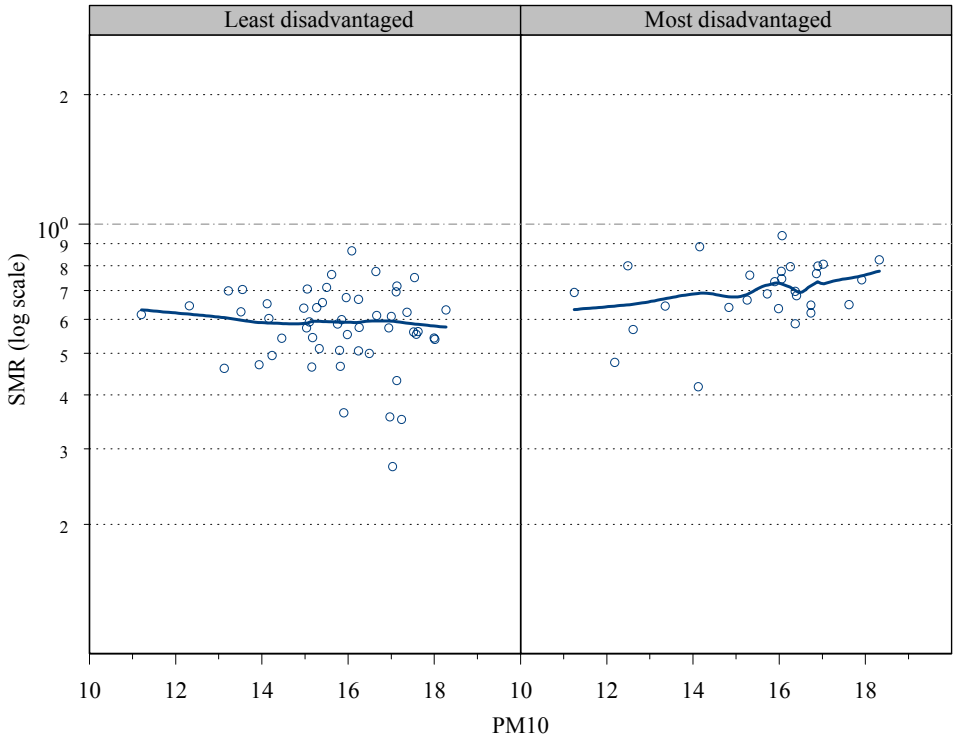


SLA winter SMR against disadvantage



POA winter most and least disadvantaged SMR against PM10

The disadvantage classes are the same as those identified from the unweighted tree of the directly standardised rates. Interesting trend in least disadvantaged areas whereas most disadvantaged is unclear.



SLA winter most and least disadvantaged SMR against PM10

Generalised linear models

Generalised linear models were used to regress the winter observed (OBS) number of deaths with the indirectly calculated expected (EXP) number used as an offset (Clements 2005b) personnel communication).

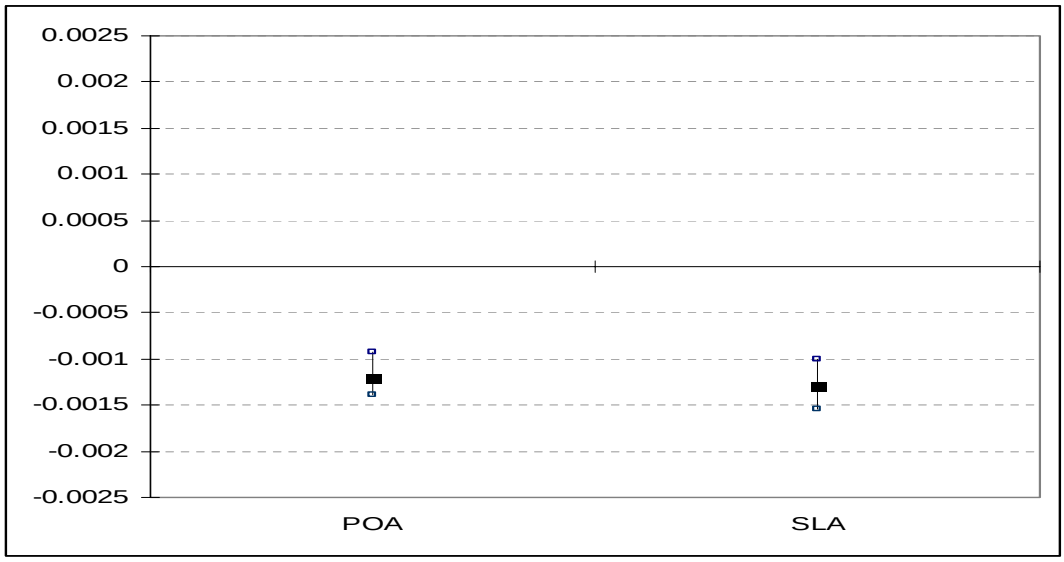
SYNTAX

```
glm(OBS ~ Disadvantage + offset(log(EXP)), family = poisson, data =
"POAorSLA", subset = SEASON == WINTER)

glm(OBS ~ PM10 + offset(log(EXP)), family = poisson, data =
"POAorSLA", subset = SEASON == WINTER & Disadvantage < "threshold")
```

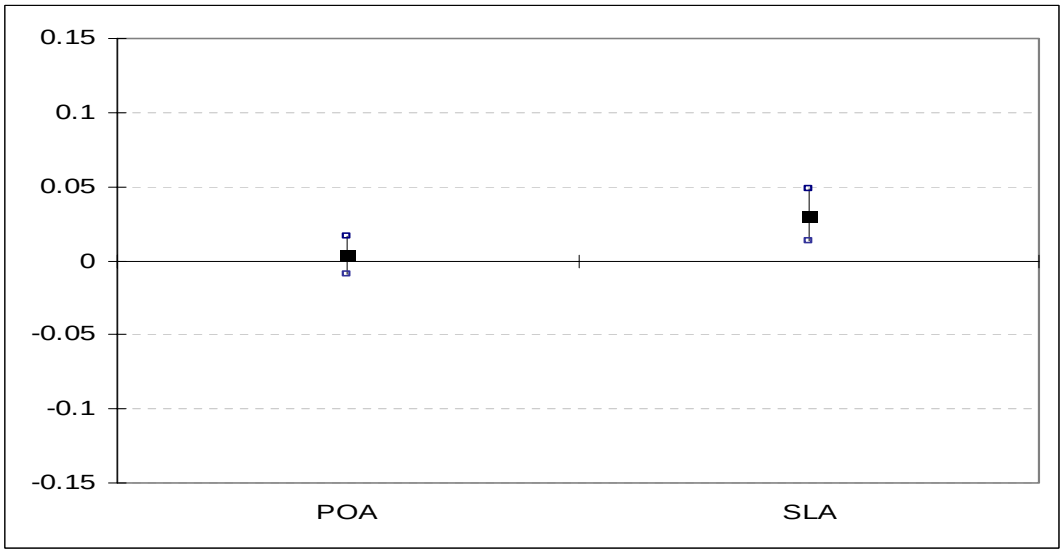
GLM of indirect SMR against disadvantage index in winter

	POA			SLA		
	Value	Std. Error	t value	Value	Std. Error	t value
Intercept	0.775	0.126	6.173	0.889	0.145	6.149
Disadvantage	-0.001	0.000	-9.573	-0.001	0.000	-9.194



GLM of indirect SMR against PM_{10} in disadvantaged areas in winter

	POA			SLA		
	Value	Std. Error	t value	Value	Std. Error	t value
Intercept	-0.451	0.103	-4.365	-0.820	0.142	-5.776
PM_{10}	0.00395	0.007	0.601	0.031	0.009	3.430



Discussion

The indirect method didn't change the direction of the regression coefficients, but the magnitude and standard error did change.

The Disadvantage relationship is significant at both spatial units and the coefficients are similar. The PM_{10} relationship is not significant at POA level while it is at SLA level.

Conclusion

POA numbers are small enough to warrant some caution regarding rate instability. However it was found here that indirect and direct (weighted by variance) give very similar results.

Appendix 3: Air Pollution Exposure Assessment

Data source

Approval from Chris Eiser to use the data was gained on the 29/7/2004. The air pollution observations are made available as edited hourly averages.

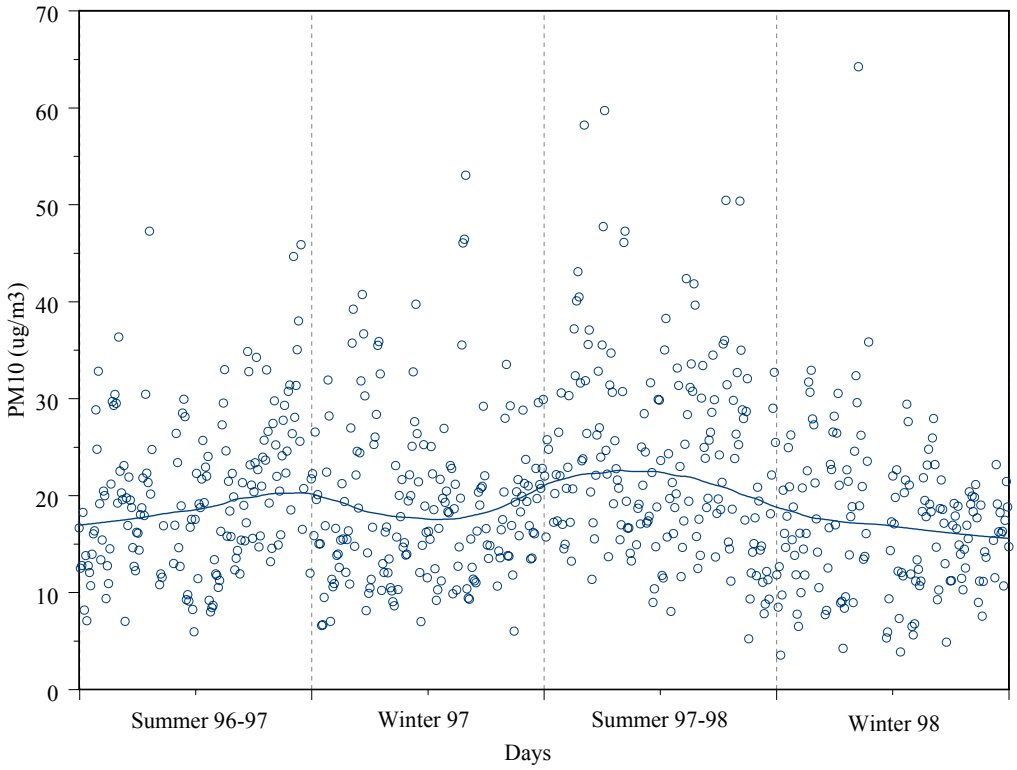
There are some missing data due to the regular 4am cleaning, instrument failure and power supply.

Spatial and temporal patterns through period 1996-1998

Spline interpolation was checked against kriging. Spline does excellent spatial smoothing and allowed adjustment of this by altering the tension of the spline surface. Exploration of spline weights led to decision to use 0.1 for all surfaces which means a smoother surface. Interpolated using longest (most decimal places) and only sites with greater than 80% records. (the State of the Environment report has a 75% limit). The following maps only display whole integer values or small number of decimal points for the monitoring sites.

PM10

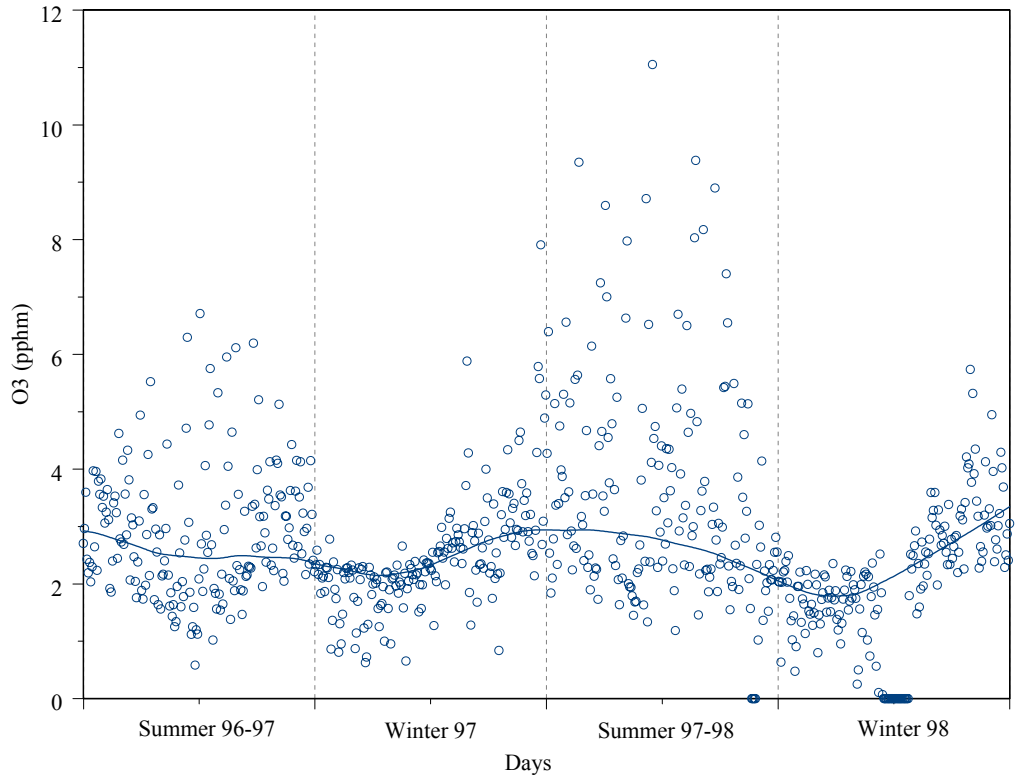
Average of 24 hour avs. Shows a summer peak. Other cities show winter peak – bushfire?



PM10 Daily Twenty-Four Hour Averages for Westmead 96-98

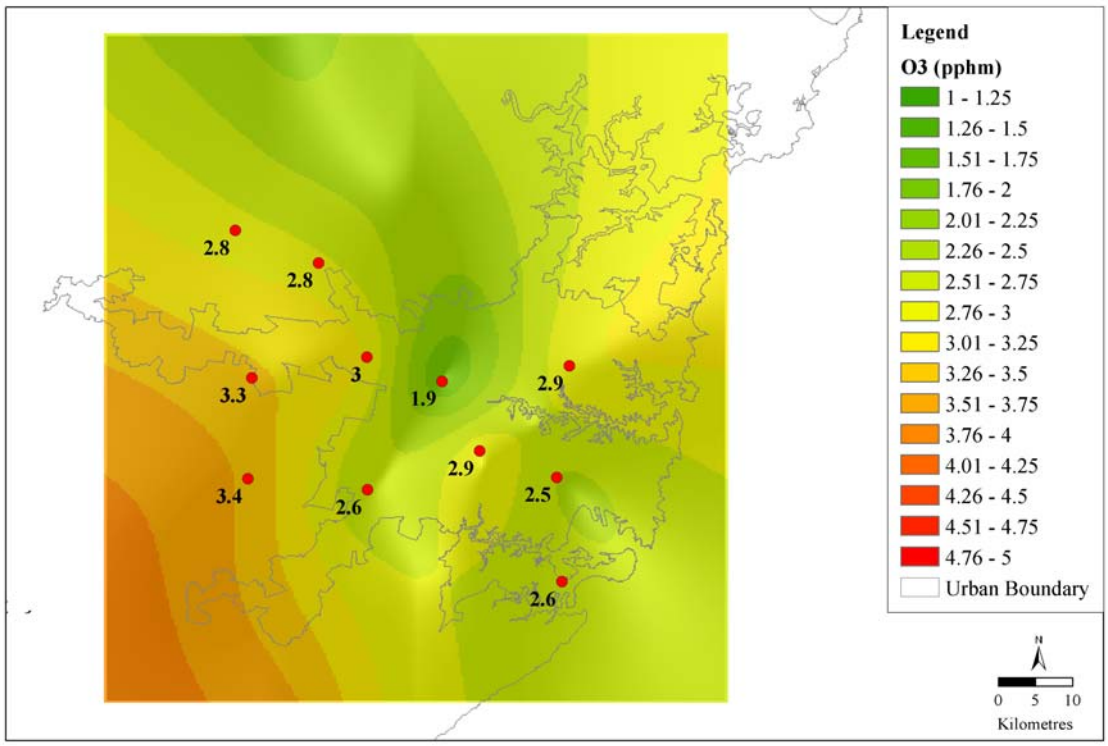
O3

O3 needs rolling four hour max calculated. Shows that this is a summer pollutant.

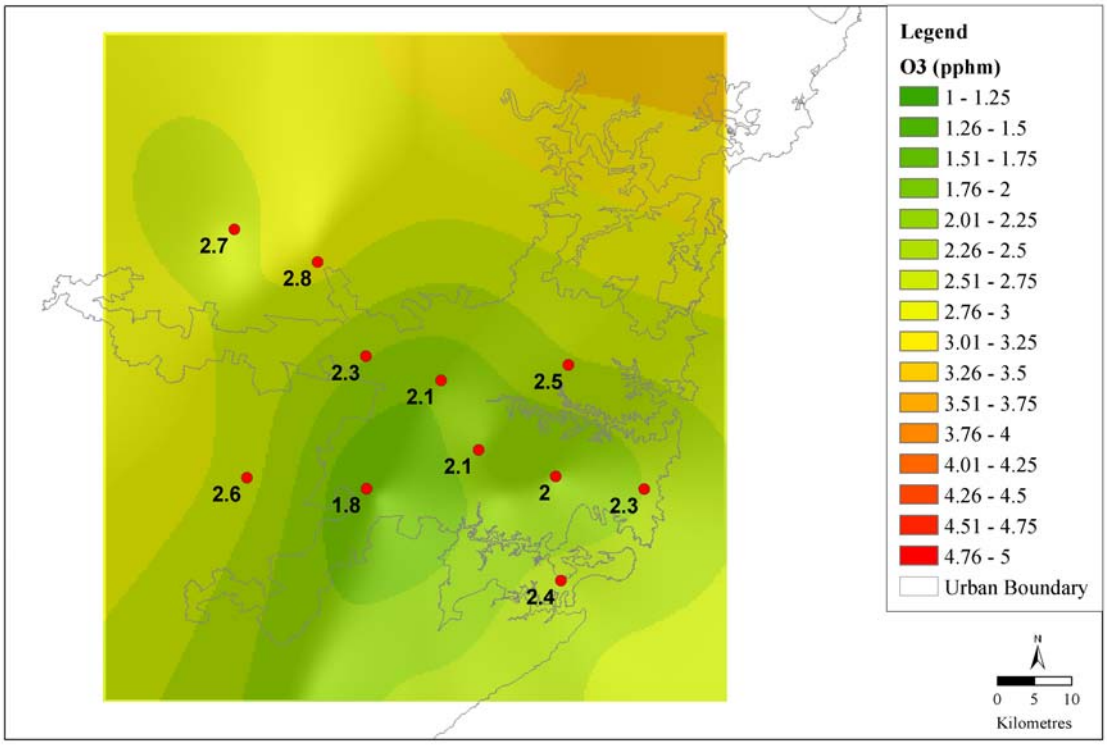


O3 Daily Rolling Four-Hour Maxima at Lindfield 96-98

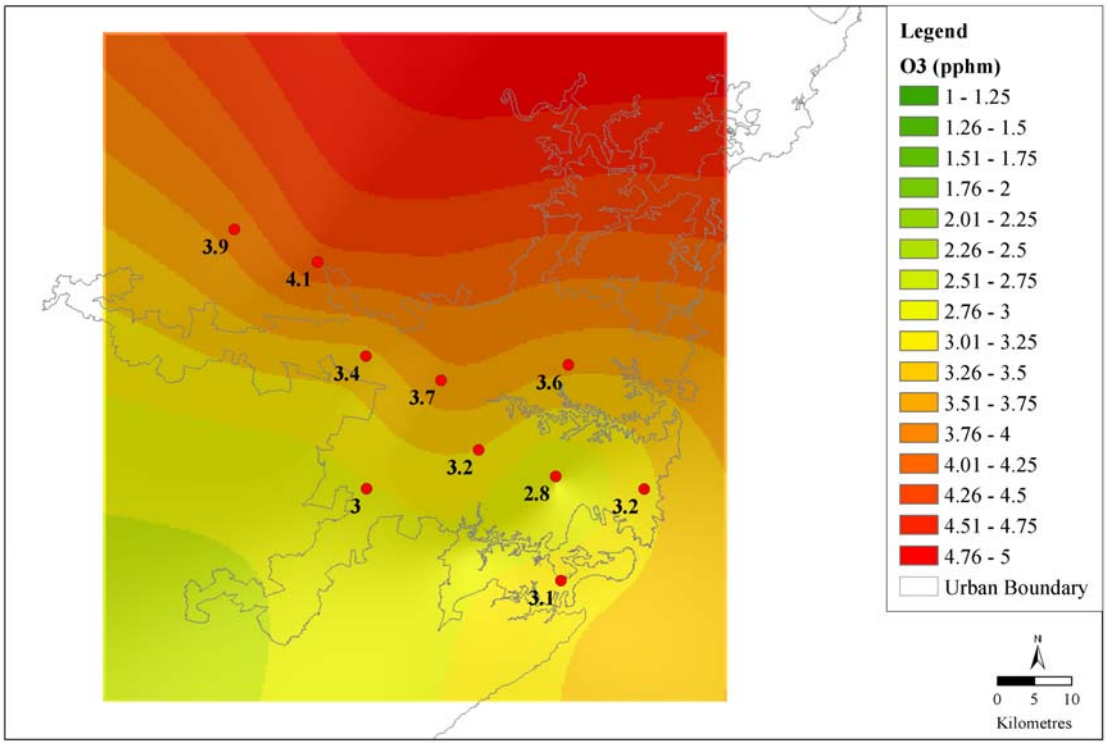
Lindfield is in the inner north. Values are higher at periphery.



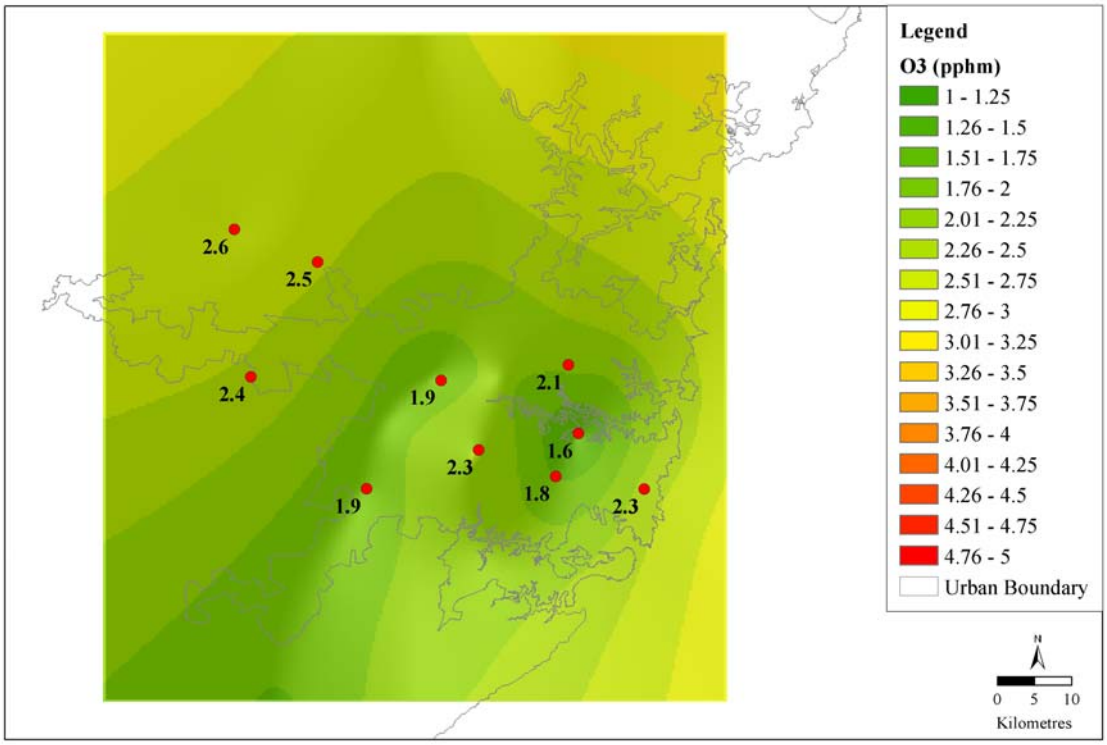
O3 summer 1996-97



Winter 1997 O3 average of max four hour avs



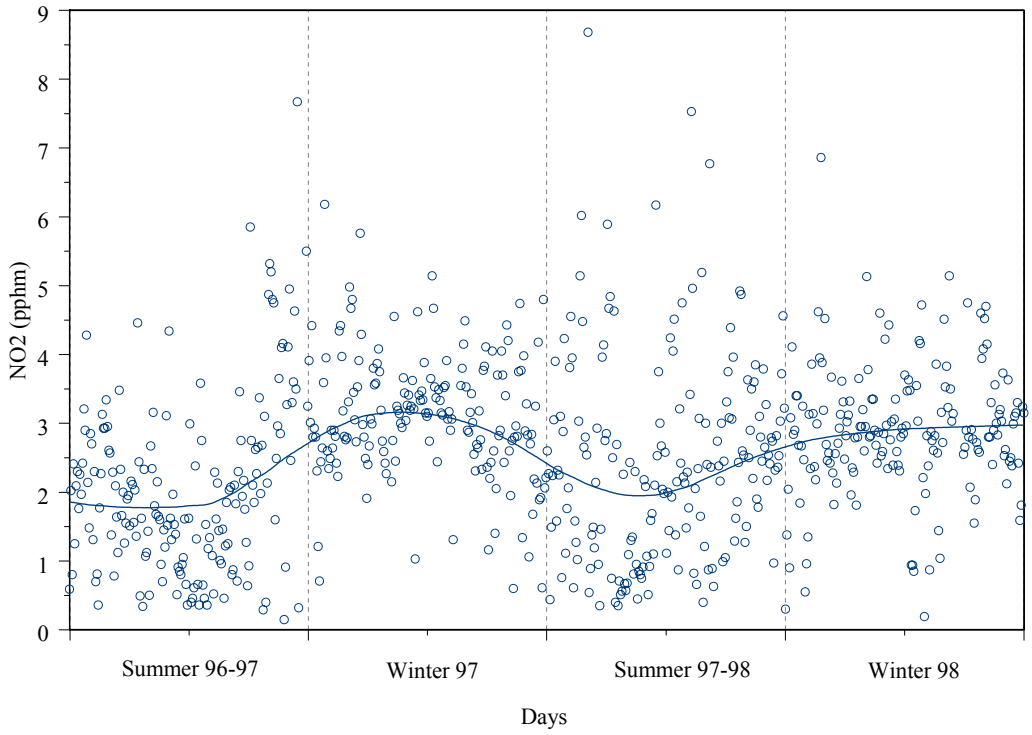
Summer 1997-98



Winter 1998

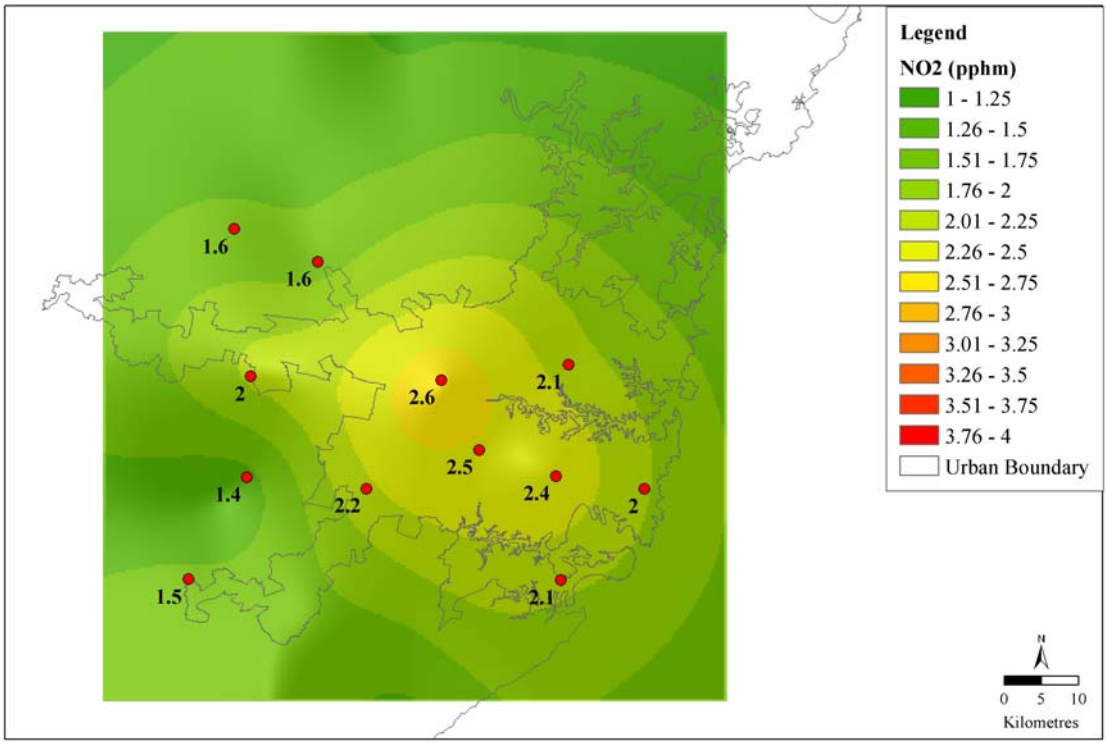
NO2

NO2 daily one-hour maxima through the period. NO2 is generally a winter pollutant

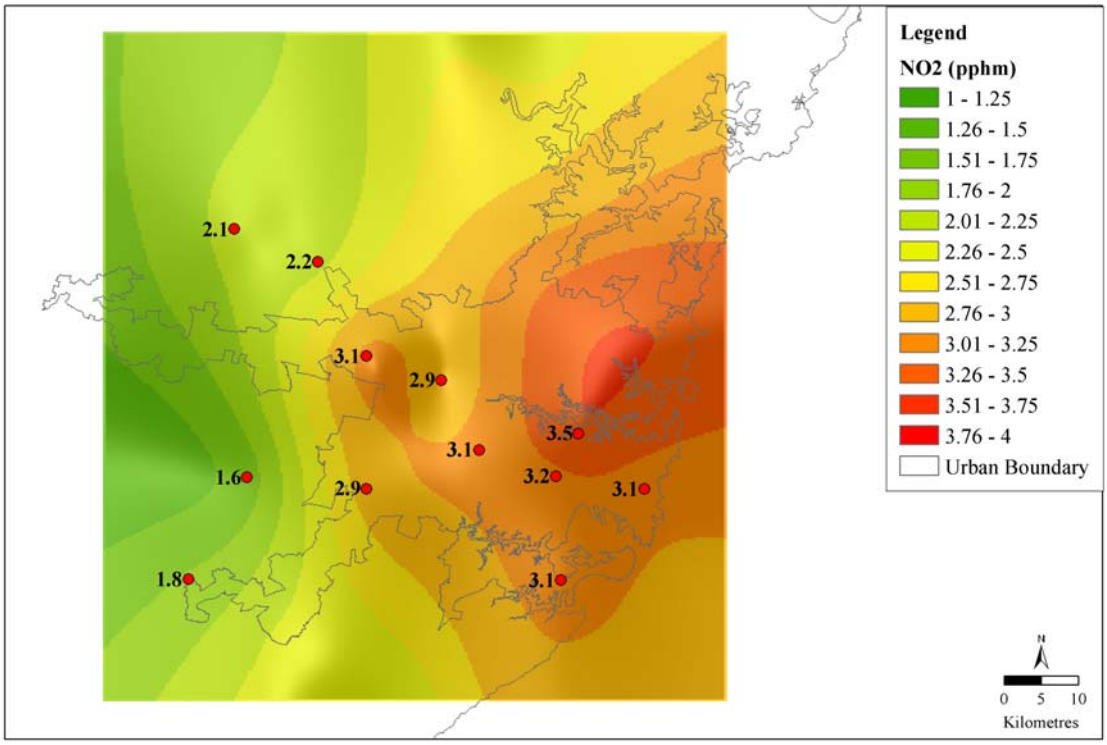


NO2 Daily One-Hour Maxima at Randwick 96-98

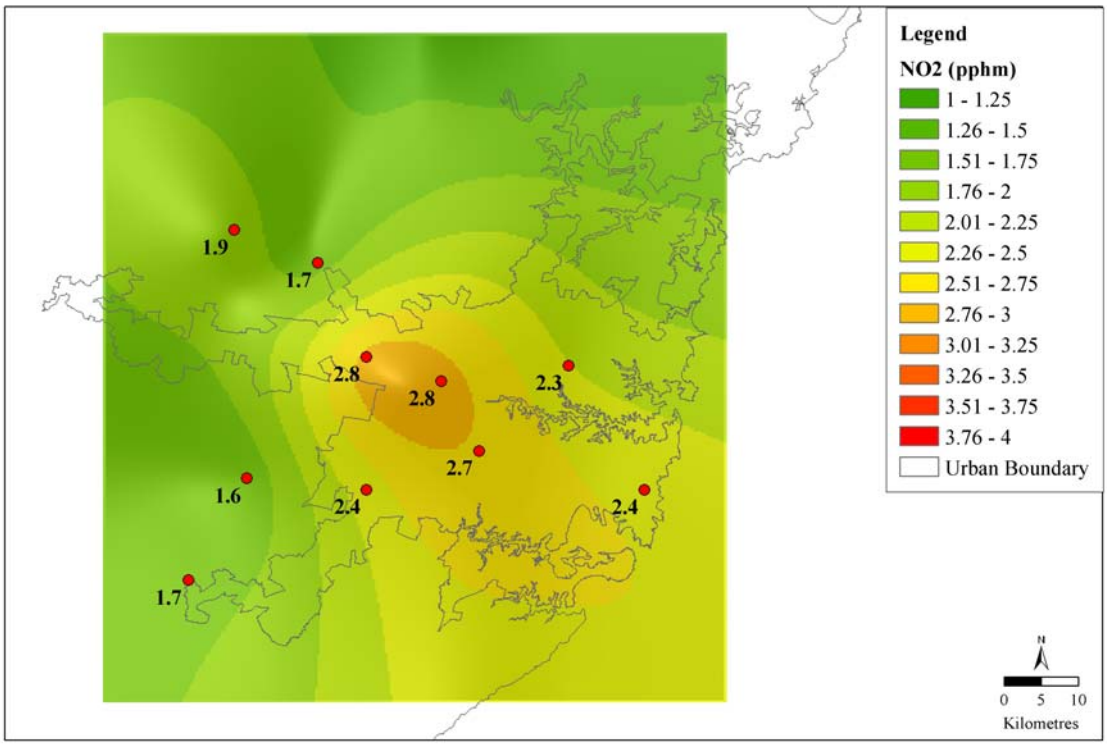
Six month averages of NO2 daily one-hour maxima .through period



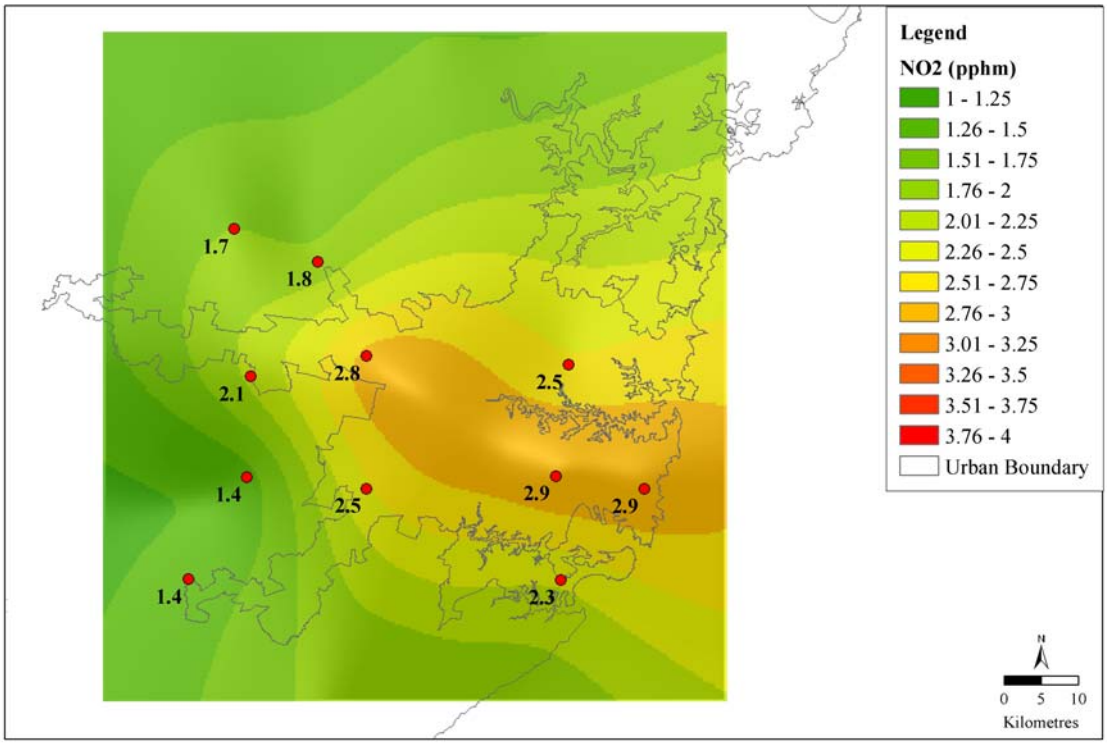
NO2 summer 1996-97



NO2 winter 1997



NO2 summer 1997-98

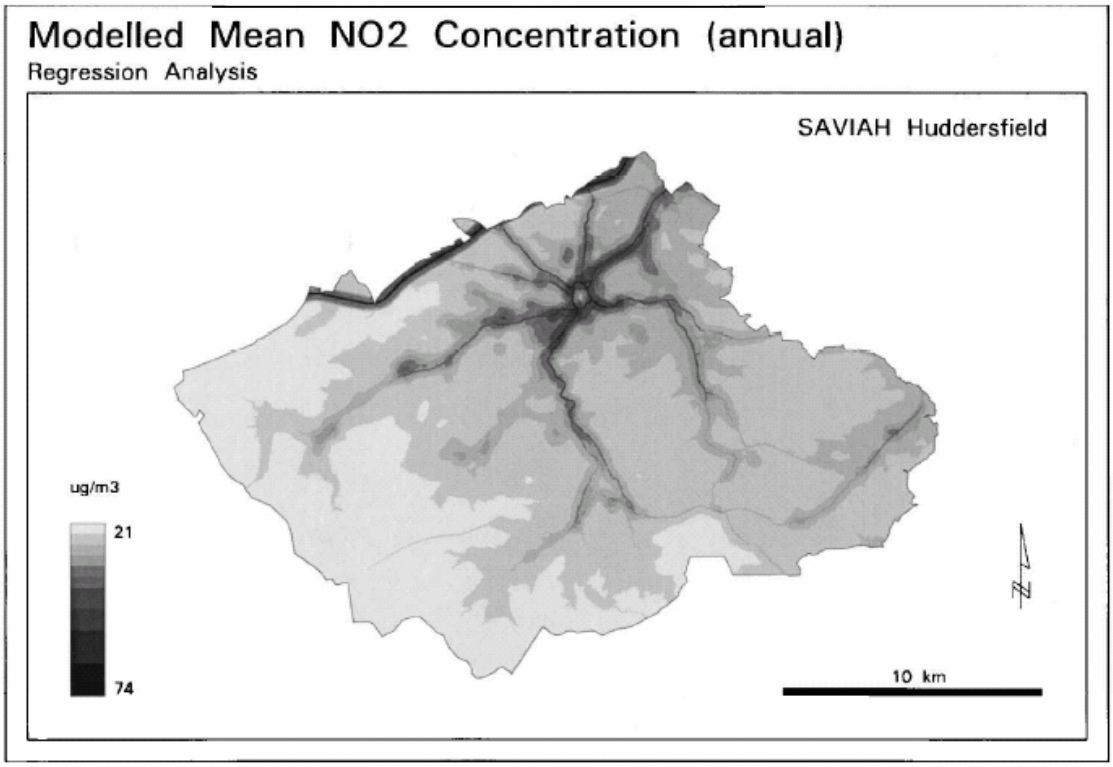


NO2 winter 1998

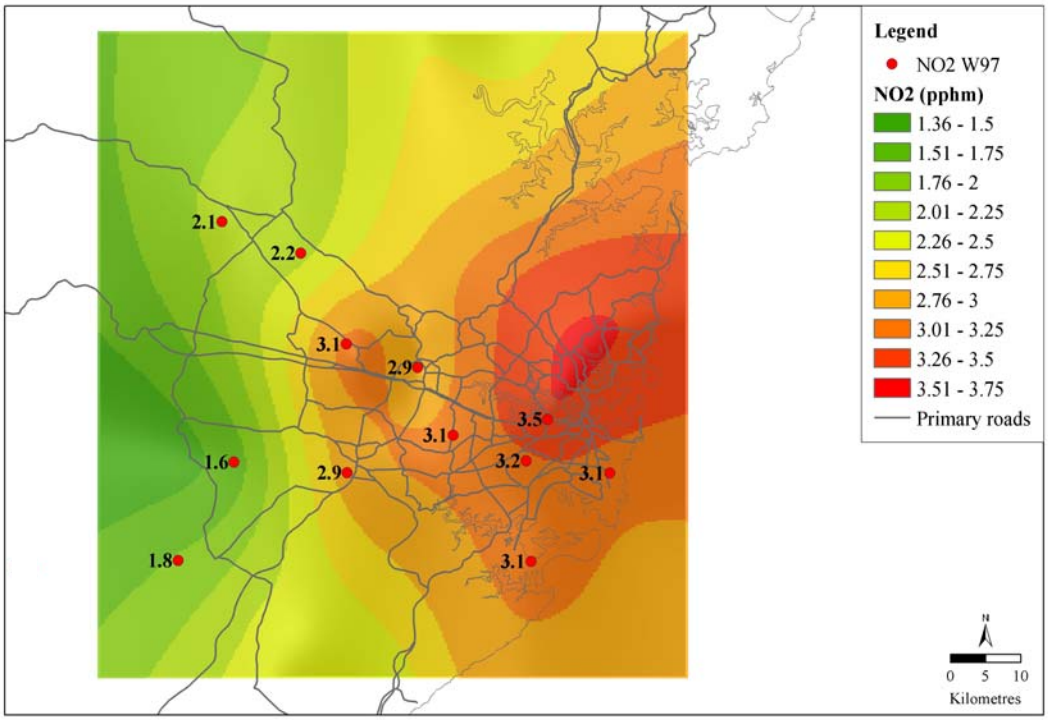
Error and Uncertainty

The AP data is prone to exposure misclassification. One reason for this is that the monitoring sites are widely spaced and surfaces are probably too smooth for these traffic related pollutants. This can be seen in the NO_2 maps presented in the figure below from Briggs et al (1997). The NO_2 surface used for

Sydney is shown in the next figure overlayed by the primary roads and highways. It should be noted that the different maps have different scales and units of measurement of concentration however these do not change the pattern with concentrations much higher along the roads than the Sydney surface indicates.



NO_2 and major roads (from Briggs et al 1997)

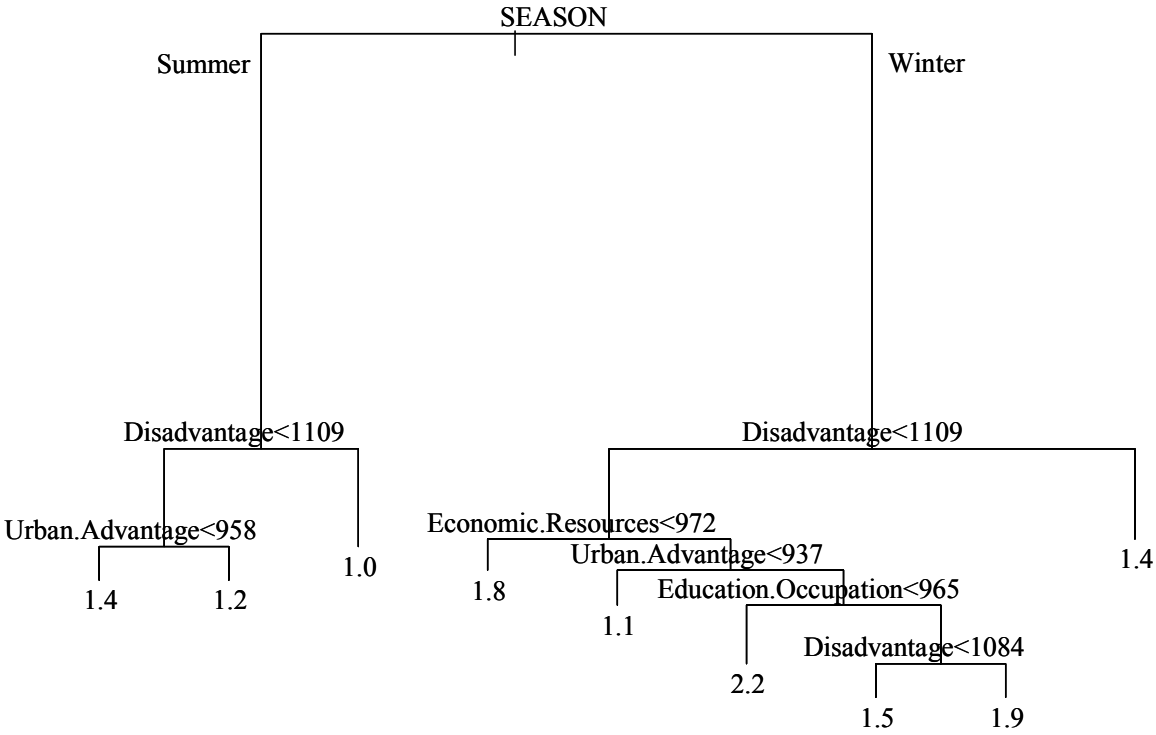


NO_2 Surface for Sydney in Winter 1997

Appendix 4: Weighted Tree Models

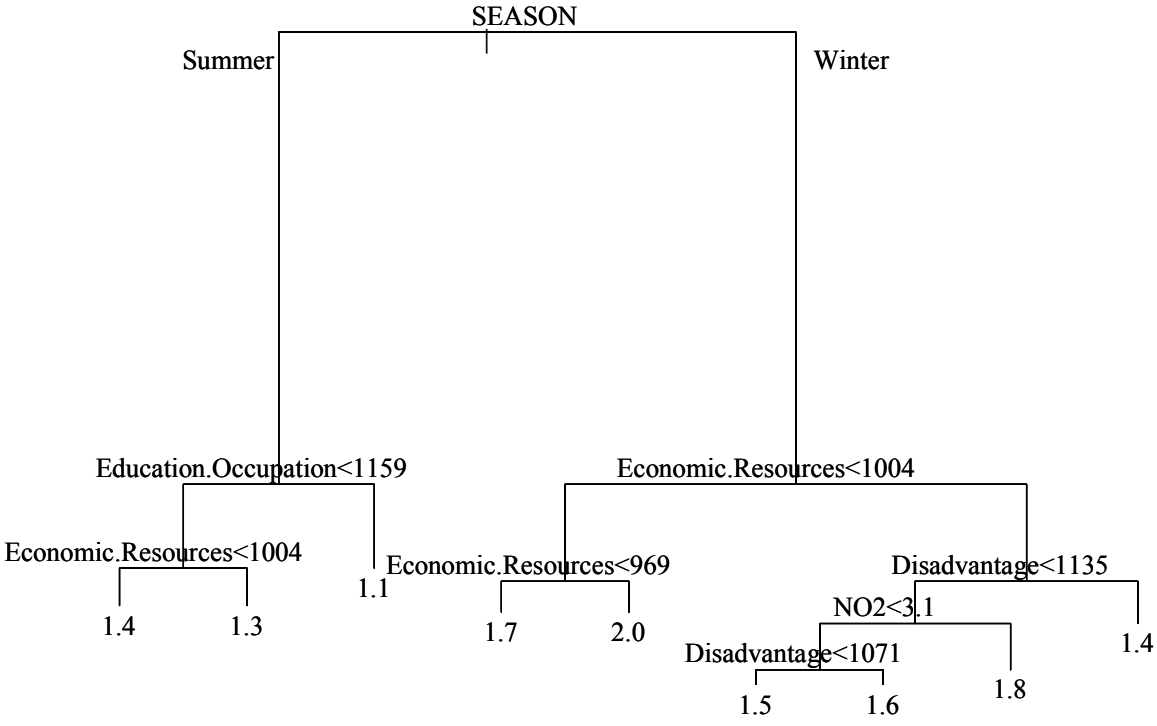
The need for weighting is discussed in previous sections. The regression tree models were run using the inverse variance as weight. These show SES is very important whilst AP is less so.

POA



POA weighted tree

SLA



SLA weighted tree

Appendix 5: Spatial Metadata

Australian Bureau of Statistics Spatial Units

Category	Element	Comment
Data	Title	Historical boundaries of the ASGC 1981-2000
	Custodian	Australian Bureau of Statistics
	Jurisdiction	Australia
Description	Abstract	These areas form a hierarchical system for collecting and publishing demographic data from the Australian Census of Population and Housing.
	Search Word(s)	SLA, CCD, SD, POA, ASGC, LGA
	Geographic Extent Name(s) **OR**	Spatial coverage includes all areas of Australia.
	Geographic Extent Polygon(s)	
Data Currency	Beginning date	1981
	Ending date	2000
Data Status	Progress	
	Maintenance and Update Frequency	Redefined approximately annually and especially each census year
Access	Access Constraints	Research purposes only
	Stored Data Format	Supplied in MapInfo MID/MIF; converted to ESRI shapefile format
	Available Format	
Data Quality	Lineage	CD boundaries are digitised from Census Collection maps at a variety of scales. Higher level units are point reduced and may not overlay exactly on lower level units. Data are as supplied except for some minor changes in codes in some years due to errors. See notes with data. 2001 data collected from the CDATA in the ANU library.
	Positional Accuracy	Varies between levels.
	Attribute Accuracy	
	Logical Consistency	
	Completeness	Complete
Contact Information	Contact Organisation	ABS geography department
	Contact Position	Director Geography, Alec Bamber or Hayley Farthing
	Mail Address	PO Box 10
	Suburb or Place or Locality	Belconnen
	State	ACT
	Postcode	2616
	Telephone	02 6252 7759
	Facsimile	02 6251 8666
	Electronic Mail Address	alec.bamber@abs.gov.au or hayley.farthing@abs.gov.au

Air Pollution Data for Sydney Monitoring Sites

Category	Element	Comment
Data	Title	The Department of Environment and Conservation (DEC NSW) - Atmospheric Science: Air Pollution Data
	Custodian	Department of Environment and Conservation (formerly EPA)
	Jurisdiction	NSW
Description	Abstract	Coordinates for the sites from http://www.epa.nsw.gov.au/air/sites.htm Data are hourly averages for all monitoring sites in Sydney Meteorology Area 94-02 and include: PM10/Pm2.5 - micrograms per cubic metre NEPH - Bsp Ozone(O3) - pphm SO2 - pphm CO - ppm NO - pphm NO2 - pphm NOX - pphm
	Search Word(s)	Air pollution, monitoring sites
	Geographic Extent Name(s) **OR**	Lower Sydney Basin
	Geographic Extent Polygon(s)	
Data Currency	Beginning date	1994
	Ending date	2002
Data Status	Progress	Complete
	Maintenance and Update Frequency	None
Access	Access Constraints	Research purposes by agreement with NSW DEC
	Stored Data Format	ASCII latitude and longitude in decimal degrees
	Available Format	
Data Quality	Lineage	From the DEC/EPA website originally but then fixed after communication with Alan Betts
	Positional Accuracy	~metres
	Attribute Accuracy	High
	Logical Consistency	
	Completeness	Fair, some sites have incomplete records
Contact Information	Contact Organisation	Department of Environment and Conservation (formerly EPA)
	Contact Position	Alan Betts, Chris Eiser or Matt Riley
	Mail Address	
	Suburb or Place or Locality	
	State	
	Postcode	
	Telephone	
	Facsimile	
	Electronic Mail Address	bettsa@epa.nsw.gov.au, Chris.Eiser@environment.nsw.gov.au or Matt.Riley@environment.nsw.gov.au

Sydney Basin Topography: Shuttle Radar Topographic Mission

Category	Element	Comment
Data	Title	3 Arc Second Shuttle Radar Topography Mission Elevation Data,
	Custodian	USGS. Reprocessing by The Global Land Cover Facility,
	Jurisdiction	America
Description	Abstract	The Shuttle Radar Topography Mission (SRTM) obtained elevation data on a near-global scale to generate the most complete high-resolution digital topographic database of Earth. SRTM consisted of a specially modified radar system that flew onboard the Space Shuttle Endeavour during an 11-day mission in February of 2000. SRTM is an international project spearheaded by the National Geospatial-Intelligence Agency (NGA) and the National Aeronautics and Space Administration (NASA). Publisher: Intellectual Property Rights: U.S. Geological Survey; use is free to all if citation is indicated as source. The U.S. Government holds the ultimate ownership. Source for this dataset was the Global Land Cover Facility,
	Search Word(s)	Elevation, topography, DEM
	Geographic Extent Name(s) **OR**	Global
	Geographic Extent Polygon(s)	
Data Currency	Beginning date	
	Ending date	
Data Status	Progress	
	Maintenance and Update Frequency	
Access	Access Constraints	None
	Stored Data Format	GEOTiff
	Available Format	
Data Quality	Lineage	Shuttle Mission, Reprocessed to GeoTIFF. Version 1.0.
	Positional Accuracy	
	Attribute Accuracy	
	Logical Consistency	
	Completeness	
Contact Information	Contact Organisation	Global Land Cover Facility (GLCF) the College Park, University of Maryland. 2004
	Contact Position	
	Mail Address	
	Suburb or Place or Locality	
	State	
	Postcode	
	Telephone	
	Facsimile	
	Electronic Mail Address	http://www.landcover.org. http://glcfapp.umiacs.umd.edu:8080/esdi/index.jsp

The Australian Mortality Database

Category	Element	Comment
Data	Title	The National Mortality Database
	Custodian	ABS
	Jurisdiction	Australia
Description	Abstract	Every death in Australia from 1964-2003. The local area of usual residence at time of death
	Search Word(s)	Death, mortality, geocoded, georeferenced, SLA, Postcode
	Geographic Extent Name(s) **OR**	Australia
	Geographic Extent Polygon(s)	
Data Currency	Beginning date	1964
	Ending date	2003
Data Status	Progress	Ongoing
	Maintenance and Update Frequency	Annual
Access	Access Constraints	Research purposes only and ANU Human Research Ethics Committee approval required
	Stored Data Format	MS Access database
	Available Format	
Data Quality	Lineage	Originally compiled by the State Registrar of Births Deaths and Marriages, de-identified by the ABS
	Positional Accuracy	Available aggregated to Postcode and SLA
	Attribute Accuracy	Unknown
	Logical Consistency	
	Completeness	Complete census of all deaths
Contact Information	Contact Organisation	National Centre for Epidemiology and Population Health, NCEPH OR Australian Bureau of Statistics
	Contact Position	Data Manager
	Mail Address	Melissa Goodwin (NCEPH) OR peter.burke@abs.gov.au (ABS)
	Suburb or Place or Locality	NCEPH, Australian National University, Building 62, Mills Road, Canberra
	State	ACT
	Postcode	0200
	Telephone	(02) 6125 2779
	Facsimile	(02) 6125 0740
	Electronic Mail Address	Melissa.Goodwin@anu.edu.au

Australian Census of Population and Housing

Category	Element	Comment
Data	Title	Australian Census of Population and Housing products: CDATA 1996 and CDATA 2001
	Custodian	The Australian Bureau of Statistics
	Jurisdiction	Australia
Description	Abstract	Enumeration as at 30th June each census year
	Search Word(s)	Population,
	Geographic Extent Name(s) **OR**	Australia
	Geographic Extent Polygon(s)	
Data Currency	Beginning date	Electronically available since 1981
	Ending date	2001
Data Status	Progress	Ongoing, next census in 2006
	Maintenance and Update Frequency	Every 5 years
Access	Access Constraints	Research purposes only
	Stored Data Format	DBF
	Available Format	
Data Quality	Lineage	
	Positional Accuracy	Aggregated to CCD, POA, SLA, SSD, SD and State
	Attribute Accuracy	
	Logical Consistency	
	Completeness	
Contact Information	Contact Organisation	ABS
	Contact Position	Demography
	Mail Address	PO Box 10
	Suburb or Place or Locality	Belconnen
	State	ACT
	Postcode	2616
	Telephone	
	Facsimile	
	Electronic Mail Address	

Socio-Economic Indexes for Areas (SEIFA) 1996

Category	Element	Comment
Data	Title	Socio-Economic Indexes For Areas (SEIFA) 1996
	Custodian	ABS
	Jurisdiction	Australia
Description	Abstract	The 5 Indexes are constructed from census variables collected in each CCD
	Search Word(s)	Socio-Economic Status
	Geographic Extent Name(s) **OR**	Australia
	Geographic Extent Polygon(s)	
Data Currency	Beginning date	1996
	Ending date	1996
Data Status	Progress	Complete
	Maintenance and Update Frequency	None
Access	Access Constraints	Research purposes only.
	Stored Data Format	
	Available Format	
Data Quality	Lineage	From the ABS. Agreement for use by NCEPH staff or students only.
	Positional Accuracy	
	Attribute Accuracy	
	Logical Consistency	
	Completeness	
Contact Information	Contact Organisation	National Centre for Epidemiology and Population Health, NCEPH
	Contact Position	Data Manager
	Mail Address	Melissa Goodwin (NCEPH)
	Suburb or Place or Locality	NCEPH, Australian National University, Building 62, Mills Road, Canberra
	State	ACT
	Postcode	0200
	Telephone	(02) 6125 2779
	Facsimile	(02) 6125 0740
	Electronic Mail Address	Melissa.Goodwin@anu.edu.au