UNIVERSITY of TASMANIA

University of Tasmania Open Access Repository

Cover sheet

Title
Where the dust settles: a spatial investigation of respiratory disease and particulate air pollution in the Tamar Valley (1992-2006)

Author
Jabbour, S

Bibliographic citation
Jabbour, S (2007). Where the dust settles: a spatial investigation of respiratory disease and particulate air pollution in the Tamar Valley (1992-2006). University Of Tasmania. Thesis.
https://doi.org/10.25959/23211233.v1

Is published in:

Copyright information
This version of work is made accessible in the repository with the permission of the copyright holder/s under the following,

Licence.

If you believe that this work infringes copyright, please email details to: oa.repository@utas.edu.au

Downloaded from University of Tasmania Open Access Repository

Please do not remove this coversheet as it contains citation and copyright information.

University of Tasmania Open Access Repository

Library and Cultural Collections
University of Tasmania
Private Bag 3
Hobart, TAS 7005 Australia
E oa.repository@utas.edu.au CRICOS Provider Code 00586B | ABN 30 764 374 782 utas.edu.au

Where the dust settles: a spatial investigation of respiratory disease and particulate air pollution in the Tamar Valley (1992-2006)

by

Samya Jabbour (BSc.)

A thesis submitted in partial fulfilment of the requirements for a Graduate Diploma in Spatial Information Science with Honours at the School of Geography and Environmental Studies, University of Tasmania (October, 2007).

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any tertiary institution, and to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signed

Samya Jabbour BSc.
19 October 2007

Table of contents

DECLARATION ..... **II**

ACKNOWLEDGMENTS ..... **X**

ABSTRACT ..... **XI**

1 INTRODUCTION **1**

1.1 Background 1
1.2 Significance of the study 2
1.3 Aims and objectives 2
1.4 Structure of the thesis 3

2 SPATIAL EPIDEMIOLOGY OF RESPIRATORY DISEASE: REVIEW **4**

2.1 Particulate air pollution and health 4
2.1.1 PM_{10} and $PM_{2.5}$ defined 4
2.1.2 PM_{10} and $PM_{2.5}$ health impacts 5
2.1.3 National and international standards 5
2.1.4 Tamar Valley air pollution 6
2.1.5 Tamar Valley respiratory health studies 8
2.2 Spatial epidemiology 9
2.2.1 Why spatial? 9
2.2.2 Confounders in spatial epidemiology 10
2.2.3 GIS and 'Environmental Justice': a case for the spatial approach 11
2.3 Air pollution exposure estimates 11
2.3.1 Exposure surfaces – the spatial approach to exposure classification 12
2.3.2 Exposure misclassification 12

3 DATA ACQUISITION AND PRE-PROCESSING **14**

3.1 Introduction 14
3.1.1 Point data vs. aggregate data 14
3.1.2 Confidentiality, geoprivacy and ethical reporting 14
3.2 Data acquisition and pre-processing 15
3.2.1 Project ethics approval 15
3.2.2 Spatial extent of the study area 16
3.2.3 Address points dataset 16
3.2.4 Census data 18
3.2.5 Population estimation 'per residence' 18
3.2.6 Medical records 19
3.3 Random perturbation of address points 20
3.3.1 De-identification of medical data 21

3.3.2 A better method: de-identification of *address* data 23
3.3.3 Ethical reporting of spatial public health data 27
3.3.4 Further masking of rural addresses 27

4 DISEASE MAPPING 29

4.1 Research Hypothesis 29
4.2 Introduction 29
4.2.1 Generalisation of point patterns into 'hot-spots' or 'clusters' 29
4.2.2 The value of disease mapping 29
4.2.3 Cluster detection techniques 30
4.3 Methods 31
4.3.1 Software requirements 31
4.3.2 Preliminary disease analysis 31
4.3.3 Kernel Density Estimation 32
4.3.4 Getis Ord Gi* statistic 33
4.3.5 Kulldorff's spatial scan statistic 33
4.4 Results 35
4.4.1 Preliminary disease analysis 35
4.4.2 Kernel Density/Intensity Estimation 36
4.4.3 Getis Ord Gi* Statistic 41
4.4.4 Kulldorff's spatial scan statistic 43
4.4.5 Summary statistics 45
4.5 Discussion 47
4.5.1 Disease clusters and statistical inference 47
4.5.2 Hospital admissions as a measure of morbidity 49
4.5.3 Spatial autocorrelation and confounders 50
4.6 Conclusion 51

5 EXPOSURE SURFACES AND DISEASE: CORRELATIONS 52

5.1 Research Hypothesis 52
5.2 Introduction 52
5.2.1 Winter air pollution in the Tamar Valley 52
5.2.2 Exposure assessment and the ecological fallacy 53
5.2.3 Air pollution dispersion modelling 55
5.2.4 Terrain analysis as proxy for cold air drainage 56
5.3 Software and data requirements 57
5.4 Methods 57
5.4.1 Non-spatial investigation of PM_{10} and disease 57
5.4.2 Spatial extent of study area 58
5.4.3 Digital elevation model 58

5.4.4 TAPM – The Air Pollution Model 59
5.4.5 LandSerf terrain analysis 60
5.4.6 Zonal statistics and data generalisation 62
5.4.7 Global statistical analysis 65
5.4.8 Geographically Weighted Regression 65
5.5 Results 66
5.5.1 Non-spatial analysis of PM_{10} and disease relationships 66
5.5.2 TAPM and LandSerf exposure surfaces 70
5.5.3 Population and disease density surfaces 72
5.5.4 Global statistical analysis 75
5.5.5 Geographically Weighted Regression 75
5.6 Discussion 78
5.6.1 Terrain features and exposure assessment 78
5.6.2 Dispersion modelling and exposure assessment 79
5.6.3 Issues of socioeconomic confounding in the Tamar Valley 81
5.6.4 Methodological limitations of this study 82
5.7 Conclusion 83

6 CONCLUSION 84

6.1 Findings from this study 84
6.2 Recommendations for further research 87

REFERENCES 88

List of figures and tables

Figure 2.1- *PM10 concentrations on the sixth-highest recorded day of each year 1991 to 2001 for Launceston and every Australian capital city. The red line indicates the NEPM 24 hour standard of 50 µgm^{-3}, which may be exceeded just five times annually under the national standard. (Source: Australian Government Department of the Environment and Water Resources, 20 Jan 2006)* _____ *6*

Figure 2.2 *– Landsat and Quickbird satellite images of all of Tasmania (left) and just the Tamar Valley (right). In the image on the left, a large band of fog is seen covering the entire Tamar Valley and extending southeast to the midlands area. The image on the right shows the area around Launceston blanketed by low cloud/fog. (Sources: SOER / Google Earth (2007))* _____ *7*

Figure 3.1 *– The white rectangle demarcates the spatial extent of the Tamar Valley study area in northern Tasmania* _____ *16*

Figure 3.2 *- Map of a section of the Tasmanian Address Points Dataset with points coloured by postcode; a digital elevation model is shown in shaded relief to display terrain features. Postcodes in the Tamar Valley cover vast areas of varying terrain, and are therefore not a suitable aggregation unit for micro-scale health studies such as this.* _____ *17*

Figure 3.3 *- The Address Points Dataset and Census collection districts, coloured according to population numbers. The white rectangle defines the spatial extent of the study area.* _____ *18*

Figure 3.4 *- Census collection districts and Address Points Dataset were combined to give an approximation of population 'per house', then clipped to the spatial extent of the study area. Digital elevation model is displayed in shaded relief.* _____ *19*

Figure 3.5 *- Illustration showing that address points could be moved further than 200 m from their 'true' location (up to 282.8 m) if maximum perturbation distances were added in both the Easting and Northing direction. Basic Pythagorean theory was applied to remedy this problem.* _____ *22*

Figure 3.6 *- Unsatisfactory results of the first de-identification attempt. Dark blue points are house locations from the Address Points Dataset, green points indicate COPD incidence. For dwellings with numerous admissions, it is clearly seen that 'cases' accumulate and become clustered in a 200 m circular buffer around the true address location. This renders address data potentially re-identifiable. It can also be seen that patterns in the organisation of urban properties make it possible to re-identify the original street address.* _ *23*

Figure 3.7 *– Successful de-identification of address points using the second macro. This diagram was developed using sample address data to check to accuracy of the macro prior to implementation.* _____ *25*

Figure 3.8 *- The Address Points dataset at two scales in its original form* _____ *26*

Figure 3.9 *- The APD at two scales after random perturbation (Source: Land Information Services Tasmania (theLIST))* _____ *26*

Figure 4.1 *– Kernel estimation, showing the effect of different bandwidths, or search windows, on both the 'smoothness' of the created density surface and the magnitude of reported density estimates (f(x)). Bandwidth increases from left to right in the diagram. (Adapted from (Silverman, 1986))* _____ *32*

Figure 4.2 *- Proportional hospital admission rates for Asthma, Bronchiolitis, Bronchitis and COPD for each year of the study period.* _____ *35*

Figure 4.3 *- Total Asthma (blue) and COPD (pink) admissions plotted individually and combined (yellow) for every year in the study period. The rise in COPD in 1999 is not entirely explained by the change from ICD-9 to ICD-10 in this year, as the increase in COPD in this year is greater than the decrease in Asthma.* 36

Figure 4.4 *– Total admissions for Asthma (blue), Bronchiolitis (pink) and Bronchitis (yellow) for every year in the study period* 36

Figure 4.5 *– Kernel Intensity Estimation of Total Population (A) and Total Asthma (B), Bronchiolitis (C), Bronchitis (D) and COPD (E) for the 1992-2006 study period. It is seen that disease patterns (B – E) are different from the spread of total population (A), indicating that disease is not occurring uniformly across the population. All legends are relative only. A 2 km search radius was used.* 38

Figure 4.6 *– Kernel Intensity Estimation showing annual variation in the spatial pattern of ABB (Asthma, Bronchiolitis and Bronchitis) for every year of the study period.* 40

Figure 4.7 *– KDE showing seasonal variation in disease occurrence for combined Asthma, Bronchiolitis and Bronchitis for each combined season of the study period 1992-2006. Note the different legend for each season.* 40

Figure 4.8 *- Results of the Getis Ord Gi* statistic (with z-score rendering) for (A) Asthma, (B) Bronchiolitis, (C) Bronchitis and (D) COPD. A 1 km radial search window was used. Only clusters that are > 1 (beige) and > 2 (red) standard deviations above the mean expected rate of disease for the study area are displayed. To protect geoprivacy, isolated points were removed prior to publication.* 41

Figure 4.9 *– Z-score results of the Getis Ord Gi* statistic for combined cases of Asthma, Bronchiolitis and Bronchitis for the entire study period (1992-2006). Left map shows only those clusters with a z-score greater than 1. The map on the right shows the Launceston area (the area within the white square in the left image), showing clusters with all z scores ranging between 2 standard deviations above and below the population mean. The localities of Rocherlea, Newnham, Ravenswood, Invermay, Waverley, St Leonards and Glen Dhu show 'significantly' higher rates of disease than the general population. ABB incidence around the Launceston CBD is at least 2 standard deviations below the mean. Isolated address points have been removed.* 42

Figure 4.10 – *SaTScan results for Asthma clusters in entire valley. Large red elliptical cluster centred around Mt Direction suggests that Asthma incidence is 'significantly' higher on the east side of the Tamar than the west (this cluster returned a p-value of <0.01). Smaller yellow cluster (p <0.01) is shown just north of Launceston CBD at Invermay. All other clusters vary in size and significance. Address points of the background population are displayed in green.* 43

Figure 4.11 *– SaTScan analysis for Asthma clusters in the George Town area (the white square shows analysis extent). Both the red and yellow clusters are significant at p<0.001. Green points are the scrambled address points of the APD; digital elevation model is displayed in shaded relief.* 44

Figure 4.12 *– SaTScan analysis for Asthma clusters in the entire study area, showing only the George Town region (the white square here is extent of display only). The red cluster is the northern tip of the large elliptical cluster in Figure 4.10; green points are scrambled address points of the APD. Digital elevation model is displayed in shaded relief.* 45

Figure 4.13 *– Geographic location of digitised population subregions 'George Town' (green), 'Hadspen' (pink), 'Invermay'(orange), 'Newnham' (light blue) and 'Waverley' (yellow).* 46

Figure 5.1 – *The spatial extent of this section of analysis is shown as the pink rectangle covering Launceston, Newnham and Hadspen.* 58

Figure 5.2 – *Demonstration of the quadratic approximation function used in LandSerf to categorise digital elevation models into the six terrain classes listed on the right (peak, pass, pit, plane, channel, ridge). (Sources: (Wood, 1996); (Fisher et al., 2004))* 60

Figure 5.3 – *Illustration of multi-scale fuzzy feature classification, showing that the same point on a landscape can be assigned different morphometric classes at different spatial scales. (Source: (Fisher et al., 2004))* 61

Figure 5.4 *– Enlarged image of a non-overlapping 200 m buffer, or zone, illustrating the variation in underlying grid cells. Zonal statistics provide summary statistics on these grid cell values and assign these to the centroid of the zone, which in this case is the address point in its scrambled location.* 62

Figure 5.5 *– Histograms showing the spread of raster values within four sample buffers of the DEM (left), channel (middle) and planar (right) grid layers. While distributions vary markedly between buffers, none of the histograms are obviously skewed, and so can be considered normally distributed.* 63

Figure 5.6 *– Relationships between 95th percentile PM_{10} concentrations (blue) and annual disease admissions (pink) for (A) Asthma, (B) Bronchiolitis, (C) Bronchitis and (D) COPD for each year in the study period.* 67

Figure 5.7 *- Combined Asthma, Bronchiolitis, Bronchitis and COPD (ABBC) plotted against Ti Tree Bend PM_{10} 95th percentile values for each year in the study period, showing reasonable correlation until 1999, then a strong departure from this trend.* 67

Figure 5.8 *- Asthma, Bronchiolitis and Bronchitis (ABB) (i.e. minus COPD) plotted against Ti Tree Bend PM_{10} 95th percentile values for each year in the study period, showing a much more consistent correlation with the removal of COPD.* 68

Figure 5.9 *- Paired graphs of daily fluctuations in (upper, pink) combined hospital admissions for Asthma, Bronchiolitis and Bronchitis (ABB), and (lower, blue) 24hr average concentrations of PM_{10} recorded at Ti Tree Bend, for the latter years of the study period (2003-2006).* 69

Figure 5.10 *– TAPM exposure surfaces of (A) 'cmax' and (B) cavg, showing the modelled PM_{10} maximum (top) and average (bottom) concentration at each grid cell.* 70

Figure 5.11 *– All terrain exposure surfaces used in the study displayed both with (right) and without (left) address points and 200 m buffers. The outline of the Tamar River is shown in blue.* 71

Figure 5.12 *– TAPM "cmax" resampled to 200 m resolution and displayed with adjusted histogram to show areas of highest PM_{10} concentration. The areas of East Launceston and Newstead show the highest modelled particulate levels.* 72

Figure 5.13 *– Density distributions of (A) total population, (B) disease admissions for all winters, and (C) disease admissions for winter 2005. All are shown with 50% isopleths demarking the area within which 50% of the density distribution falls. The Tamar River outline is shown in blue. Grids are in 200 m resolution.* 73

Figure 5.14 *- TAPM cmax grid for 2005 winter maximum PM_{10} concentrations overlaid with 50% isopleths of 2005 winter ABB admissions (green), and total population (purple). It is seen that the TAPM output closely follows the total population distribution (upon which emission levels were based). The areas of highest PM_{10} concentration around East Launceston and Newstead (see Figure 5.12) show high disease rates, while a large proportion of disease cases also occur in the less populated areas of Ravenswood, Waverley and Hadspen. The outline of the Tamar River is shown in blue. Isopleths were derived using Hawth's kernel density estimator in Spatial Ecology extensions for ArcGIS (Beyer, 2004).* 74

Figure 5.15*- Three-dimensional view of the Launceston region seen in Figure 5.14, looking northwest through the North Esk and Tamar River valleys from Relbia. Terrain height has been exaggerated slightly for*

illustrative purposes; Tamar River outline is shown in blue. It is seen that a large proportion of houses with recorded winter admissions for ABB (light green) occurred in areas outside of the major population centre (purple). These were predominantly on the eastern side on the North Esk valley (on the right in this image). 74

***Figure 5.16** – Results of Geographically Weighted Regression analysis for parameter values. Density of disease incidence for all winters was modelled against each exposure surface, and total population. Spatial non-stationarity exists in all relationships, though most notably with the terrain analysis layers of channel, planar and elevation. Light shaded areas correspond to locations where the independent variable has a higher influence on disease incidence. 50% isopleth of all winter disease density in shown in yellow; the Tamar River outline is shown in blue.* 77

***Table 3-1** - General format of the Address Points Dataset (APD).* 21

***Table 3-2** - General format of the medical records datasets prior to de-identification. Admissions for each disease (Asthma, Bronchiolitis, Bronchitis and COPD) were summarised in separate tables. (All displayed data are hypothetical.) *LOS = length of stay associated with hospital admission.* 21

***Table 3-3** - General format of the Address Points Dataset (APD) after random perturbation with the second macro. Street addresses have been removed and true Easting and Northing values have been replaced by the results of the random perturbation. Incremental counts of each disease have been generated (AS = Asthma, BL = Bronchiolitis, BR = Bronchitis, COPD = Chronic Obstructive Pulmonary Disease).* 24

***Table 3-4** - General format of the medical records datasets after de-identification with the second macro. Street addresses have been replaced with the adjusted Easting and Northing values from the APD. (All data are hypothetical.) *LOS = length of stay associated with hospital admission.* 25

***Table 4-1** – Global statistics of disease admissions for the entire study area. All disease numbers reflect 15 years of accumulated hospital admissions. (*ABB = combined Asthma, Bronchiolitis and Bronchitis.)* 36

***Table 4-2** – Analysis of disease incidence in local subregions of the study area, showing up to 65% increase in disease incidence in these areas relative to the total population of the study area. Geographic locations used to define subregions are shown in Figure 4.13. (* Calculated as disease cases / population) (** Calculated as local disease incidence of subregion / disease incidence of total study area x 100)* 46

***Table 5-1**- An section of the table used to run Geographically Weighted Regression, containing centre coordinates of each grid cell and all corresponding raster values. 'Case05' corresponds to the density surface of disease cases for winter 2005; 'wscase' is disease cases for all winters (1992-2006).* 64

***Table 5-2** – Global statistical relationships between binary disease data (0 = non-cases, 1 = disease cases) for 2005 winter disease data and TAPM outputs, and for all winters disease data and all terrain-based exposure surfaces. Population per house was also measured against all winter disease data.* 75

***Table 5-3** – Test for non-stationarity in the relationship between disease incidence for all winters and each of the independent variables tested. Non-stationarity is indicated if the inter-quartile range of the local distribution is greater than 2 standard deviations of the global distribution; all relationships in this study exhibit non-stationarity.* 76

Acknowledgments

This project could not have been possible without the collective expertise, guidance and general goodwill of many good people. A very huge THANKYOU must first go to Arko Lucieer, *Super*-visor extraordinaire! Your calm encouragement and guidance made each week of this experience that much more possible. Thankfully, it will never be known just where my project may have wandered to if it weren't for our weekly meetings….! And to Richard Mount, the very best secondary *Super*-visor anyone could hope for! Your enthusiasm for this project and your general encouragement throughout it all meant a whole lot to me.

A big thankyou also goes to Manuel Nunez for great support in the early months of this study, and for putting me in touch with some other quite knowledgeable folk (before disappearing to Spain to lounge about in the sun!) A big thankyou also goes to Rob Musk for extensive help (and patience) while I was straining to get my head around spatial statistics. Your brain is an enigma.

To Bill Wood, Andreas Ernst and Jim Markos in Launceston for each providing considerable input and encouragement in the early stages, and for tolerating my vagueness as I struggled to find the right path for this study!

To Dr Richard Wood-Baker from the Respiratory Research branch of the Menzies Research Institute – this project would probably never have gotten off the ground in its current form without your considerable help with data acquisition, and general encouragement. Your ongoing support throughout the year has been very much appreciated. A very big thankyou must go to Mike Power at the Environment Division of DTAE, Tasmania for enormous amounts of help with everything to do with air pollution dispersion modelling in the Tamar Valley, and for so generously OFFERING TO RUN TAPM FOR ME!!! which added a very important dimension to my work that I simply wouldn't have had time to do myself. Is there enough chocolate in the world to repay you?? Huge thankyous also go to Darren Turner (UTAS) and Tony Miller (Eighty Options) for extensive help converting a concept from my head into various computer programs that drove the process of de-identification of address points.

And of course the biggest thankyou goes to Dom! For helping to keep me mostly sane this year, and for putting up with me for the rest of the time too… for nurturing our garden when I was stuck to the computer… and above all for encouraging me to follow my heart and produce meaningful work. You are an inspiration.

Abstract

The detrimental health effects of particulate air pollution have been well established through environmental health research worldwide. Fine or 'respirable' particulate matter derived from combustion sources has been linked to both acute and chronic respiratory and cardiovascular conditions, and premature death in the most susceptible of a population. The Tamar Valley in northern Tasmania has a significant winter air pollution problem. Launceston is the largest population centre in the valley (population approx. 67,000) and despite its size this small city has regularly recorded the highest levels of particulate pollution levels of any city in Australia. This is due largely to complex geographic and climatic processes that support cold air drainage and the formation of night-time temperature inversions in the valley over winter months. Under these conditions ground temperature drops and air pollution becomes trapped at ground level under a layer of dense cold air. Fine particulate matter from domestic wood heating contributes to around 88% of particulate load in Launceston compared to 65% in other Australian cities. Concern has therefore been raised for the respiratory health of Tamar Valley residents in recent years. Previous studies have assumed homogeneity of pollution exposure, and disease risk, across the landscape. This assumption is unrealistic, as recent research indicates that both the distribution of disease and the dispersal of particulate air pollution exhibit considerable spatial variation.

This is the first study to look in detail at the spatial relationships between particulate air pollution and respiratory disease distribution in the Tamar Valley. Disease clustering was investigated and various environmental processes were explored in detail to explain the spatial disparity of disease distribution. Patterns of respiratory disease occurrence in the Tamar Valley were investigated through spatial analysis of 15 years (1992-2006) of de-identified hospital admissions records. Issues of confidentiality and geoprivacy in spatial public health studies were discussed in detail. Spatial distributions of Asthma, Bronchiolitis, Bronchitis and Chronic Obstructive Pulmonary Disease (COPD) were explored individually and in combined form. Data were explored for annual variations in disease distribution. This revealed that, while disease incidence generally declined over the study period, this decline was most noticeable around George Town in the north of the valley. Further analysis revealed little spatial variation in seasonal spatial patterns of disease occurrence across the valley, though disease cases generally were more numerous in winter. COPD incidence was found to be highly clustered in a small number of address locations thought to correspond to nursing homes and aged care facilities across the valley. It was therefore believed that COPD

was more closely correlated with the locations of these facilities than with any geographic or climatic processes. Three techniques for the detection of disease clusters were applied (kernel density function, Getis Ord Gi* statistic and Kulldorff's spatial scan statistic). Areas around George Town and the North Esk valley east of Launceston consistently showed elevated disease levels. However, considerable variation in the reporting of 'significant' clusters was noted between methods, and also with the same method at different spatial scales. Issues of statistical inference were therefore discussed.

Several 'exposure surfaces' were created to approximate the winter dispersion of particulate air pollution in Launceston. Modelled air pollution concentrations were derived from TAPM (The Air Pollution Model), a prognostic air pollution dispersion model currently in use in Tasmania for environmental monitoring purposes. A digital elevation model was also classified into terrain features that are known to accumulate high levels of particulate pollution through the process of cold air drainage (i.e. low-lying channels and river flats). Spatial relationships between disease incidence and these air pollution 'proxies' were then explored in detail. Weak relationships were found between disease incidence and terrain features representing small channel and valleys. A 'significant' relationship was found between disease incidence and the valley floor, though issues of statistical inference were again discussed in this context. Spatial non-stationarity was detected in all relationships, indicating that global statistics inadequately define these relationships. A strong *inverse* relationship was found between modelled air pollution concentrations and disease incidence, indicating that disease rates were generally higher in areas outside the modelled air pollution plume derived by TAPM. TAPM concentrations were also found to closely mirror the underlying population distribution. The inability of TAPM to adequately predict pollution levels in areas outside major population centres, and various issues of socioeconomic confounding were discussed as possible explanations for this finding.

Results generally revealed considerable variation in the spatial relationships between disease incidence and air pollution proxies used in this study. These results argue strongly for the spatial analysis of air pollution relationships to health outcomes, and the continued refinement of methods. None of these findings could have resulted from a purely temporal (non-spatial) investigation.

1 Introduction

1.1 Background

The Tamar Valley in northern Tasmania has a long-standing air pollution problem. Despite its small population and relative absence of large scale industry, Launceston has regularly recorded the highest air pollution levels of any city in Australia (DEH, 2004). This is largely due to the topographic and climatic characteristics of the Tamar river valley, which support the formation of temperature inversions over the winter months (Lyons and expert working party, 1996; Power, 2001).

The movement and dispersion of air pollution within the Tamar Valley has been extensively studied, most comprehensively for the Tamar Valley Airshed Study (TVAS) (Power, 2001). It is known that light katabatic (gravity-propelled) winds roll down the surrounding hill slopes and settle in the valley floor on still winter nights; this produces a layer of dense cold air at ground level that remains trapped under a layer of displaced warmer air until it is dispersed by solar radiation or wind, usually the following morning (Nunez, 1991; Power, 2001; Sturman and Tapper, 2006). Air pollution produced at ground level, typically wood smoke from residential heating (Lyons and expert working party, 1996); (Ayers et al., 1999), therefore accumulates and concentrates within this layer. Seasonal fluctuations in the spatial dispersal of pollution are governed largely by seasonally variable meteorology (Power, 2001).

Particulate air pollution is commonly measured as PM_{10} (particulate matter with an aerodynamic diameter equal to or less than 10 microns). The small particles that comprise PM_{10} are known to penetrate deep into the lungs and beyond into the circulatory system (Dockery and Pope, 1994). Both PM_{10} and the smaller $PM_{2.5}$ have consistently been linked to morbidity and mortality in studies worldwide (Dockery and Pope, 1994; Forastiere, 2004; Pope III, 2000; WHO, 2006a). A 10 μgm^{-3} increase in daily mean PM_{10} is now known to correspond to increases of approximately 1.4% and 3.4% in cardiovascular and respiratory disease respectively, and a 1% increase in mortality from all causes (Dockery and Pope, 1994).

Launceston's air quality has improved in recent years, due largely to local government initiatives and increased public awareness of the health risks of wood heating. In particular, the *Launceston Woodheater Replacement Program* launched in 2001 has contributed to a substantial reduction in wood smoke pollution from residential wood heaters (DEH, 2005b). However, Australia's National Environment Protection Measure (NEPM) standard has been exceeded every year in Launceston until 2005 (DEH,

2005b); 2006 was the only year on record that Launceston's particulate air pollution levels were within allowable national limits.

1.2 Significance of the study

This is the first study to look in detail at the spatial spread of respiratory disease and air pollution in the Tamar Valley. The use of Geographic Information Systems (GIS) is rapidly expanding in the field of public health research as spatial analysis of health data is capable of revealing localised anomalies that can otherwise be hidden within statistical averages of a population. With further industrial developments currently proposed for the upper Tamar Valley, it is vital that we better understand what is happening in terms of public health in surrounding population centres.

A recent time-oriented study conducted by medical researchers (Mesaros et al., 2007) found a weak relationship between air pollution levels and respiratory admissions for bronchitis/bronchiolitis in the Tamar Valley. This study assumed, however, that levels of pollution were uniform across the valley, which may have reduced the strength of any association. There is now strong evidence to suggest that spatial disparity in particulate air pollution concentration is significant even at the neighbourhood scale, and that assumptions of homogeneity can lead to false conclusions being drawn about both the magnitude and significance of pollution-related health outcomes (Maheswaran and Craglia, 2004; Monn, 2001; Wilson et al., 2004; Wilson et al., 2006; Wilson and Zawar-Reza, 2006)

1.3 Aims and objectives

This study is a detailed pre-epidemiological study predicated on understanding the spatial distribution of respiratory disease and its relationship to the dispersal of particulate air pollution in the Tamar Valley.

The objectives of the study are to:

- Effectively deal with issues of confidentiality and geoprivacy in the analysis and publication of spatial public health information;
- Investigate the spatial patterns of acute respiratory disease occurrences, and thereby identify areas of increased relative risk of disease in the valley;
- Investigate annual and seasonal fluctuations in spatial patterns of disease;
- Examine various exposure surfaces derived from modelled air pollution dispersion data and terrain analyses as proxies for winter cold air drainage in the Tamar Valley, and thereby identify possible areas of increased relative risk of pollution exposure in the valley;

- Explore the spatial relationships between pollution exposure surfaces and disease locations, and thereby comment on the spatial relationships between environmental air pollution and acute respiratory illness in the Tamar Valley;
- Explore the uncertainties inherent in this type of public health research.

1.4 Structure of the thesis

Chapter one provides some background information and outlines the importance of conducting this research.

Chapter two contains a review of relevant epidemiological and ecological studies, with an emphasis on the spatial interrogation of medical data and its use in air pollution exposure analyses.

Chapter three details the pre-processing of data to ensure geoprivacy of medical records. Aspects of ethical use of spatial public health data are addressed.

Chapter four provides a detailed spatial investigation of acute respiratory disease in the Tamar Valley based on hospital admissions records from 1992 to 2006. This chapter utilises de-identified individual-level data for a high-resolution spatial analysis of disease clustering in the valley.

Chapter five combines spatial disease data with various 'exposure surfaces' derived from air pollution modelling and terrain analysis to investigate spatial relationships between respiratory disease occurrence and climatic/geographic factors known to influence PM_{10} concentration.

Finally, chapter six concludes the thesis and offers suggestions for further research.

2 Spatial epidemiology of respiratory disease: review

2.1 Particulate air pollution and health

The adverse health effects of particulate air pollution are well documented. Research worldwide has consistently shown that increased levels of particulate air pollution are associated with increased morbidity and mortality, most noticeably from respiratory and cardiovascular causes (Abramson, 2001; Ackermann-Liebrich et al., 1997; Brook et al., 2007; Chen et al., 2006; Christie et al., 1992; Dockery and Pope, 1994; Forastiere, 2004; Fusco et al., 2001; Gilliland et al., 2005; Goldberg et al., 2006; Hales et al., 1999; Jerrett et al., 2005b; Leem et al., 2006; McGowan et al., 2002; Medina et al., 2004; Mesaros et al., 2007; Moolgavkar, 2000; Oyana et al., 2004; Peel et al., 2006; Pope III, 2000; Salvaggio, 1994; Sheppard et al., 1999; Simpson et al., 1997; Ulirsch et al., 2007; Wordley et al., 1997; Yunesian et al., 2006). The historic case of a catastrophic fog in London in 1952 is the most dramatic example of this relationship, with drastically elevated pollution levels causing widespread respiratory problems and the premature deaths of an estimated 12,000 people (Bell and Davis, 2001; De Angelo, 2006).

Along with sulphur dioxide, nitrogen dioxide and ozone, particulate air pollution is known to exacerbate or cause diseases of the respiratory and cardiovascular systems, resulting in both acute and chronic disease conditions and premature death for the most susceptible in the population (WHO, 2005). Acute respiratory conditions with short latency periods, such as Asthma, Bronchitis and Bronchiolitis, are well suited to investigations of the relationships between air pollution and health. These diseases are especially responsive to fluctuations in particulate air pollution, which gives less opportunity for bias to occur from other diseases (Elliott and Wakefield, 2000).

2.1.1 PM_{10} and $PM_{2.5}$ defined

Particulate matter with an aerodynamic diameter equal to or less than 10 microns, or ten thousandths of a millimetre, is commonly referred to as PM_{10} or 'inhalable' particles. A measure of PM_{10} includes quantities of all particles equal to and smaller than 10 microns (10 μm). $PM_{2.5}$ refers to particles 2.5 μm in diameter and smaller and is referred to variably as the 'fine' or 'respirable' fraction of PM_{10}. Airborne concentrations of PM_{10} and $PM_{2.5}$ are typically expressed in micrograms per cubic metre (μgm^{-3}).

Particulate air pollution is most commonly expressed as the highest or average mass of PM_{10} per volume of sampled air per day, or per year. For public health reasons however, there is debate around the optimal size fraction that should be monitored

(i.e. PM_{10}, $PM_{2.5,}$ PM_1 etc), whether to measure mass or particulate counts, and the most appropriate temporal averaging period (Elliott et al., 2000a).

Ambient PM_{10} concentrations have been measured at a permanent monitoring station at Ti Tree Bend, northern Launceston, since 1992 and this facility has recently been upgraded to also record ambient concentrations of $PM_{2.5}$. There is also a meteorological station at this site.

2.1.2 PM_{10} and $PM_{2.5}$ health impacts

Particles of different sizes impact on the human body in different ways. Research has shown that particles greater than or equal to 10 microns in diameter ($\geq PM_{10}$) are deposited in the upper airways and nasopharynx region of the respiratory tract, while finer particles ($\leq PM_{2.5}$) penetrate deep into the lungs and beyond into the bloodstream (Salvaggio, 1994). This information has been useful in determining relative air pollution types and sources in epidemiological studies where specific respiratory symptoms are known. For example, $PM_{2.5}$ is known to cause irritation and inflammation of the lower respiratory tract and is also known to originate from smoke (Dockery and Pope, 1994; Pope III, 2000), while irritations of the upper respiratory tract (nose and throat) are generally caused by dust particles of mechanical origin, which constitute the larger size fraction of PM_{10} (WHO, 2005). By studying the specific respiratory symptoms of participants in a study (i.e. presence of wet cough, dry cough, sneezing, wheezing, etc.) it is possible to deduce the relative atmospheric conditions to which they were exposed, with respect to appropriate latency periods.

The Australian Fine Particles study of 1996-1997 (Ayers et al., 1999) systematically analysed the physical and chemical properties of particulate air pollution and corresponding disease occurrences in several Australian cities and reported that, "...the statistical relationships noted between health outcomes and PM_{10} are most likely caused by the $PM_{2.5}$ component of the size fraction…", and, "...stronger relationships might be observed between $PM_{2.5}$ and health outcomes, than are currently observed between PM_{10} and health outcomes, if more $PM_{2.5}$ data were available" (Ayers et al., 1999 p89). As dedicated monitoring of $PM_{2.5}$ has only commenced in the last couple of years in Australia, PM_{10} is still being used for retrospective epidemiological studies.

2.1.3 National and international standards

The World Health Organisation (WHO) has recently updated guideline levels for air pollutants, recommending across-the-board reductions in pollutant levels globally (WHO, 2006b). Allowable concentrations of PM_{10} have now been set at 50 μgm^{-3} for 24 hour average and 20 μgm^{-3} for annual average levels; 24 hour and annual mean levels of $PM_{2.5}$ have been set at 25 and 10 μgm^{-3} respectively (WHO, 2005). It is believed

that this reduction in allowable limits will, if met, reduce worldwide air pollution-related mortality by around 15% (WHO, 2006a). However, while there is currently no recommended maximum exposure dosage (Simpson et al., 1997; WHO, 2006a), it has been suggested that the relationship between particulate air pollution and disease is linear, and as such there is no threshold of particulate air pollution below which there would be no observable health effects (McGowan et al., 2002).

In line with WHO guidelines, and under the *National Standards for Criteria Air Pollutants*, Australia's National Environment Protection Measures (NEPM) air pollution standards are currently set at 50 μgm^{-3} for 24 hour average concentrations of PM_{10}; $PM_{2.5}$ limits are set at 25 μgm^{-3} for daily maximum levels and 8 μgm^{-3} for annual averages (DEH, 2005a). Under this standard, the 24-hour average concentrations are allowed to be exceeded five times annually (five allowable exceedances).

2.1.4 Tamar Valley air pollution

The Ti Tree Bend air monitoring station at Invermay, northern Launceston, has been measuring daily and annual average PM_{10} concentrations since 1992 (though initial measurements were taken only every six days). $PM_{2.5}$ measurements commenced in June 2005. Results have demonstrated that Launceston, despite its low population, has regularly experienced some of the highest air pollution events in the country. For example, in 1992 the highest recorded PM_{10} concentration at Ti Tree Bend was 186 μgm^{-3}, or close to four times what is now the allowable limit, and in 1997 the NEPM 24-hour average standard of 50 μgm^{-3} was exceeded on 51 days, or 46 more than the standard now allows (data from Department of Primary Industries and Water (DPIWE, 2007)). Figure 2.1 shows a plot of the sixth highest 24-hour average concentrations of PM_{10} (i.e. the next most polluted day after the five allowable exceedances) in each capital city in Australia and Launceston over a decade, and shows the magnitude of Launceston's air pollution problem.

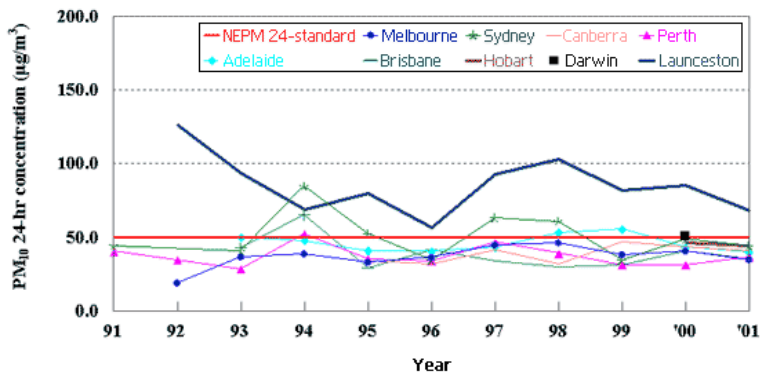


Figure 2.1- PM10 concentrations on the sixth-highest recorded day of each year 1991 to 2001 for Launceston and every Australian capital city. The red line indicates the NEPM 24 hour standard of 50 μgm^{-3}, which may be exceeded just five times annually under the national standard. (*Source: Australian Government Department of the Environment and Water Resources, 20 Jan 2006*)

Leading studies have found that wood smoke is the main contributor to particulate air pollution in Launceston in winter (Ayers et al., 1999; DEH, 2005b; Lyons and expert working party, 1996), while other major Australian cities of larger population experience higher proportions of automotive and industrial emissions (Ayers et al., 1999). Accordingly, the proportion of $PM_{2.5}$ in PM_{10} was found to be up to 88% in Launceston in winter, compared to around 65% in other cities (i.e. Sydney, Melbourne, Adelaide and Brisbane). In light of this, the new NEPM $PM_{2.5}$ 24-hour standard of 25 μgm^{-3} coming into effect in 2008 will effectively require a further halving of particle pollution concentrations in the Tamar Valley (Power, 2007).

While particulate matter from residential combustion sources is known to be the primary source of air pollution in the Tamar Valley, emission rates alone do not account for the elevated air pollution levels; unfavourable geographical and meteorological factors are known to compound the problem (Lyons and expert working party, 1996). In particular, stable atmospheric conditions associated with anticyclonic weather patterns in winter are now known to be responsible for the accumulation and concentration of PM_{10} in the Tamar Valley. And most notably, cold air drainage transports pollutants downhill, which has been predicted to result in concentrations becoming highest in the low-lying areas of the North Esk valley, Glen Dhu and southern Launceston (Lyons and expert working party, 1996; Nunez, 1991).

Figure 2.2 – Landsat and Quickbird satellite images of all of Tasmania (left) and just the Tamar Valley (right). In the image on the left, a large band of fog is seen covering the entire Tamar Valley and extending southeast to the midlands area. The image on the right shows the area around Launceston blanketed by low cloud/fog. *(Sources: SOER / Google Earth (2007))*

In 1996, at the completion of an extensive study into the air pollution problem in Launceston, it was recommended that, "all possible measures should be taken by the community and authorities to reduce the causes of particulate air pollution in the local

environment, as a matter of urgency" (Lyons and expert working party, 1996). Following from this study, the situation in Launceston and the Tamar Valley has improved in recent years, due in large part to public education and local government initiatives. In particular, the *Wood Heater Replacement Program* established in 2001 has greatly reduced the particulate air pollution load in the valley. However, concern has been raised that this improvement in concentrations may have reached a plateau at what is still an unacceptably high level (ABC Northern Tasmania, 2004).

2.1.5 Tamar Valley respiratory health studies

Launceston's elevated PM_{10} levels and high $PM_{2.5}/PM_{10}$ ratio, coupled with the known relationships between $PM_{2.5}$ and diseases of the lower respiratory tract, have prompted considerable focus on the respiratory health of Tamar Valley residents over the years. A study of the geographical spread of childhood asthma across Tasmania was completed in 1980 and, using the location of primary schools that were attended by participants as the location of 'cases', this study found that "the most severely affected asthmatics were most prevalent in a wide area around the Tamar Estuary in the North of the state. The proportion of cases in this area was over twice the norm; this was significant at p=<0.01" (Giles, 1980). Giles (1980) observed that the number of cases per population in the Tamar Valley in this class of "sever, chronic" Asthma was 17/209, or 8.3%; the observed Tasmanian state average was 2.9%.

An expert working party consisting of scientific and medical researchers was formed in 1991 to further investigate the effects on respiratory health of what was then a worsening air pollution problem in Launceston and the upper Tamar Valley (Lyons and expert working party, 1996). Launceston General Hospital (LGH) records were examined and evidence suggested an increasing pattern of respiratory disease since 1980, which was worse in winter. Hospital records between 1992/93 were subject to a more detailed analysis of correlations with field PM_{10} measurements, though small population numbers made statistical inferences difficult. International evidence of similar studies was cited as indicating that a lack of statistical significance in small-population studies such as this did not necessarily preclude the existence of a relationship (Lyons and expert working party, 1996).

A more recent study attempted to remedy the problems of small population numbers by analysing accumulated hospital admissions and Ti Tree Bend monitoring data from 1992 to 2002 (Mesaros et al., 2007). This study found a significant relationship between winter PM_{10} concentrations and hospital admissions for bronchitis/ bronchiolitis. A weak association was also found between hospital admissions and air pressure (Mesaros et al., 2007), which would seem to be associated with the anticyclonic conditions that drive pollution build up in the valley. In addition, every 10

μgm^{-3} increase in PM_{10} was found to result in a 4% rise in hospital admissions. This study agrees with the findings of an extensive review conducted by Dockery and Pope (1994) of similar studies, which found an average increase of 3.4% in respiratory disease for every 10 μgm^{-3} rise in PM_{10}.

To date, studies of the relationship between air pollution and disease in the Tamar Valley have been either non-spatial (e.g. (Lyons and expert working party, 1996; Mesaros et al., 2007) or of very coarse spatial resolution (e.g. Giles, 1980) and have therefore assumed homogeneity of exposure across the valley. As recent literature suggests, however, this assumption may in fact cause pollution exposure, and disease rates, to be underestimated in some areas and overestimated in others (Brindley et al., 2004; DEH, 2005b; Durand and Wilson, 2006; Kingham et al., 2006; Wilson et al., ; Wilson et al., 2006; Wilson and Zawar-Reza).

2.2 Spatial epidemiology

Spatial epidemiology, or medical geography, has existed for centuries – the most widely reported study being Snow's 19th century investigation of cholera outbreaks in London (Snow, 1854). With the development of Geographic Information Systems (GIS) in the past few decades, however, spatial epidemiology has become vastly more accessible (Koch and Denike, 2004; Maheswaran and Craglia, 2004).

2.2.1 Why spatial?

The principle advantage of a spatial approach to epidemiology over traditional statistical analyses lies in the fact that spatial studies are sensitive to geographic anomalies which in non-spatial studies can be hidden within statistical averages of a study population (Koch and Denike, 2004). A spatial study therefore is capable of revealing local or regional 'clusters' of elevated or reduced disease incidence that are otherwise absorbed into generalised population estimates. Maheswaran and Haining (Maheswaran and Craglia, 2004) comment that, "a substantial amount of public health information has an intrinsic spatial component, which is often not realized. The use of GIS....is strongly recommended to bring an added dimension to public health intelligence."

Elliott *et al* (2000a) describe four different realms of spatial epidemiology, the most relevant to this study being 'disease mapping' and 'geographical correlation studies' (elsewhere called 'ecological studies'). Disease mapping is described as an essentially descriptive analysis method used to "… summarise spatial and spatio-temporal variation in risk", while geographic correlation studies go some way toward explaining disease aetiology (causes or origins of disease). The key difference here lies between studying the incidence of disease, and studying the relationship between disease

incidence and another factor (Koch and Denike, 2004; Sexton et al., 2002). It should be appreciated, however, that a true aetiological study requires a considerable amount of background information relating to the personal risk factors – or confounders - of individual 'participants' in the study, with the ultimate goal being to estimate the true exposure risk to communities in scenarios of both exposure and non-exposure (Sexton et al., 2002). We are alerted to the dangers of jumping to conclusions with cause-and-effect studies here, as Schwartz *et al* (1996) point out that many correlated events may not be at all causative or related (e.g. HIV AIDS and stock market increases).

2.2.2 Confounders in spatial epidemiology

There are a great many confounding factors in epidemiological studies, the effects of which are many and varied. Both identifying and controlling for confounders, however, can be complicated (Greenland and Robins, 1986; Maheswaran and Craglia, 2004). A recent study in Tehran for example identified the following confounders in a study of atmospheric air pollution and morbidity: daily temperature, humidity, day of the week, season, 'holiday effect', smoking history (several sub-categories identified), exposure hours, occupational exposure, home heating method, presence of common cold in household, history of asthma, age, gender, education and economic status (Yunesian et al., 2006). Others include marital status, family size, unemployment or rank of employment, various measures of indoor air quality, and housing proximity to pollution sources (Gilliland et al., 2005; Jerrett et al., 2003; Pope III, 2000; Schwartz et al., 1996). Some illnesses can even act as confounders for the disease of interest, such as diabetes and respiratory illnesses being confounders for cardiovascular disease (Peel et al., 2006). Long term cigarette smoking also produces a range of respiratory symptoms similar to those associated with exposure to fine particulate matter, though the effects of smoking are greater (Pope III, 2000).

Spatial epidemiology then introduces its own unique confounder: spatial autocorrelation (Jerrett et al., 2003; Lipton et al., 2005). Positive spatial autocorrelation dictates that objects that are close together are more similar than objects that are further apart. In contrast to correlation – which is a measure of similarity between two different datasets – autocorrelation is a measure of how similar the individual components of a single dataset are to each other and how these relationships change with distance. In an urban context, spatial autocorrelation exists in the uneven organisation of houses across the city, as well as in the heterogenous social structure of the population (Hayes, 2003). Failure to account for spatial autocorrelation in epidemiological studies can influence both the magnitude and significance of any relationships drawn between air pollution and health effects (Jerrett and Finkelstein, 2005).

2.2.3 GIS and 'Environmental Justice': a case for the spatial approach

Many studies have shown that within a given population, regions of lower socio-economic status show generally higher rates of morbidity and mortality than more affluent areas (Carstairs, 2000; Hayes, 2003; Jerrett et al., 2003; Koch and Denike, 2004; Lipton et al., 2005; Maantay, 2002; Snow, 1854; Yunesian et al., 2006). Compared to the more affluent in society, the deprived are more likely to live in poorly insulated houses and in less favourable geographical locations, experience higher rates of occupational exposure to industrial pollutants, smoke more heavily and experience generally diminished levels of overall health.

In North America in particular, studies of environmental justice are consistently showing that certain geographical sections of the population, often based on racial and socioeconomic divides, are more disadvantaged than others in terms of both exposure to environmental pollutants and potential impacts on health (Lipton et al., 2005; Maantay, 2002). Elliott and Wakefield (2000) comment that this socio-economic confounding can present a real problem in studies interested in disease aetiology, as socio-economic factors are strongly correlated with both disease occurrence and levels of pollution and industry (as the poorest in society suffer more health problems generally, and often work in and/or live near sources of industrial pollution). Thus, an association between industry and disease will be detected even in the absence of air pollution effects (Elliott and Wakefield, 2000). This has been described as a special case of spatial autocorrelation.

Adjusting for deprivation, however, may be inappropriate in a study interested in the effects of air pollution per se. Regardless of whether the issue is one of environmental equity, or true air pollution exposure, an elevated incidence of disease may be reason enough for further investigation and/or intervention (Elliott and Wakefield, 2000; Lipton et al., 2005).

2.3 Air pollution exposure estimates

Studies investigating the relationships between PM_{10} and disease occurrences require some way of determining the level of pollution exposure experienced by the study group. A majority of non-spatial studies have traditionally relied on PM_{10} monitoring data from a single or few locations in the vicinity of the study area to test for correlations with disease rates (Goldberg et al., 2006; McGowan et al., 2002; Medina et al., 2004; Mesaros et al., 2007; Ulirsch et al., 2007). This approach is now being cautioned against though, particularly for studies of urban environments (Wilson et al., 2006).

Spatial studies on the other hand, require the generation of an 'exposure surface' as a means of estimating the geographical variations in pollution exposure of a population or an individual for the area and time period of interest.

2.3.1 Exposure surfaces – the spatial approach to exposure classification

Various techniques have been employed for creating exposure surfaces, ranging in sophistication from interpolating measured PM_{10} data from a network of monitoring stations (e.g. (Brindley et al., 2004; Leem et al., 2006; Liao et al., 2006)), to modelling air pollution dispersion within the region based on local meteorology, topography and several other known variables (e.g. (Scoggins et al., 2004; Stedman, 2007; Wilson and Zawar-Reza, 2006)). Some small-scale studies have also supplied participants with personal monitoring devices that directly measure individual exposures across time (Monn, 2001; Zeger et al., 2000). Other researchers have employed various proxies for exposure estimation such as land use type, terrain features, distance to known pollution sources, and land cover and thermal signals from satellite imagery (Corburn, 2007; Elliott et al., 2000a; van de Kassteele et al., 2006; Weng and Yang, 2006).

Inevitably, uncertainty exists in the estimation of all exposure surfaces, however, as pollution exposure varies between individuals and for any one individual across time (Wilson et al., 2004; Wilson et al., 2006; Zeger et al., 2000). It is therefore challenging to produce a valid and comprehensive exposure layer, even for small geographic areas (Elliott et al., 2000b).

2.3.2 Exposure misclassification

Exposure misclassification has been described as the weakest link in epidemiological studies (Elliott et al., 2000b). This arises when false assumptions are made about either the spatial distribution of pollution across the study area or the assumed level of exposure of each individual. The impacts of exposure misclassification on studies of disease aetiology can be considerable (Ito et al., 1995; Wilson et al., 2004), thus it is important to recognise the inherent value and limitations of a study before making categorical assumptions about disease aetiology based on apparent correlations between disease occurrences and exposure surfaces.

This study utilises the de-identified residential address of respiratory patients as the presumed point of exposure to outdoor air pollution. While this is a common approach in ecological studies, it has also been strongly criticised (e.g. (Briggs, 2005)). Certainly, a failure to account for other forms of exposure (e.g. occupational, indoor, vehicular, etc.) leaves this study vulnerable to exposure misclassification. However, a fundamental assumption being made here relates to the fact that winter air pollution in

the Tamar Valley is worst at night ((Lyons and expert working party, 1996; Nunez, 1991; Power, 2001), and this is a time when most people are assumed to be at home.

3 Data acquisition and pre-processing

3.1 Introduction

This chapter details the pre-processing of Launceston General Hospital admissions data in preparation for disease cluster detection. De-identification of medical records, required to protect patient privacy, is explained in detail. The dataset spans a 15 year period from 1992-2006 and includes abridged records of all hospital inpatients whose ICD-9/ICD-10 (International Classification of Disease) discharge code corresponded to Asthma, Bronchitis, Bronchiolitis or Chronic Obstructive Pulmonary Disease (COPD) as the primary disease condition.

3.1.1 Point data vs. aggregate data

Spatially represented medical data is most often available to researchers in aggregate form. That is, 'point' address locations of disease occurrence are grouped into some larger, often arbitrary, unit of space. The size of aggregation unit varies with the spatial scale of the study, though for micro scale studies such as this, data are commonly aggregated into local government areas, 'urban'/'rural' divisions, Census collection districts or postcodes. It is becoming increasingly evident, however, that aggregate data is often not sufficient to answer important spatial questions about environmental effects on public health (Armstrong et al., 1999; Boulos et al., 2006; Kwan et al., 2004; Wheeler, 2007).

Ideally, researchers would have access to address data in point form (or 'case event data'), though confidentiality requirements often limit access to this information (Boulos et al., 2006; Maheswaran and Craglia, 2004). Issues of confidentiality and geoprivacy are addressed more thoroughly in section 3.1.2 below. Using point data that has undergone some level of masking to preserve confidentiality is the most preferable method for small area ecological studies (Armstrong et al., 1999).

Following ethics approval, this study was granted access to a dataset of much higher spatial resolution than the majority of comparable studies. Patient address locations were requested and supplied in point form rather than aggregate form, which greatly enhanced the potential for detailed spatial analysis. However, working with point data instead of aggregate data introduces issues of patient confidentiality and ethical reporting that are not generally considered in studies of coarser spatial resolution.

3.1.2 Confidentiality, geoprivacy and ethical reporting

In public health research, confidential information is generally accepted to mean any demographic or medical information that can potentially identify an individual (Quinn,

1997). 'Geoprivacy' then refers to the protection of information that could identify an individual's home, place of work, or path of travel (Kwan et al., 2004). In order to preserve patient privacy in studies of spatial health data, researchers must employ some method of ensuring that both privacy and geoprivacy of individuals is maintained. These can be implemented prior to data analysis (e.g. (Wheeler, 2007)), prior to publication of maps (e.g. (Curtis et al., 2006)), or both.

The publication of spatial health data in map form presents real concerns for maintaining public privacy and confidentiality. The rapid rise in popularity of web mapping (Google Earth) has had a powerful impact on communities lacking the means to otherwise use GIS (i.e. either the hardware, software or technical ability) (Curtis et al., 2006). Researchers are required to act responsibly. Curtis *et al* (2006) warn that "...any map containing point data, even when little secondary spatial information is presented, is vulnerable to being re-engineered [post-publication] to reveal the actual addresses associated with the points. It is therefore vital that some masking occurs of the original point data."

Three core methods of preserving geoprivacy through the application of geographical masks are reported in the literature: aggregation, affine transformations and random perturbations (Armstrong et al., 1999; Kwan et al., 2004). The spatial accuracy lost through aggregation of point data has already been discussed; this is perhaps the least preferable option of the three. Affine transformations impose a systematic shift in the point dataset as a whole that maintains the positions of the points relative to each other but shifts them all in some way within the study area (i.e. rotation, re-scaling or moving) (Armstrong et al., 1999).

Random perturbation then is an introduction of random error into the dataset that shifts *each point* a random distance, within a defined upper limit, from its true location (Armstrong et al., 1999; Kwan et al., 2004). Random perturbation, or 'random scrambling', was chosen as the method of de-identification most suited to this current study.

3.2 Data acquisition and pre-processing

3.2.1 Project ethics approval

Ethics approval was granted for this project from the University of Tasmania Social Science Ethics Committee on 19 May 2007 (project number H9382). The committee granted permission to utilise de-identified patient address data in point form sourced from medical records obtained through the Department of Health and Human Services, Tasmania (DHHS). De-identification methods are detailed in section 3.3 below.

3.2.2 Spatial extent of the study area

All data in the study were analysed and displayed in the projected coordinate system GDA 94 UTM Zone 55S (MGA). All spatial analyses and map production was done using ESRI software ArcGIS v 9.1. The study area was aligned to coincide with the spatial extent of the Tamar Valley Airshed Study (TVAS) for ease of comparison with this work (see (Power, 2001)). This bounding box had the following coordinates: Top: 5468324.32, Left: 464316.71, Right: 535415.74, Bottom: 5392170.94 (see Figure 3.1). A rectangular area with these coordinates was digitised and all datasets were 'clipped' to this extent at various stages of analysis.

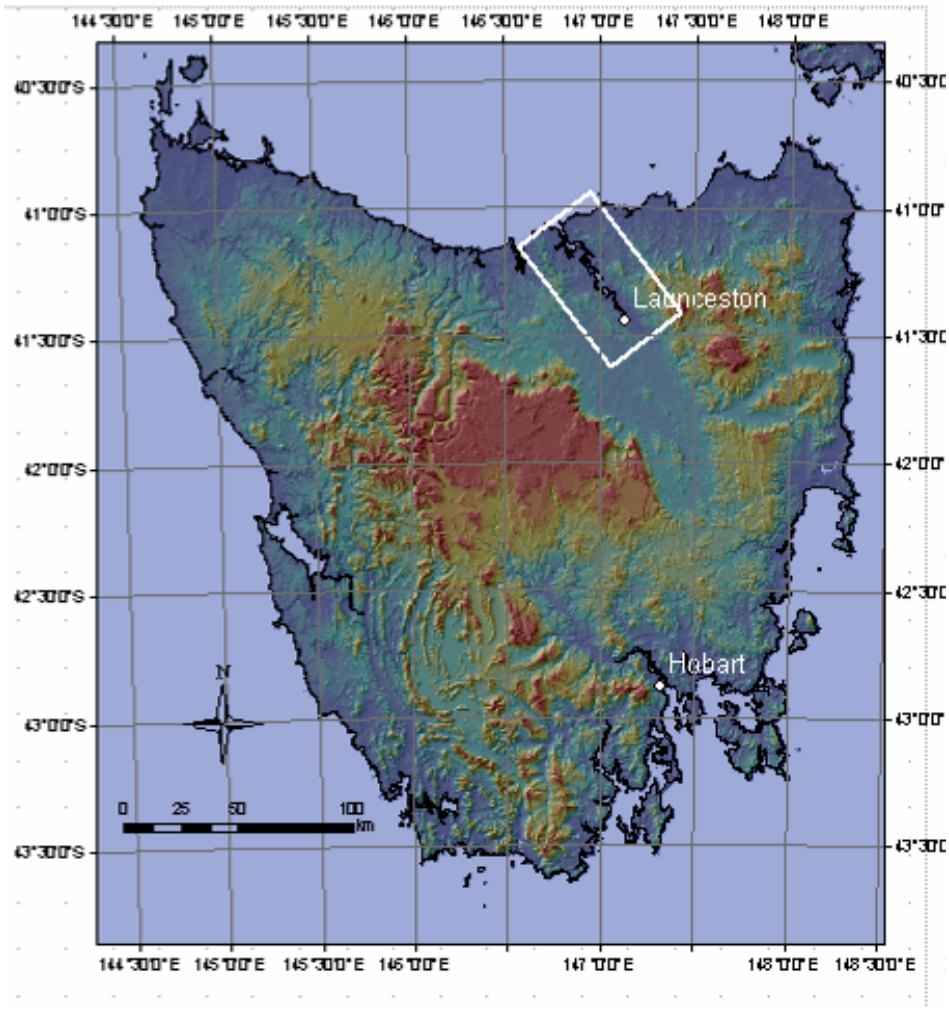


Figure 3.1 – The white rectangle demarcates the spatial extent of the Tamar Valley study area in northern Tasmania

3.2.3 Address points dataset

A section of the Address Points Dataset (APD) was provided by Land Information Services Tasmania (theLIST). This dataset in its full spatial extent contains, among other things, the street address and centroid XY co-ordinate of every residence in Tasmania (theLIST, 2007).

For urban and rural properties less than 1 ha in area, each point location in the APD corresponds to the centroid (spatial mean point) of the cadastral parcel. Rural properties greater than 1 ha in area are defined by a point location 50 m perpendicular to the road midline at the point of access to the property (theLIST, 2007). Spatial 'accuracy' of these points is reported as being within 5 m for properties less than 1 ha and 50 m for properties greater than 1 ha. Properties with multiple residences (e.g. a block of flats) are listed as a single address point, corresponding to the street address of the parent property (theLIST, 2007).

While points in the APD do not always correspond to the absolute spatial location of the house on a given property, it was believed that the dataset gave an adequate estimation of 'place of residence' for the study population. It also provided a platform from which to launch the de-identification process detailed in section 3.3 below. Figure 3.2 shows the APD coloured by postcode and illustrates that using individual address points offers vastly greater spatial resolution than if data were aggregated by postcode.

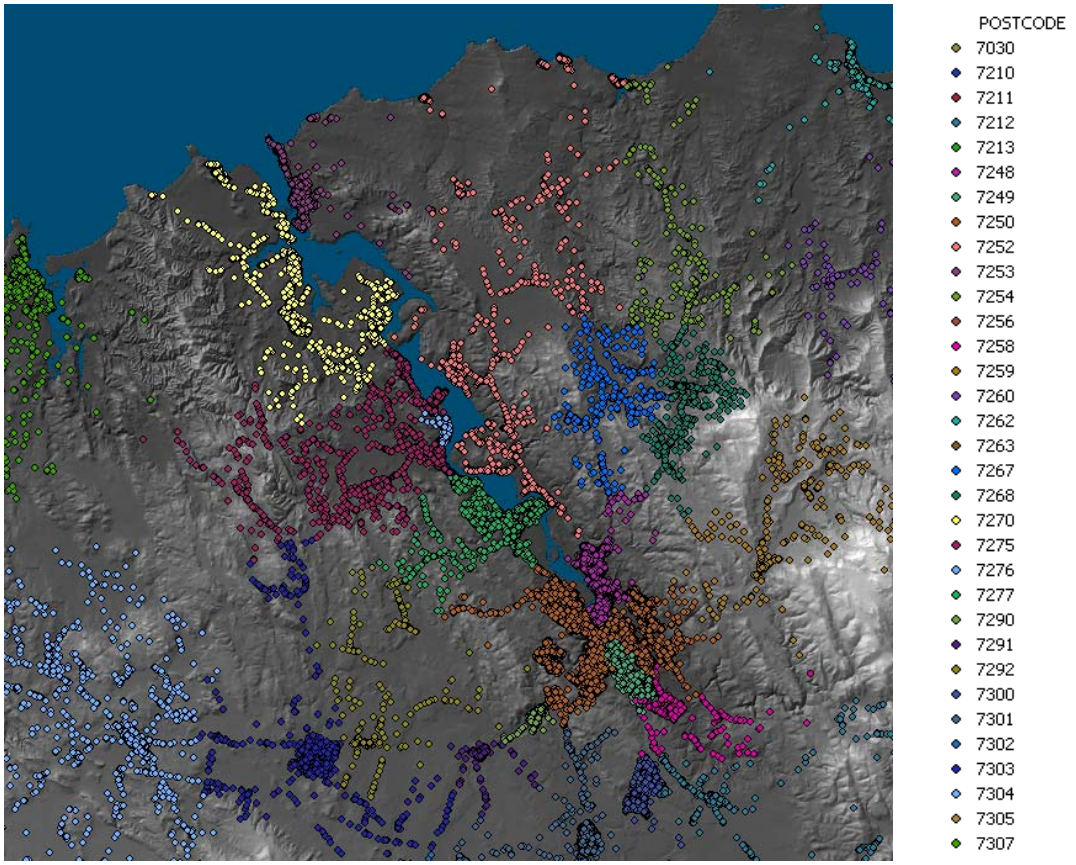


Figure 3.2 - Map of a section of the Tasmanian Address Points Dataset with points coloured by postcode; a digital elevation model is shown in shaded relief to display terrain features. Postcodes in the Tamar Valley cover vast areas of varying terrain, and are therefore not a suitable aggregation unit for micro-scale health studies such as this.

3.2.4 Census data

The Australian Bureau of Statistics (ABS) provided spatially referenced population data for 2001 aggregated to collection district level. There are on average 200 dwellings within each collection district (Australian Bureau of Statistics, 2001), which results in these areas varying greatly in size across both urban and rural regions. Census data for the Greater Launceston region was also sourced in non-spatial format for 1996, 2001 and 2006. Figure 3.3 shows the Address Points Dataset and Census collection districts of the Tamar Valley and surrounding region. A star-shaped census district in the Tamar River, representing expatriates, travellers and people who were otherwise absent from their place of residence on Census night, was deleted from the Census layer.

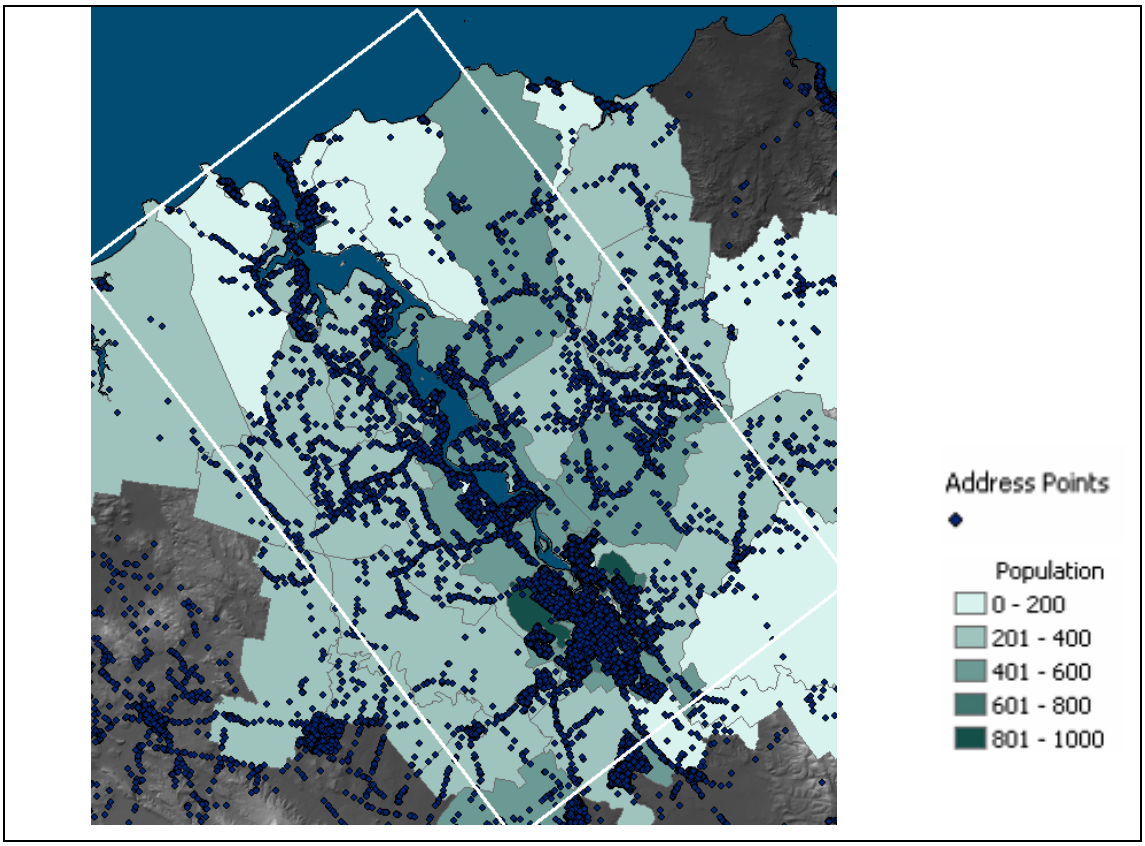


Figure 3.3 - The Address Points Dataset and Census collection districts, coloured according to population numbers. The white rectangle defines the spatial extent of the study area.

3.2.5 Population estimation 'per residence'

An estimation of population 'per residence' was calculated by spatially joining the Address Points Dataset to the 2001 Census collection districts layer in ArcGIS. The total population of each collection district was then divided by the number of 'houses' within it. This gave a crude approximation of population at each address point in the study area. Once this calculation was complete, both the Census layer and the APD were 'clipped' to the spatial extent of the study area (Figure 3.4). Resulting occupancy

numbers ranged from 0.3 to 4.2 occupants per house. In summary, this gave a total population of 94,984 spread over 42,997 addresses for the entire study area.

The population of the Tamar Valley has been described as relatively static (Power, 2001), and statistical records confirm this (Australian Bureau of Statistics, 2007). However, between 1996 and 2001 both Launceston's and George Town's population declined slightly (-0.5%; -1.3%), then increased slightly (+0.7%; +0.8%) from 2001 to 2006 (Australian Bureau of Statistics, 2007). Although these fluctuations are small, the available 2001 spatial data reflects a time when population was at its lowest for the three Census years. In addition to this, census data generally is known to slightly underestimate total population (Diamond, 1997). Therefore all population counts are likely to be modestly underestimated for the study period.

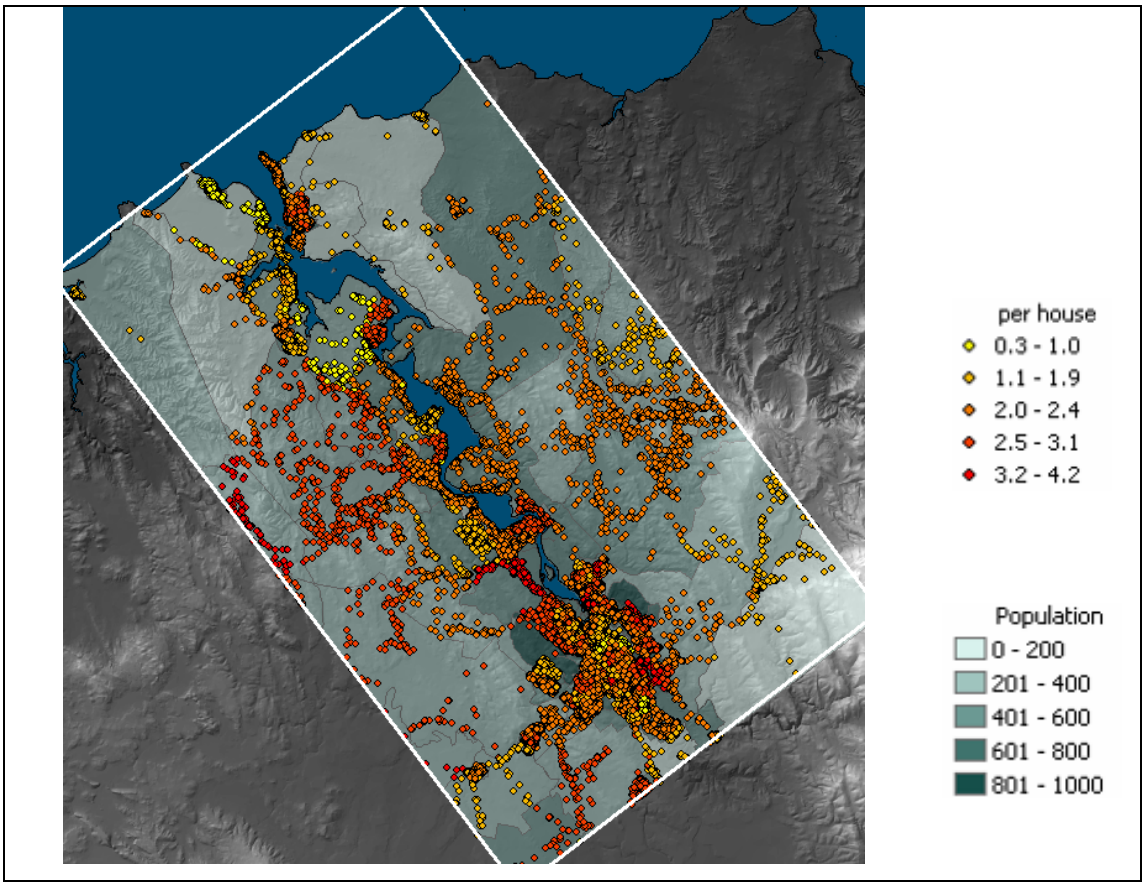


Figure 3.4 - Census collection districts and Address Points Dataset were combined to give an approximation of population 'per house', then clipped to the spatial extent of the study area. Digital elevation model is displayed in shaded relief.

3.2.6 Medical records

Abridged medical records of respiratory admissions to the Launceston General Hospital (LGH) between 1992 and 2006 were sourced from the Department of Health and Human Services, Tasmania (DHHS). These electronic datasets contained a record number, date of admission, length of stay and street address of every admission within

this period. LGH is by far the largest public hospital in the Tamar Valley, with George Town Hospital in the north currently servicing just 15 acute-care beds. As admissions for both hospitals are recorded centrally at LGH, it was believed that the requested dataset was representative of all Tamar Valley inpatients admitted into the public system over the study period.

For reasons of patient confidentiality these datasets could not be accessed directly, and it was therefore necessary to devise a method of 'de-identifying' patient addresses remotely. De-identification is a process whereby personal information is altered or truncated to preserve the privacy of individuals. This process is detailed in section 3.3 below.

Age and gender data of each patient were not requested with the medical records as it was thought that this could facilitate the possible re-identification of individuals. For example, if for a given address there was an admission for a 52-year-old male in 1992 and a 64-year-old male in 2004, then it could reasonably be assumed that that same address now (in 2007) may be home to a 67-year-old male, with a history of respiratory disease. This was believed to be an undesirable outcome. Unfortunately, the absence of age and gender information from the dataset also meant that these demographic confounders could not be accounted for.

3.3 Random perturbation of address points

This section details the de-identification of patient address data through a process of random perturbation, or scrambling. Each point location was shifted a uniform random direction and distance from its true location up to a defined maximum radius. This radial distance was determined intuitively through examining both the APD and a digital elevation model (DEM) of the study area. A distance of 200 m was found to be mutually suitable for maintaining the spatial integrity of the dataset and protecting patient geoprivacy.

Given the large distances between some rural properties, it was considered that rural properties may require scrambling a greater distance than urban ones to adequately preserve geoprivacy. However, this was decided against on the grounds that the spatial locations of rural addresses were already less 'accurate' than urban locations in the Address Points Dataset (as discussed in section 3.2.3 above); the additional spatial error that would have been introduced by shifting rural points even further from their 'true' location was not desirable. It was therefore decided instead to perform the analyses based on a 200 m maximum radial perturbation and remove rural points from maps prior to publication.

The general format of both the medical records (Table 3-1) and address points datasets (Table 3-2) can be seen below.

Table 3-1 - General format of the Address Points Dataset (APD).

Geo-ID	Geo type	Easting	Northing	Street number	Street name	Street type	Locality	Postcode
904232	urban	511248	5414868	23	Smith	Rd	Launceston	7250
1014089	rural	510014	5416897	101	Fred	St	Mt Direction	7277
950296	urban	461707	5446182	49	North	Ave	George Town	7253

Table 3-2 - General format of the medical records datasets prior to de-identification. Admissions for each disease (Asthma, Bronchiolitis, Bronchitis and COPD) were summarised in separate tables. (All displayed data are hypothetical.) *LOS = length of stay associated with hospital admission.

Record number	Admission date	LOS*	Street number	Street name	Street type	Locality	Postcode	Primary Disease
12345-1	1992-06-05	2	***	***	***	***	7250	ASTHMA
54321-4	2005-11-23	12	***	***	***	***	7249	ASTHMA
98765-1	1994-02-19	1	***	***	***	***	7278	ASTHMA

3.3.1 De-identification of medical data

Initially, a macro was designed that accessed both the medical records and the Address Points Dataset (APD) simultaneously. This macro was developed in Visual Basic by Darren Turner (UTAS) and worked in five stages:

1. An address row was read from the medical records file (e.g. 23 Smith Rd Launceston);

2. This was matched to the corresponding address in the ADP, and the relevant true X,Y coordinates were found;

3. A uniform random number generation with an absolute upper limit of 200 (i.e. +/- 200) was applied to both the X (Easting) and Y (Northing) co-ordinates of the true address location;

4. It was identified that points could potentially be moved more than 200 m (Euclidean distance) from their true location if high values were added to both the Easting and Northing value, as illustrated in Figure 3.5, resulting in a square buffer rather than circular.

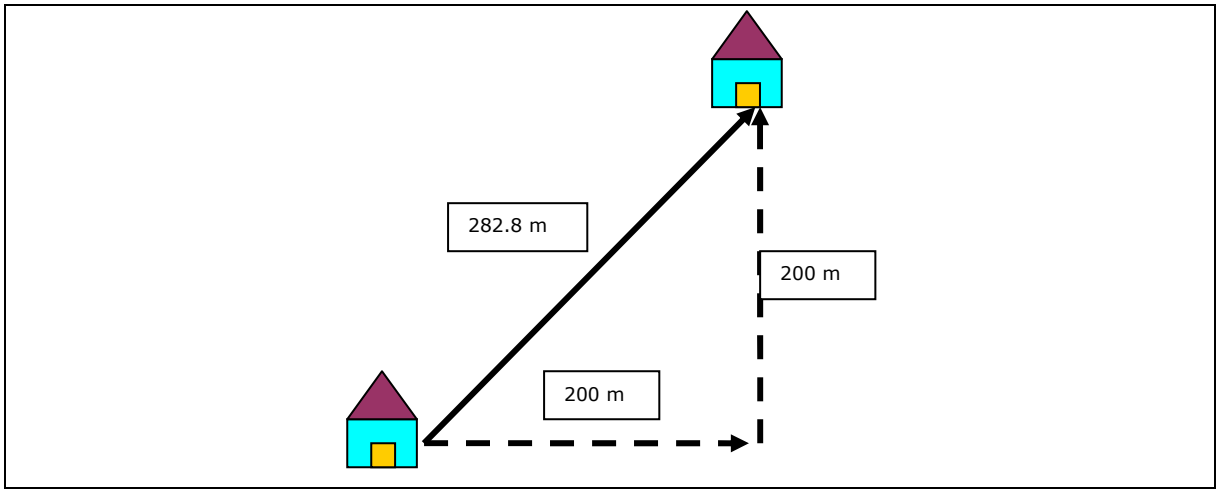


Figure 3.5 - Illustration showing that address points could be moved further than 200 m from their 'true' location (up to 282.8 m) if maximum perturbation distances were added in both the Easting and Northing direction. Basic Pythagorean theory was applied to remedy this problem.

To instead ensure a circular buffer with 200 m maximum radius, Pythagorean theory was applied: $E^2 + N^2 = c^2$; where E is the distance in metres that the true location will shift along the Easting axis, N is the distance in metres that the true location will shift along the Northing axis, and c is total distance shifted (maximum = 200). After randomly generating both E and N the equation was solved for c. An expression was considered 'true' if c <= 200 and 'false' otherwise; if false, step 3 was repeated until true.

5. The scrambled X,Y location was written into the appropriate row of the medical records file and the street address columns were deleted.

At the completion of this 5-step process, medical records could then be accessed which contained a scrambled address location in place of the street address. Unfortunately, this method proved unsatisfactory for three reasons. Firstly, applying a new perturbation to every address in the medical records dataset meant that none of the geographic locations of the resulting disease cases aligned exactly with the locations of 'houses' in the APD. This not only considerably raised the number of data points and therefore the processing speed required for analysis, but also resulted in all of the 'true' address locations being free from disease cases. Secondly, in cases with numerous admissions from the same street address, as was most commonly the case with COPD, the resulting data points were so neatly, randomly arranged around a single address point that the true address location was easily re-identifiable. And thirdly, by randomly moving only the case data and leaving the population data in its true location, it was thought that address locations could potentially be re-identified by analysing street patterns of address points, particularly in urban areas. Figure 3.6 illustrates these issues.

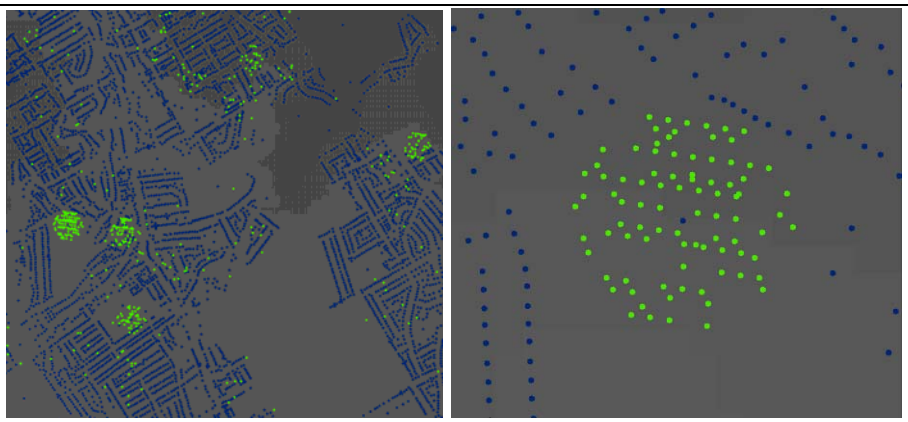


Figure 3.6 - Unsatisfactory results of the first de-identification attempt. Dark blue points are house locations from the Address Points Dataset, green points indicate COPD incidence. For dwellings with numerous admissions, it is clearly seen that 'cases' accumulate and become clustered in a 200 m circular buffer around the true address location. This renders address data potentially re-identifiable. It can also be seen that patterns in the organisation of urban properties make it possible to re-identify the original street address.

In light of the above observations, it was determined that a new method of de-identification was needed that met the following two criteria: disease 'cases' must align exactly with address points; and, patterns of address points must be altered to disallow possible re-identification.

3.3.2 A better method: de-identification of *address* data

A second macro was then designed that first scrambled the points in the *Address Points Dataset* and then assigned these scrambled co-ordinates to the medical records. This macro was developed in Visual Basic by Tony Miller (Eighty Options, Tasmania) and worked in five stages:

1. A uniform random perturbation was performed on all Easting and Northing values in the Address Points dataset (ADP). To once again ensure that the buffer was circular rather than square, and that the circle had a maximum radius of 200 m, Pythagoras' law was applied, though slightly differently this time:

$$N^2 + E^2 = c^2;$$

where N is the amount in metres to be added to the Northing value, E is the upper limit of the amount in metres to be added to the Easting value, and c is the maximum radius of the circular buffer (200). A number between -200 and 200 was first randomly generated to be added to the Northing value. Pythagoras' law was then solved for E, given that c = 200. For example, if N has been generated as -64, then solving for E gives:

$$\begin{aligned} E^2 &= c^2 - N^2 \\ &= 200^2 - (-64)^2 \\ &= 40000 - 4096 \end{aligned}$$

= 35904

E = +/- 189.5

This value for E is an absolute value that provides the range within which the second number will be randomly generated. That is, a number between -189.5 and 189.5 is then randomly generated to be added to the Easting value of that true address point in the APD.

This step was applied to every address location in the APD such that all address locations were then randomly shifted within a 200 m radius, in any direction, form their true location.

2. A street address was read from the medical records file;

3. This street address from medical records was matched with a street address in the APD;

4. The newly generated (scrambled) X,Y location corresponding to that address was written into the medical records file (just as a look-up table). In this way, multiple admissions from the same street address were all assigned the same X,Y coordinates.

5. The primary disease name (i.e. ASTHMA, BRONCHIOLITIS, BRONCHITIS or COPD) was read from each of the medical records files, and for every 'case' of every disease registered at that address, a number was added incrementally to the corresponding column in the APD.

The resulting format of the Address Points Dataset and medical records is shown in Table 3-3 and Table 3-4, respectively. Figure 3.7 demonstrated that address points were successfully 'scrambled' a random direction and distance up to 200 m from their original location.

Figure 3.8 and Figure 3.9 show the spatial outcome of this de-identification process, which proved a far superior method to the first attempt.

Table 3-3 - General format of the Address Points Dataset (APD) after random perturbation with the second macro. Street addresses have been removed and true Easting and Northing values have been replaced by the results of the random perturbation. Incremental counts of each disease have been generated (AS = Asthma, BL = Bronchiolitis, BR = Bronchitis, COPD = Chronic Obstructive Pulmonary Disease).

Geo-ID	Geo type	Adjusted Easting	Adjusted Northing	Postcode	AS	BL	BR	COPD
904232	urban	511214	5414903	7250	2	0	1	0
1014089	rural	510164	5416842	7277	0	0	0	0
950296	urban	461709	5446116	7253	1	0	0	1

Table 3-4 - General format of the medical records datasets after de-identification with the second macro. Street addresses have been replaced with the adjusted Easting and Northing values from the APD. (All data are hypothetical.) *LOS = length of stay associated with hospital admission.

Record number	Admission date	LOS	Adjusted Easting	Adjusted Northing	Postcode
12345-1	1992-06-05	2	502175	5529683	7250
54321-4	2005-11-23	12	492758	5418294	7249
98765-1	1994-02-19	1	512734	5467325	7278

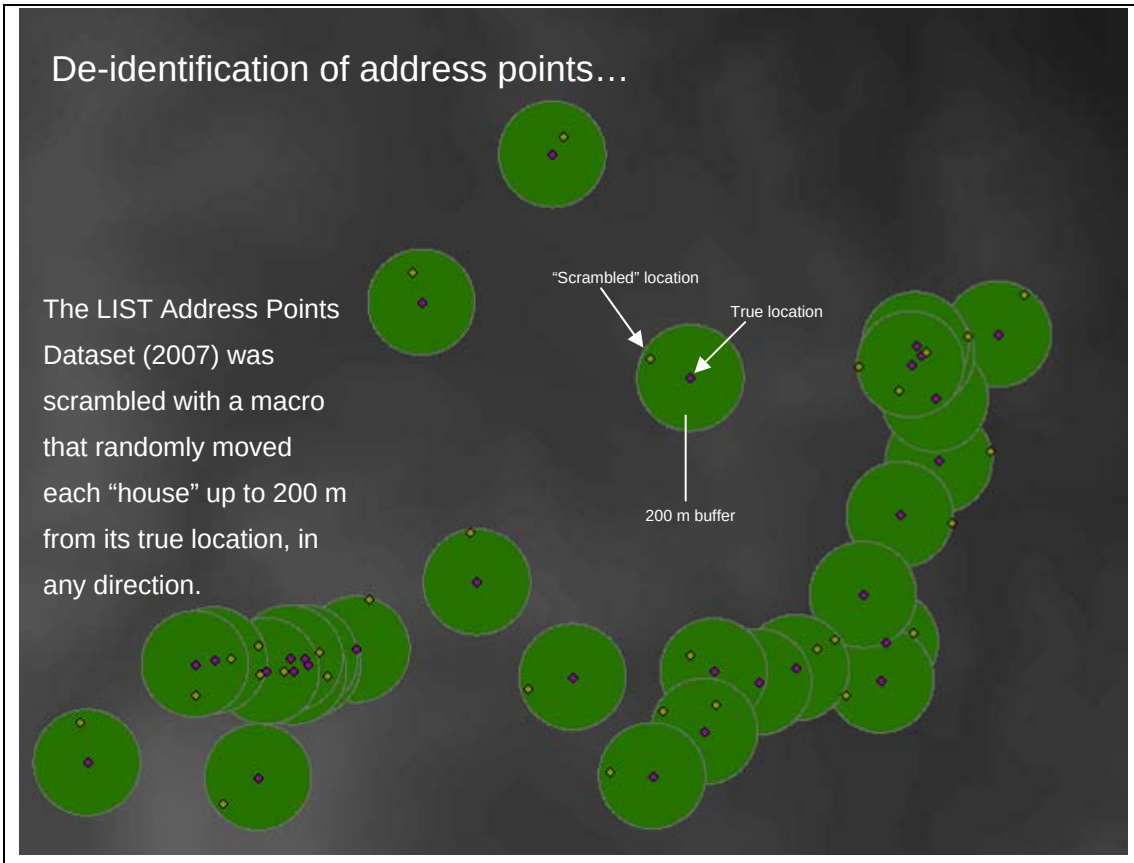


Figure 3.7 – Successful de-identification of address points using the second macro. This diagram was developed using sample address data to check to accuracy of the macro prior to implementation.

Figure 3.8 and Figure 3.9 then show the Address Points Dataset at two spatial scales before and after random perturbation. This de-identification process proved a far superior method to the first attempt. A comparison of Figure 3.8 and Figure 3.9 shows that the random perturbation of address points resulted in addresses at the neighbourhood scale becoming unrecognisable, though at a coarser resolution (i.e. the whole study area) there is very little change. This was precisely the outcome that was hoped for; spatial integrity of the dataset was not compromised at the coarse resolution of the entire study area, and geoprivacy was maintained at the neighbourhood scale.

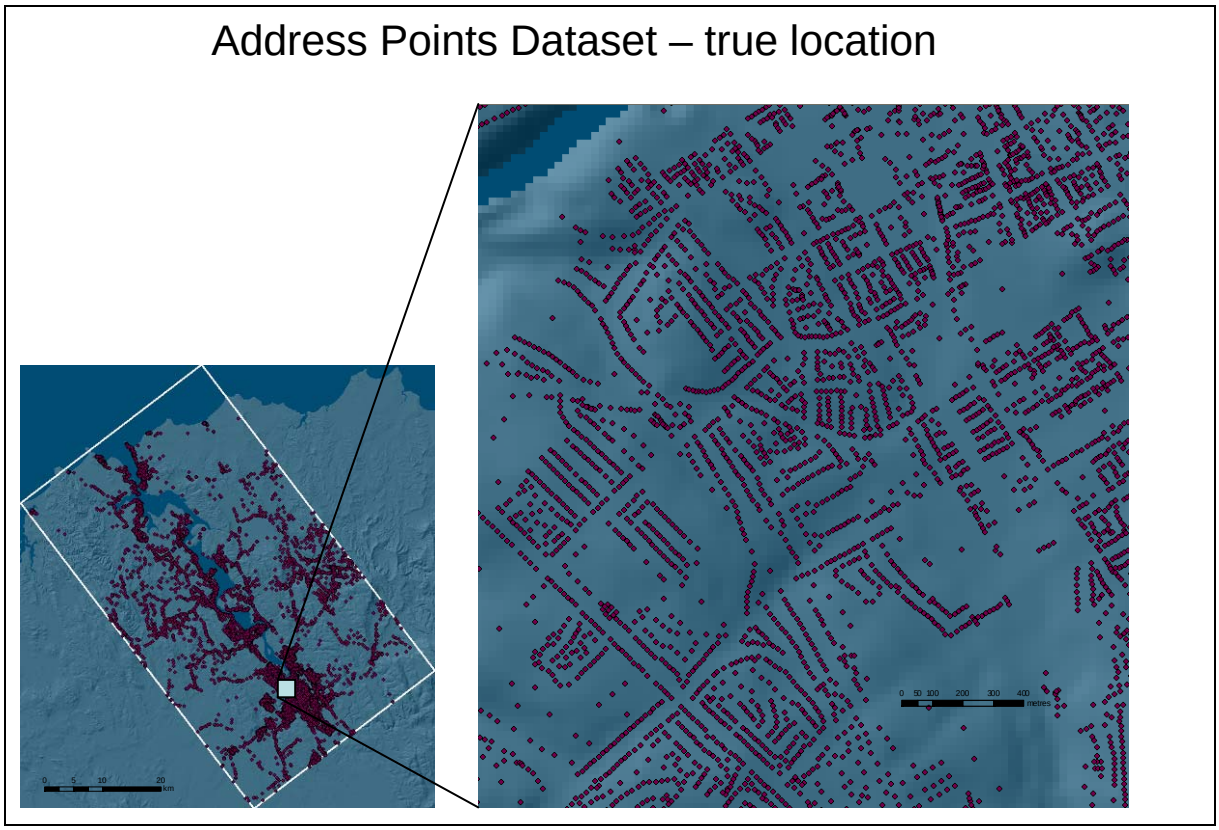


Figure 3.8 - The Address Points dataset at two scales in its original form

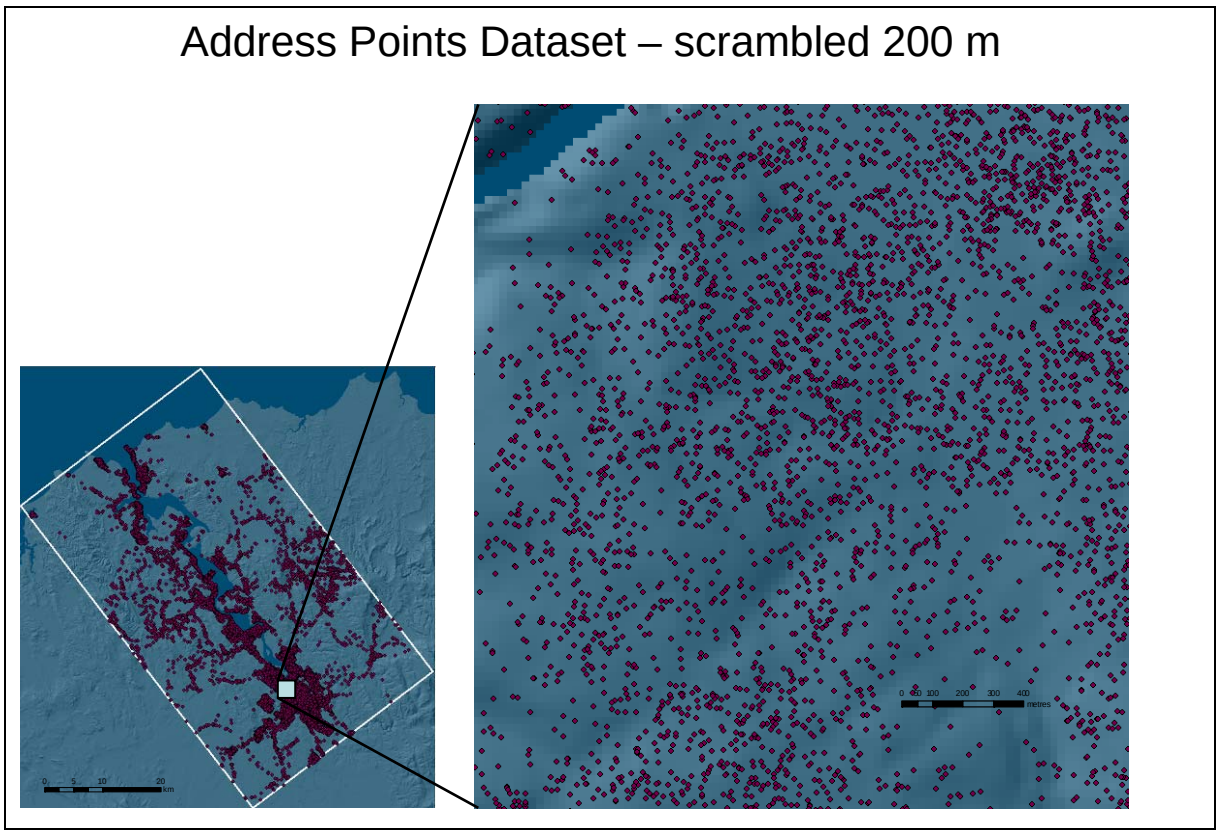


Figure 3.9 - The APD at two scales after random perturbation *(Source: Land Information Services Tasmania (theLIST))*

3.3.3 Ethical reporting of spatial public health data

It must be acknowledged that while there is clear value in analysing health data spatially, the interpretation and presentation of results requires great care. Inappropriate reporting of disease clusters can cause unnecessary alarm and distress, particularly among the at risk population (Fefferman et al., 2005; Sabel and Loytonen, 2004), and this must be considered along with any presumed benefit of identifying public health concerns.

The publication of spatial health data holds real concerns about geoprivacy and confidentiality, yet no standards currently exist for the mapping of health data in point form. Studies have shown that it is sometimes possible to 're-engineer' modified address data from published maps to re-identify some of the residences/residents, even if background cartographic elements have been removed (eg roads) (Curtis et al., 2006).

The U.S. Department of Health and Human Services (HHS) has produced guidelines for the publication (mapping) of health data in aggregate form. They state that maps should not be produced at the residential level for areas with a population of less than 20,000 (Curtis et al., 2006); this is somewhat open to interpretation. It is noted that there is no mention of the sparsity of address points in rural vs. urban settings, nor the scale at which maps can be published. It is also not clear whether this guideline refers to *aggregation units* of 20,000+, or to the total population of the study area. In the absence of clear guidelines for ethical reporting, a combination of standards for aggregate data and common sense should be applied.

3.3.4 Further masking of rural addresses

As mentioned above, the considerable distances between many rural properties in the study area raised some concerns about the risk of possible re-identification of these addresses post-publication. As seen by comparing the coarse resolution images on the left of FiguresFigure 3.8 andFigure 3.9, there is little change in the relative position of some rural address points after the 200 m perturbation as some addresses are considerably more than 200 m apart. Other researchers recommend dealing with this problem by removing sparse data points, aggregating point data into a generalised surface, or refraining from publishing maps altogether (Curtis et al., 2006).

To further ensure geoprivacy of rural residents in this study, all address points that were further than 200 m from their nearest neighbour were removed from the dataset prior to map production. Some points with only one or two close neighbours, that were thought to be possibly re-identifiable, were also deleted. This was achieved by first creating a 200 m buffer around each address point and selecting all those with no

overlapping buffers (i.e. no neighbours within 200 m). The dataset was then visually checked for any other address points that were considered potentially recognisable at a scale of 1:400,000 (the scale at which the majority of maps were produced for this study). Remote points were thus included in the *analyses* of this study, but excluded from the published maps.

4 Disease mapping

4.1 Research Hypothesis

- "Spatial and/or temporal clusters of respiratory disease are evident within the Tamar Valley"
- H_{null}: "There is no clustering of respiratory disease in the Tamar Valley at any spatial or temporal scale." i.e. "There is complete spatial and temporal randomness."

4.2 Introduction

This chapter details the analysis of de-identified patient address data for the detection of areas of elevated risk of respiratory disease in the Tamar Valley. Three methods of cluster detection are applied and explained in detail.

4.2.1 Generalisation of point patterns into 'hot-spots' or 'clusters'

In its simplest form, point data may be observed 'as is' for a rough look at disease occurrences across a study area. It is useful then to generalise point data of disease occurrence into areas of elevated or reduced risk of disease. 'Cluster detection' is a common term in the literature referring most often to the detection of areas of elevated disease risk. It has been suggested that the term 'spatial variation in risk' is more appropriate than 'cluster' in studies of respiratory disease as it is argued that the latter implies a genetic or infectious component to the disease under study (and therefore non-independence of observations) (Diggle, 2000; Sabel and Loytonen, 2004). However, 'cluster' has a relatively generic meaning across the majority of literature and so both terms are used concurrently here.

4.2.2 The value of disease mapping

Throughout its long history, disease mapping has proven an invaluable tool in studies looking to answer a fundamental question; "where is disease most prevalent?" By effectively answering this question, further questions of social/environmental justice and disease aetiology can be investigated.

On a global scale, disease mapping by the World Health Organization has been singularly instrumental in detecting the presence and spread of communicable diseases such as HIV AIDS, H5N1 avian influenza (bird flu) and severe acute respiratory syndrome (SARS). This initial knowledge of disease geography allows for the identification of at-risk populations and the development of highly targeted, cost-effective intervention strategies (WHO, 2007).

Sabel and Loytonen (2004) pose a series of generic questions that could be answered with a cluster analysis, including:

- Is the observed clustering a result of natural background variation? (i.e. is the background population itself spatially clustered?)
- At what scale does clustering occur?
- Are clusters a result of some existing heterogeneity in the study area (i.e. autocorrelation)?
- Are clusters associated with proximity to roads or other pollution sources?
- Are spatial clusters also aggregated in time?

These research questions guided much of the current study.

4.2.3 Cluster detection techniques

Identifying the presence and/or location of disease clusters in a study area are central components of many spatial epidemiological and pre-epidemiological studies (Kelsall and Diggle, 1998; Maheswaran and Craglia, 2004; Sabel and Loytonen, 2004). Cluster detection techniques can be global or local, focused or general and use aggregate or individual-level data (Elliott et al., 2000a; Fotheringham et al., 2002b; Maheswaran and Craglia, 2004).

Global cluster detection determines the presence/absence of spatial clustering in an entire dataset and can provide useful initial information about the study area. Non-spatial epidemiological studies can be thought of as using global test statistics, as the presence and magnitude of disease rates (and/or relationship to environmental factors) are considered for the entire population as a whole (Fotheringham et al., 2002a). Examples of spatial global cluster detection techniques include the K function, Pearson's chi-squared statistic, kernel intensity function ratio summary, and Cuzick and Edward’s method (Cuzick and Edwards, 1990). Global statistics designed to detect spatial similarity (autocorrelation) include Moran's I and Geary's c (Sabel and Loytonen, 2004).

Local clusters then can be defined as specific areas within the dataset where elevated or reduced rates of disease occur; this is the particular domain of spatial data analysis (Fotheringham et al., 2002a). Local clustering techniques include: kernel intensity function; Getis Ord statistics, and Kulldorff’s spatial scan statistic (Wheeler, 2007).

Moving window methods are a type of local test able to detect the spatial location of excess cases. The search ‘window’ can be defined in a number of ways (e.g. circle, grid cells, annulus etc) and these methods generally assess where the "number of cases within a window exceeds that expected by chance" (Sabel and Loytonen, 2004). Examples include Openshaw's method, Besag and Newell's method, Kulldorff’s scan

statistic, and Cuzic and Edward's method. Moving window methods are particularly recommended for sparse data (Wartenberg and Greenberg, 1993, in: (Sabel and Loytonen, 2004).

Tests for focused clusters concentrate on a particular known pollution source and look for proximal disease clusters. General, or unfocused, tests on the other hand assume nothing of relative proximity to pollution sources and test the whole population for local areas of elevated disease risk (Kulldorff and Nagarwalla, 1995).

Global clustering of respiratory disease has already been observed and reported for Launceston and the Tamar Valley as a whole (i.e. (Giles, 1980; Lyons and expert working party, 1996; Mesaros et al., 2007)). This current study therefore employed three local unfocused cluster detection techniques to investigate the spatial variability in disease rates at the micro-scale: Kernel Density Estimation, the Getis Ord Gi* statistic, and Kulldorff's spatial scan statistic, implemented in SaTScan. Each is explained in detail below.

4.3 Methods

4.3.1 Software requirements

ESRI software ArcGIS version 9.1 was used for the majority of spatial data analysis and map generation in this section of the study.

SaTScan v 7.0.3 was used to implement Kulldorff's spatial scan statistic for cluster detection. SaTScan is a public access software package developed by Kulldorff (2006) and was designed explicitly for the purpose of detecting local disease clusters within a population, though it is applicable to a broad range of research fields. SaTScan incorporates Poisson, Ordinal, Bernoulli, Exponential, Normal and Space-Time Permutation statistical models suitable for various research questions and data types. Data can be searched for purely spatial, purely temporal or spatio-temporal clusters of disease.

4.3.2 Preliminary disease analysis

Disease data were first analysed non-spatially to observe global temporal trends in disease occurrence. Various combinations of Asthma, Bronchiolitis, Bronchitis and COPD admissions numbers were analysed and graphed for comparison. Throughout the study, disease 'houses' were used to represent address points where one or more disease cases was recorded; disease 'cases' refers to total hospital admissions and can include multiple admissions from the same address.

4.3.3 Kernel Density Estimation

Kernel Density Estimation (KDE) was used in this investigation primarily as an exploratory data analysis tool. KDE can be conceptualised as the application of a moving 'window' that systematically scans a spatial dataset to determine varying density across the study area. Kernel *Intensity* Estimation refers to the same technique used to analyses the distribution of *attributes* within the data (e.g. rather than the number of houses occurring in a given area, the number of house occupants in a given area) (Waller and Gotway, 2004; Wheeler, 2007). Kernel Intensity Estimation is a more correct definition for its use in the current study, though 'Kernel Density Estimation' is commonly used to describe both functions in the literature.

The output of KDE is a unit-less measure of density that corresponds to the function of the mathematical curve created by the kernel window passing over the study area; this is illustrated in Figure 4.1. The integral of x is always equal to 1, so as the radius of the search window increases, the maximum value of $f(x)$ decreases. The optimal search window can either be determined intuitively or through a mathematical investigation, mostly incorporating the Mean Integrated Squared Error (MISE) which is a measure of agreement between the kernel density estimate and the true probability density estimate (Fotheringham et al., 2002b). As KDE was used here largely for exploratory purposes, intuitive selection of the kernel window was thought to be adequate.

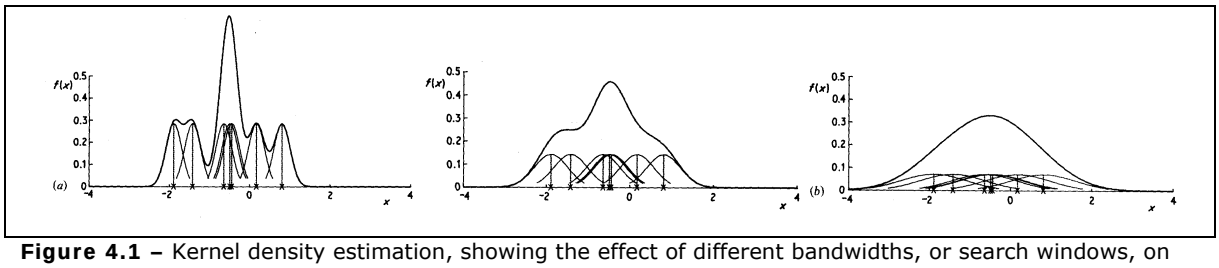


Figure 4.1 – Kernel density estimation, showing the effect of different bandwidths, or search windows, on both the 'smoothness' of the created density surface and the magnitude of reported density estimates ($f(x)$). Bandwidth increases from left to right in the diagram. (Adapted from (Silverman, 1986))

KDE was applied using the ArcGIS Spatial Analyst extension, with the analysis environment set to the spatial extent of the study area. It is usual to produce a map layer of expected disease incidence and then compare this with the observed disease layer to compare the differences (Maheswaran and Craglia, 2004). The function was therefore first applied to the population as a whole – using the created population 'per residence' attribute of the APD – to illustrate the spatial pattern that would be expected if disease occurrences were evenly distributed across the population.

KDE was then applied to data from each disease individually (i.e. the point locations of Asthma, Bronchiolitis, Bronchitis and COPD accumulated over the 1992-2006 study period), and to all diseases collectively. Data were grouped over the entire 15 year

study period for initial investigations. Temporal variation in disease patterns where then explored using the Asthma dataset for each year, and each season, to test for robustness of spatial patterns over time.

4.3.4 Getis Ord Gi* statistic

The Getis Ord Gi* statistic with z-score rendering was applied using the ArcGIS Spatial Statistics extension. Separate disease cluster analyses were conducted for each of Asthma, Bronchiolitis, Bronchitis and COPD using cumulative data for the entire study period (1992-2006). Combined disease incidence for Asthma, Bronchiolitis and Bronchitis was also tested for. A 1 km fixed distance search radius was chosen intuitively as the distance that would adequately account for the shift introduced by the 200 m random perturbation of address points.

The Getis Ord Gi* statistic is similar to KDE in that it scans the dataset with a moving window of fixed size (in this case) to detect density variations; it differs from KDE on two important counts. Firstly, rather than scanning the entire dataset indiscriminately, each point is visited individually and the density/intensity of observations is calculated within a defined radial distance (of each address point). Secondly, a measure of density inside each search window is calculated relative to that expected outside the window (i.e. the rest of the population). In this way, the Gi* statistic is computed that compares local averages with a global average and thus identifies 'hot spots' (Getis and Ord, 1996). A Gi* statistic is calculated for each point in the dataset and corresponds to the density of observations within the search radius around that point (including the point itself) (Getis and Ord, 1996). Thus, address points with no disease cases may be assigned a high Gi* score if neighbouring points (within the search radius) have a high disease incidence.

The output of the Getis Ord Gi* statistic is a point dataset that is a replicate of the original, with a Gi* score attached to each point. The dataset can then be displayed in graduated colour to highlight the variation in disease density. Rather than publish individual point data, however, z-score rendering was first applied. Z-score rendering applies a z-score to each point, thereby identifying clusters that are various standard deviations from the mean of the study area (i.e. Z-scores of +1 and +2 relate to clusters that are 1 and 2 standard deviations above the mean respectively). Clusters of just one or two isolated points were removed from the dataset prior to map production to aid geoprivacy.

4.3.5 Kulldorff's spatial scan statistic

SaTScan is a free software tool developed by Kulldorff (2006) that can be used to search for purely spatial, purely temporal or spatiotemporal clusters. Kulldorff's spatial

scan statistic incorporated in SaTScan extends the work of Openshaw *et al* (1987) and Cuzick and Edwards (1990) to incorporate an unfocused (general), local cluster detection method. Like the Gi* statistic, a search window is centred on each point in the dataset. Here though, the search window can be either circular or elliptic and, critically, of ever increasing radius from zero until up to 50% of the population is incorporated (Kulldorff and Nagarwalla, 1995). In this way, the maximum search radius is not a fixed Euclidian distance but rather is variably dependent on the spatial distribution of the background population. The optimal radius is determined separately for each cluster based on the window size (and orientation if elliptic) that contains the highest number of cases per population.

The 'ellipses' used in SaTScan vary not only in size but also in shape and orientation. A circle and five ellipses are used, with shapes defined by the ratio of the long to short axis (ratios are 1.5, 2, 3, 4 and 5); each ellipse shape is then rotated through a different number of spatial orientations (corresponding to 4, 6, 9, 12 and 15 for each ellipse shape, respectively). The north-south orientation is always included and the other orientations are evenly spaced around the compass (Kulldorff, 2006). Within each window, observed/expected disease frequency is then calculated for *each cluster of points*, rather than each point. This process therefore assigns the same value to every point within a shared cluster, which can include both cases and non-cases. The statistic calculates the expected disease intensity within each possible window variant, based on the population size contained within the cluster through Monte Carlo simulations of (1000 in this case) random subsets of the remaining population. Thus, the statistical significance of each possible cluster is calculated; the *cluster* with the lowest p-value for that *point* is then reported in the output. For large datasets, this is clearly a highly computationally demanding process. For example, the full dataset used in this study consisted of around 43,000 points, which required 2 Gb of memory and took over 8 days to run on a Linux machine, using two processors.

SaTScan was first programmed to analyse clustering of the Asthma dataset in a spatial subset of the George Town area (as KDE and the Gi* statistic both showed an elevated incidence of Asthma in George Town). The dataset was scanned for purely spatial clusters of elevated disease rates under a Poisson probability model. Elliptical search windows were selected to incorporate up to 50% of the population. Individual address points were treated as having a population of one, thus this was a test for clustering of 'Asthma houses' (houses where Asthma occurred over the study period 1992-2006).

SaTScan was then run on a dual processor machine to analyse Asthma clustering in the entire dataset (a powerful computer was required to manage the entire dataset at once). The same parameters were used as for the George Town analysis, though the

population 'per residence' field was used as a measure of population. This analysis of the entire dataset was therefore a test for clustering within the *population*, rather than a test for clustering of Asthma houses.

The resulting cluster maps were edited prior to publication to remove address points that were further than 200 m from their nearest neighbour or were otherwise thought to be possibly re-identifiable.

4.4 Results

4.4.1 Preliminary disease analysis

Graphs of general disease rates are shown below. Figure 4.2 shows the proportional hospital admission rates for Asthma, Bronchiolitis, Bronchitis and COPD for each year in the study period. Figure 4.3 shows a comparison between Asthma and COPD admissions, and Figure 4.4 shows total admissions for Asthma, Bronchiolitis and Bronchitis for every year in the study period. It is seen in these graphs that COPD admission rates began to rise sharply in 1999. This is probably due to the changeover from ICD-9 to ICD-10 discharge codes between 1998 and 2000, which brought some transfer of diagnoses from chronic Asthma to COPD (Wisconsin Department of Health and Family Services, 2004). However, Figure 4.3 shows that this increase is not entirely explained by the ICD-9 to ICD-10 changeover, as the increase in COPD was higher than the decrease in Asthma. Figure 4.4 shows the annual fluctuations in disease admission rates for Asthma, Bronchiolitis and Bronchitis, where it is seen that Asthma rates alone dropped markedly in 2004 (which may relate again to the sharp peak in COPD admissions in this year). Global statistics for the study area are summarised in
Table 4-1.

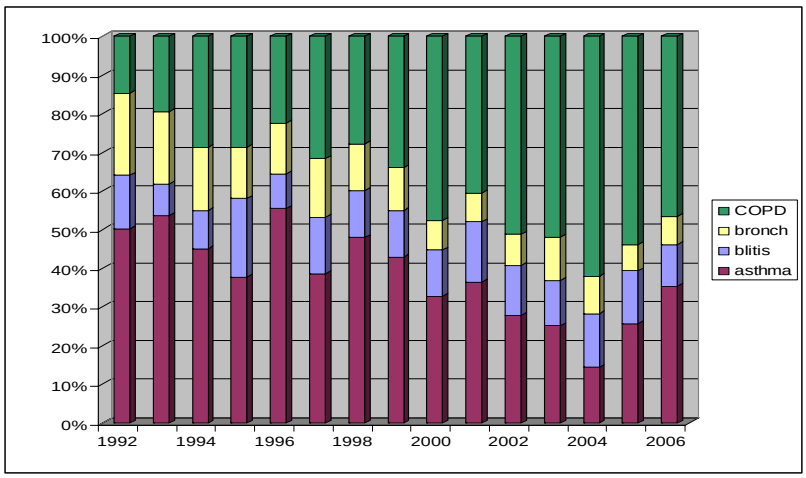


Figure 4.2 - Proportional hospital admission rates for Asthma, Bronchiolitis, Bronchitis and COPD for each year of the study period.

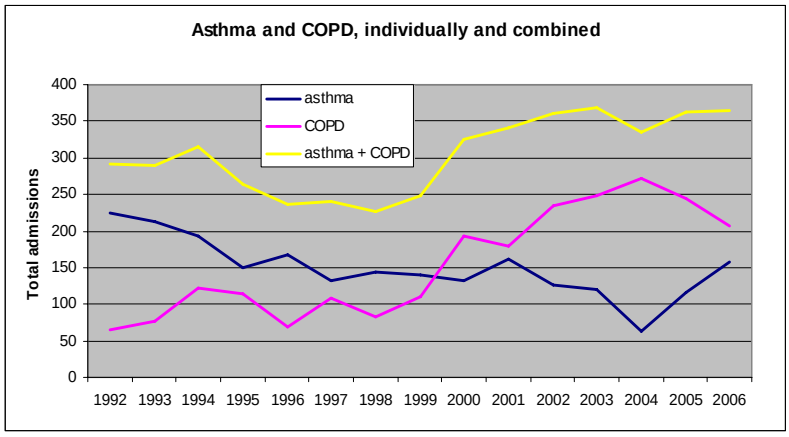


Figure 4.3 - Total Asthma (blue) and COPD (pink) admissions plotted individually and combined (yellow) for every year in the study period. The rise in COPD in 1999 is not entirely explained by the change from ICD-9 to ICD-10 in this year, as the increase in COPD in this year is greater than the decrease in Asthma.

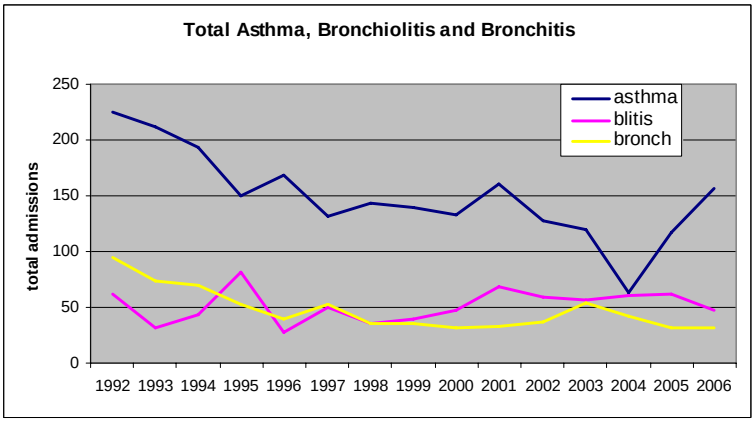


Figure 4.4 – Total admissions for Asthma (blue), Bronchiolitis (pink) and Bronchitis (yellow) for every year in the study period

Table 4-1 – Global statistics of disease admissions for the entire study area. All disease numbers reflect 15 years of accumulated hospital admissions. (*ABB = combined Asthma, Bronchiolitis and Bronchitis.)

	Address points	Total population	Asthma	Bronchiolitis	Bronchitis	COPD	Total disease	ABB*
	42997	94985	2240	773	717	2337	6067	3730
% of total population	---	100	2.36	0.81	0.75	2.46	6.39	3.92

4.4.2 Kernel Density/Intensity Estimation

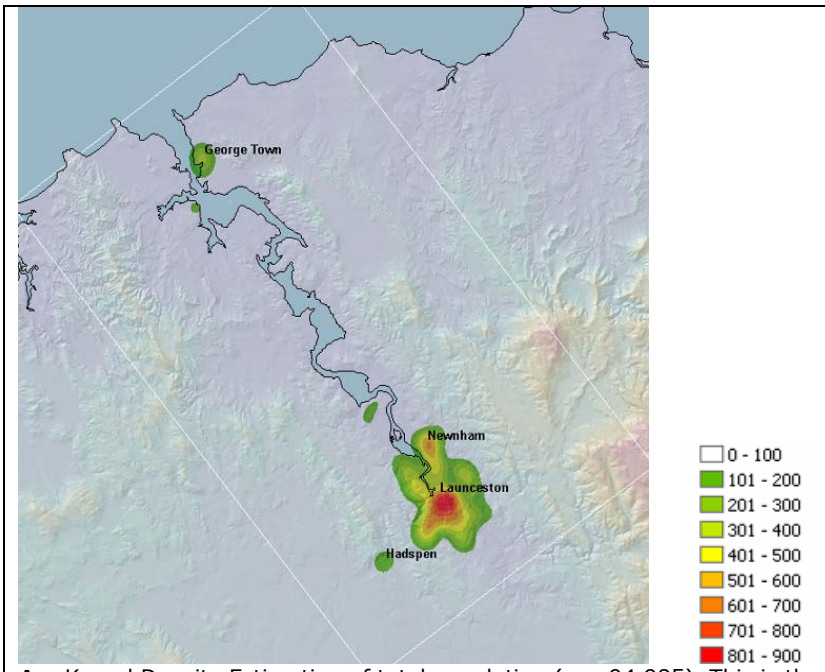
A 2 km kernel window (search radius) was found to neither over-smooth nor create 'spikes' around data points, which was the desired outcome (Fotheringham et al., 2002b). However, this bandwidth also meant that only the larger population centres around Launceston, George Town and Hadspen were detected. The spatial pattern of the underlying population of the study area is shown in Figure 4.5 (A) where it can be seen that population density is highest in southern Launceston. Smaller population densities are seen in Newnham north east of Launceston, George Town to the north of

the valley and Hadspen to the south west. This is the spatial density distribution that would be expected if disease occurrences were evenly distributed across the population. Figure 4.5 (B-D) shows that Asthma, Bronchiolitis and Bronchitis exhibit spatial patterns quite different from the total population, which suggests that disease occurrence may not be spatially random. Again, KDE produces a unit-less measure of density, so all legends are relative only (see Figure 4.1 and section 4.3.3).

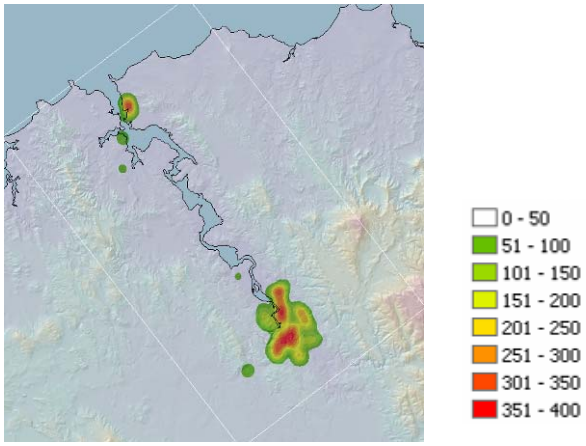
Figure 4.5 (E) shows that COPD has a spatial pattern that is highly clustered in southern Launceston. Both spatial and non-spatial observations of COPD incidence showed that this disease was occurring in large numbers at a handful of addresses across the Tamar Valley (up to 137 cases from a single address over the 15 year period); these addresses were most likely nursing homes and aged care facilities. As such, it was believed that the spatial occurrence of COPD was largely biased by the location of these facilities, rather than influenced by any geographic or climatic processes. It was therefore decided to remove COPD from subsequent analyses of 'total disease', which from here on includes only Asthma, Bronchiolitis and Bronchitis, abbreviated to 'ABB'.

To test the robustness of disease clusters over time, combined Asthma, Bronchiolitis and Bronchitis (ABB) incidence was analysed annually (1992-2006). Results are displayed in Figure 4.6 where it is seen that there is some variation in total disease pattern across the study years. Given the higher rate of Asthma compared to both Bronchiolitis and Bronchitis (n = 2440 vs. 773 and 717, respectively), these results are heavily influenced by Asthma fluctuations. Disease rates generally declined over the study period. Most notably, disease rates in George Town were very high in the first years of the study period (1992-1995). 2006 showed one extreme cluster of disease northeast of Launceston and closer inspection of the dataset revealed that this was due to 47 admissions from just 3 addresses.

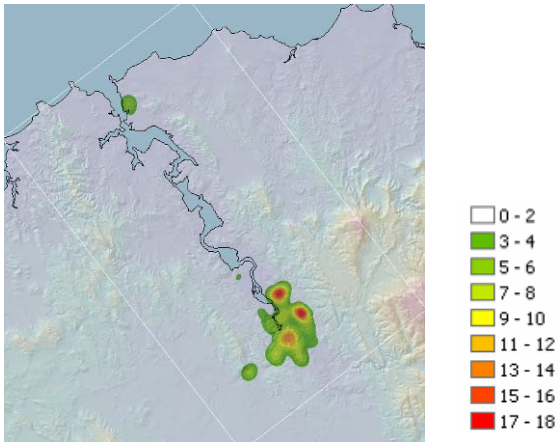
Seasonal variation in total disease was then investigated. Results are shown in Figure 4.7, where it can be seen that seasonal variations are not as pronounced as annual fluctuations. Note, however, that the magnitude of disease occurrence fluctuates across the seasons, with winter displaying the highest disease rate of all seasons.



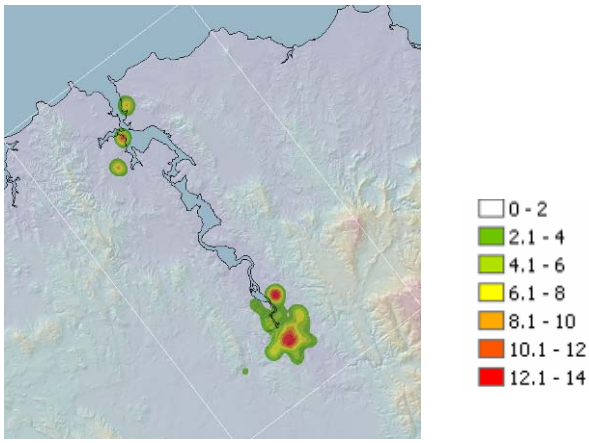
A – Kernel Density Estimation of total population (n = 94,985). This is the general density pattern that would be expected of each disease, if disease occurrences were distributed evenly across the population.



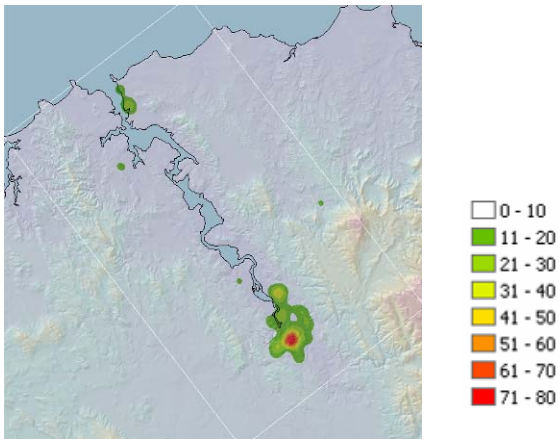
B – Asthma (n = 2240)



C – Bronchiolitis (n = 773)

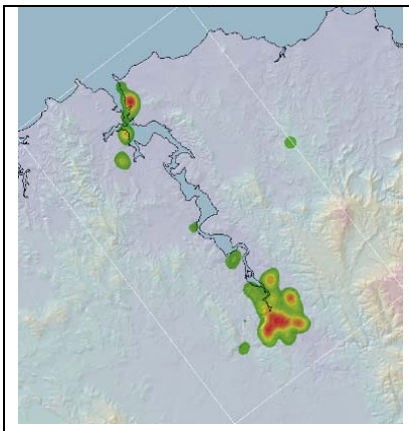


D – Bronchitis (n = 717)

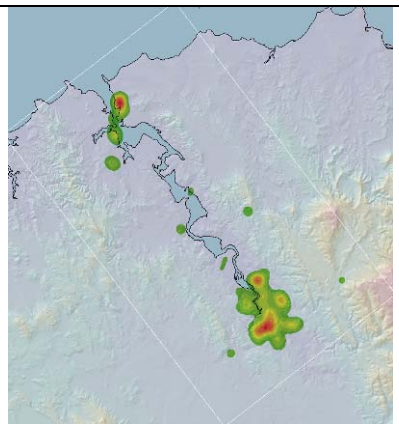


E – COPD (n = 2327)

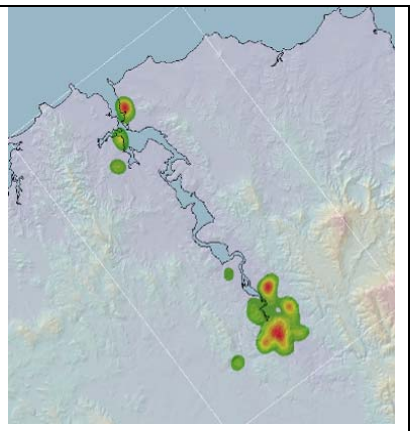
Figure 4.5 – Kernel Intensity Estimation of Total Population (A) and Total Asthma (B), Bronchiolitis (C), Bronchitis (D) and COPD (E) for the 1992-2006 study period. It is seen that disease patterns (B – E) are different from the spread of total population (A), indicating that disease is not occurring uniformly across the population. All legends are relative only. A 2 km search radius was used.



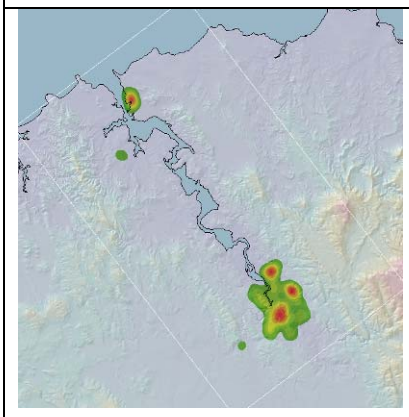
1992 (n = 382)



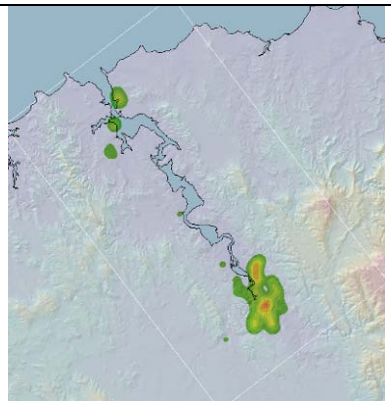
1993 (n = 318)



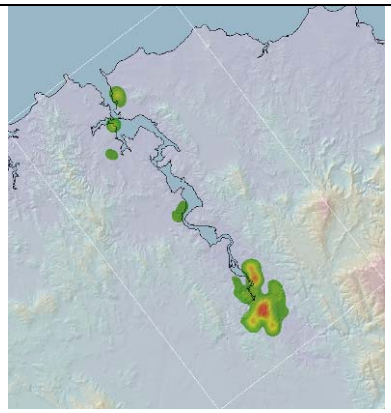
1994 (n = 306)



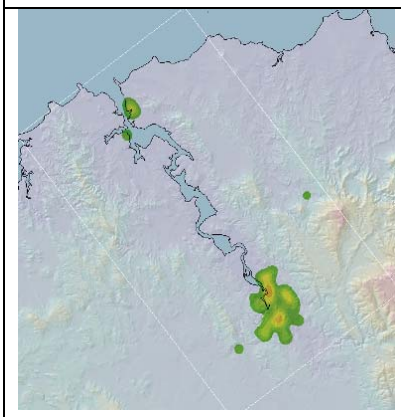
1995 (n = 284)



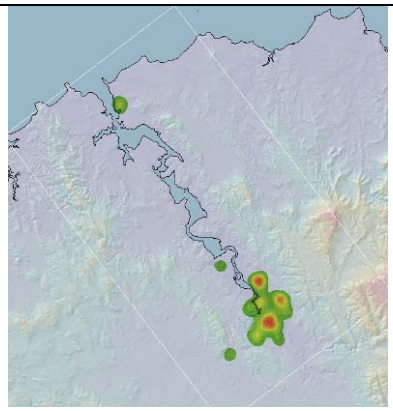
1996 (n = 235)



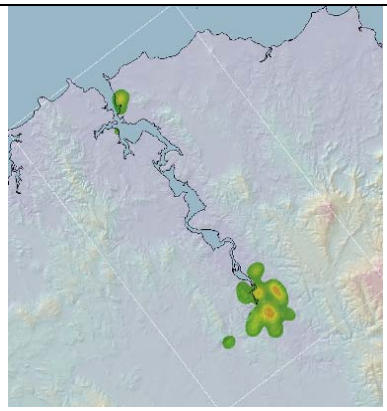
1997 (n = 234)



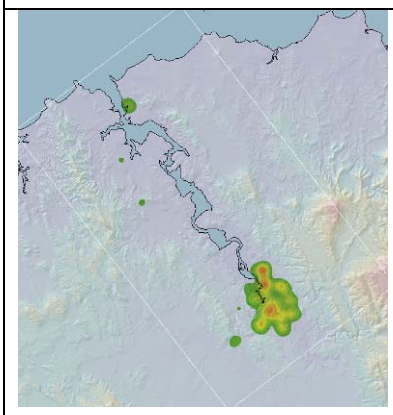
1998 (n = 215)



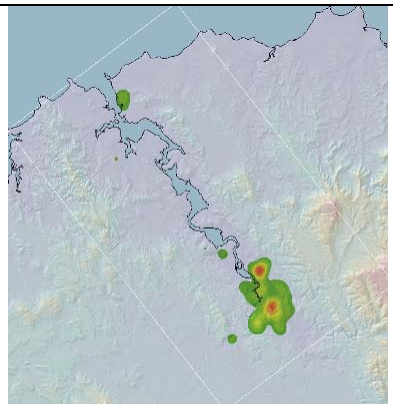
1999 (n = 214)



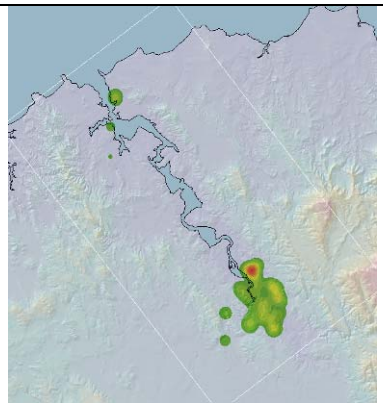
2000 (n = 212)



2001 (n = 263)



2002 (n = 223)



2003 (n = 230)

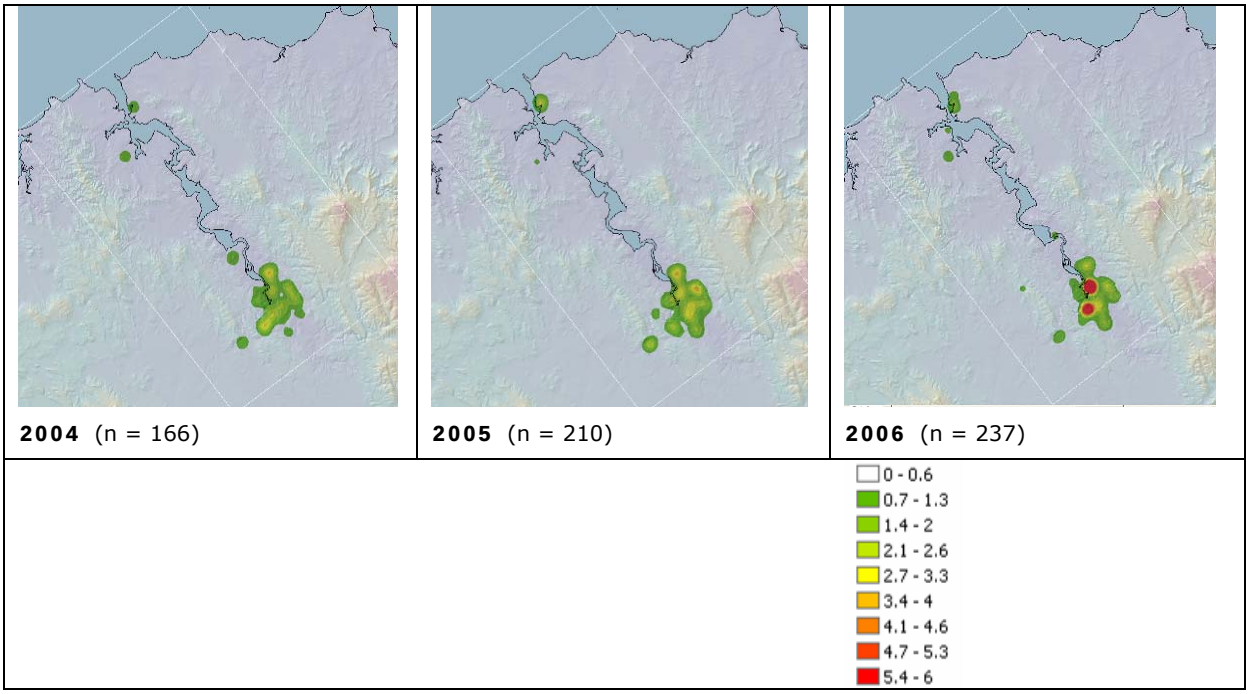


Figure 4.6 – Kernel Intensity Estimation showing annual variation in the spatial pattern of ABB (Asthma, Bronchiolitis and Bronchitis) for every year of the study period.

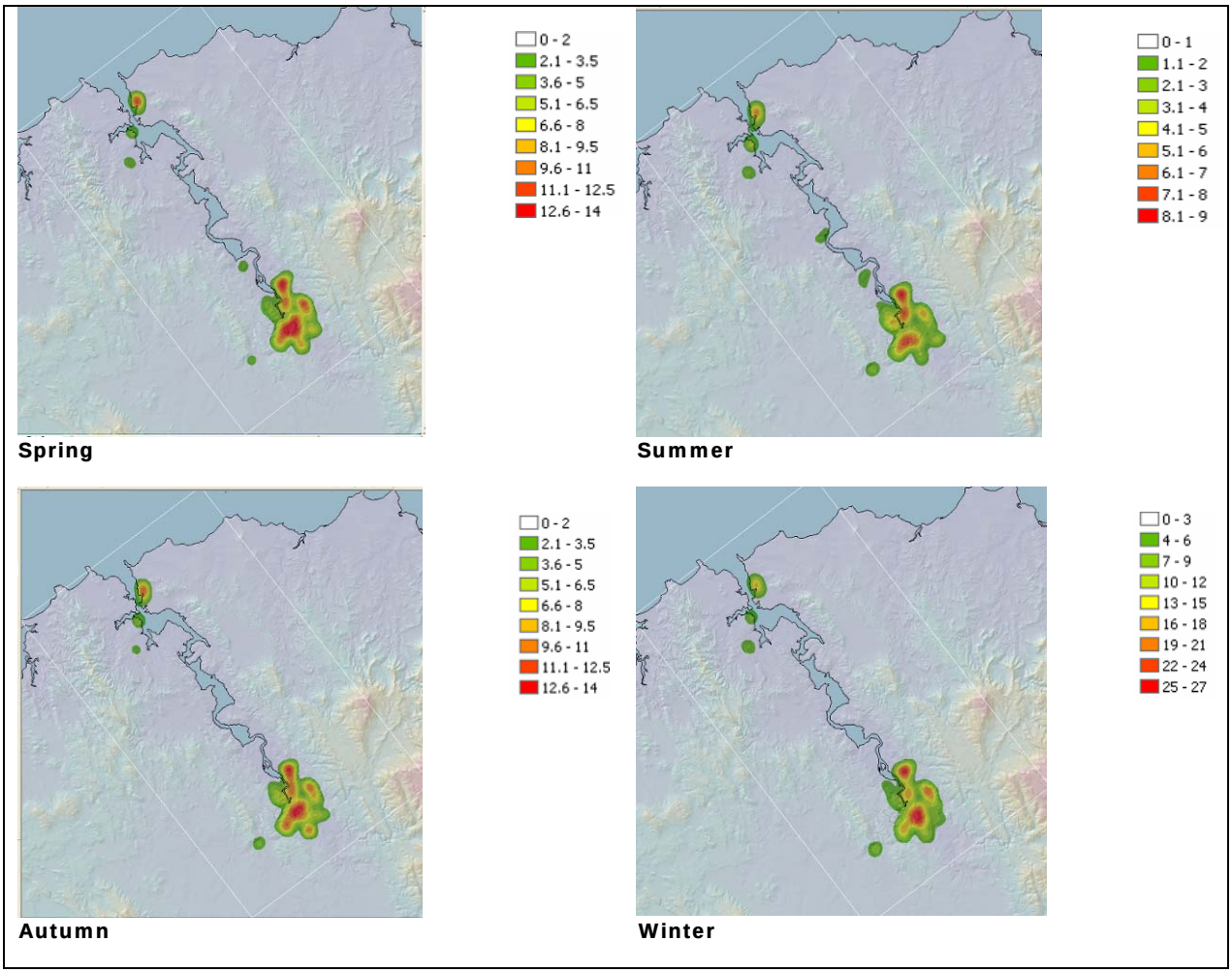


Figure 4.7 – KDE showing seasonal variation in disease occurrence for combined Asthma, Bronchiolitis and Bronchitis for each combined season of the study period 1992-2006. Note the different legend for each season.

4.4.3 Getis Ord Gi* Statistic

Results of the Getis Ord Gi* statistic, with displayed z-score rendering, are shown in Figure 4.8 below. Areas coloured beige and red correspond to points that are 1 and 2 standard deviations above the mean, respectively. Here it can be seen that Asthma is most prevalent around George Town and eastern and northern Launceston; Bronchiolitis predominates around Hadspen and northern and eastern Launceston; and, Bronchitis is quite prevalent in the northern sections of the study area around George Town, Beauty Point and Beaconsfield. COPD is seen to be highly clustered in areas around southern Launceston, Legana, Lilydale and Low Head to the north of George Town. As mentioned, these areas correspond to the location of aged care facilities, and so were not considered indicative of any geographical or climatic correlate.

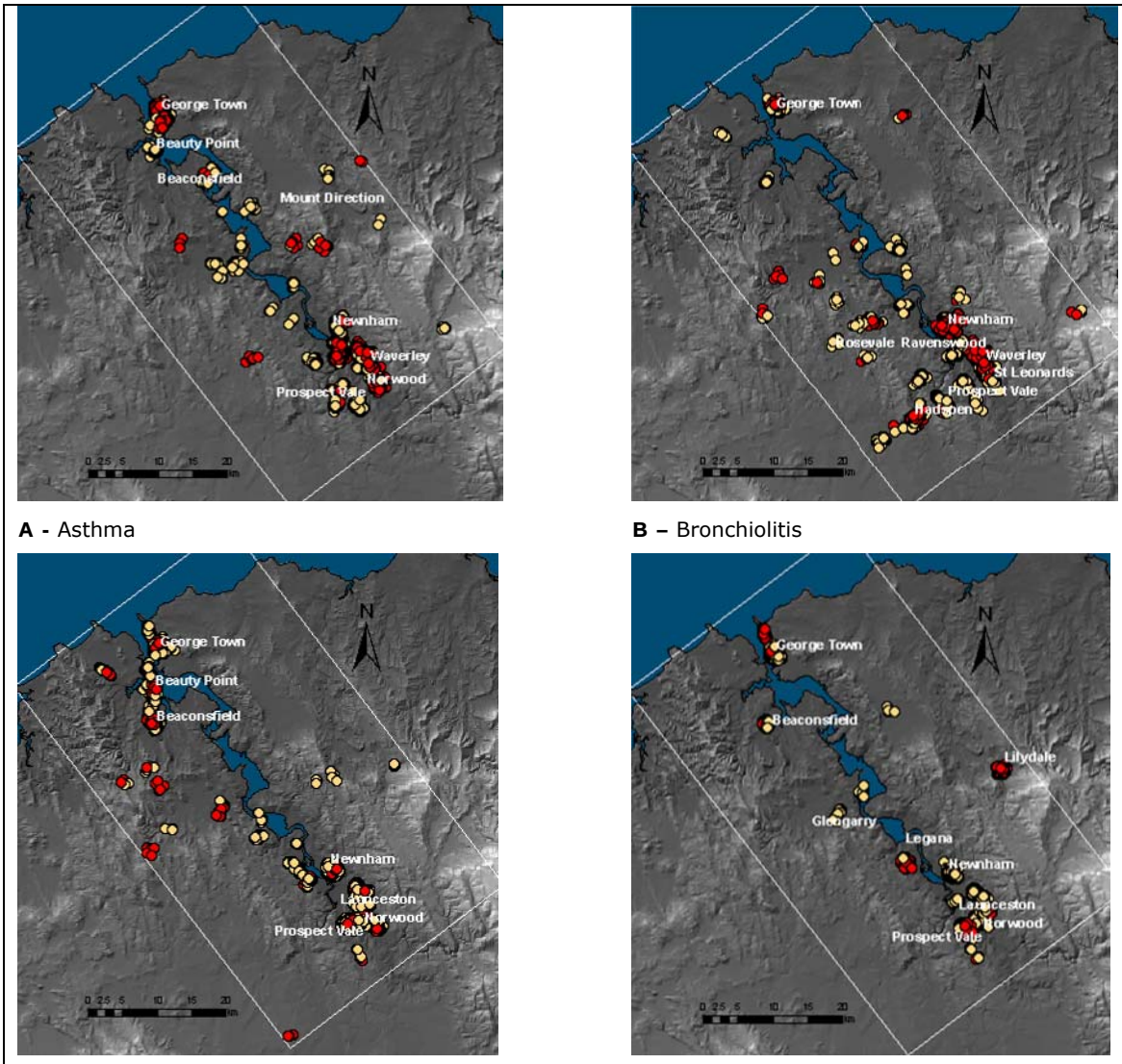


Figure 4.8 - Results of the Getis Ord Gi* statistic (with z-score rendering) for (A) Asthma, (B) Bronchiolitis, (C) Bronchitis and (D) COPD. A 1 km radial search window was used. Only clusters that are > 1 (beige) and > 2 (red) standard deviations above the mean expected rate of disease for the study area are displayed. To protect geoprivacy, isolated points were removed prior to publication.

The Getis Ord Gi* statistic was then applied to identify combined clusters of Asthma, Bronchiolitis and Bronchitis (ABB). The results, at two spatial scales, can be seen in Figure 4.9. The map of the entire dataset (on the left) shows that total disease is elevated in George Town, Beauty Point and Beaconsfield in the north, and the eastern suburbs of Launceston in the south. Scattered areas around Rosevale, Mount Direction and Hadspen also showed elevated levels.

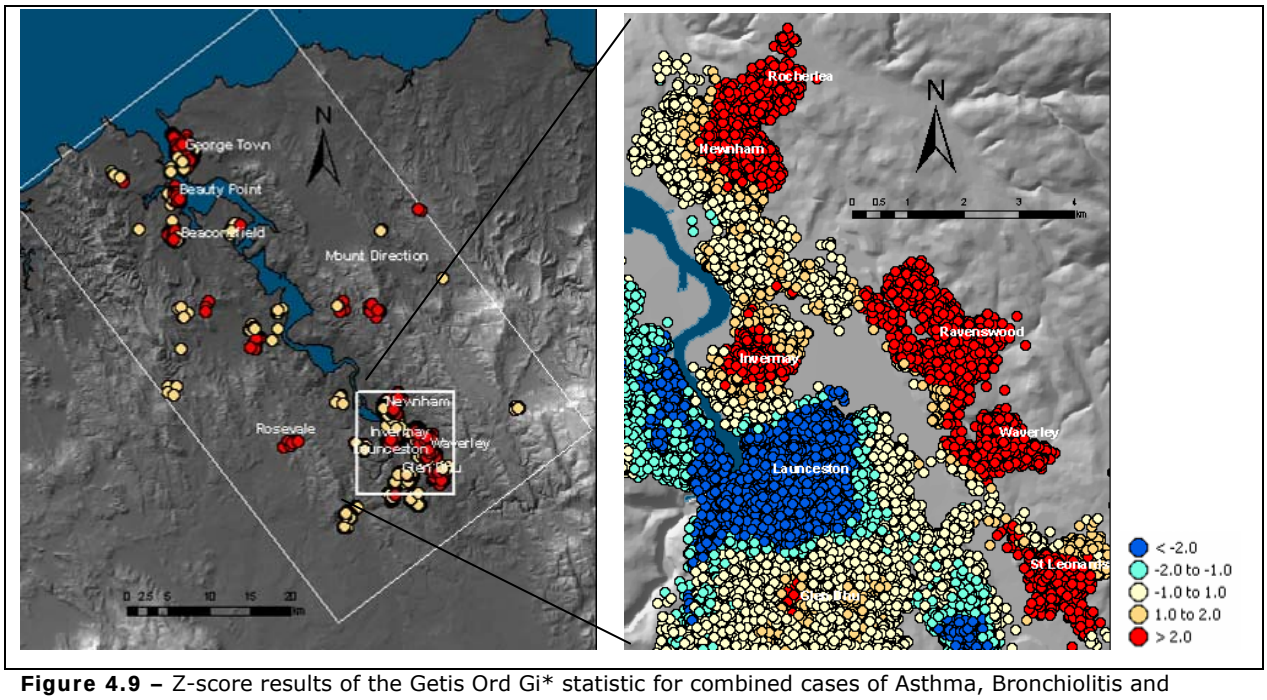


Figure 4.9 – Z-score results of the Getis Ord Gi* statistic for combined cases of Asthma, Bronchiolitis and Bronchitis for the entire study period (1992-2006). Left map shows only those clusters with a z-score greater than 1. The map on the right shows the Launceston area (the area within the white square in the left image), showing clusters with all z scores ranging between 2 standard deviations above and below the population mean. The localities of Rocherlea, Newnham, Ravenswood, Invermay, Waverley, St Leonards and Glen Dhu show 'significantly' higher rates of disease than the general population. ABB incidence around the Launceston CBD is at least 2 standard deviations below the mean. Isolated address points have been removed.

A close-up view of Launceston, seen on the right in Figure 4.9, reveals that the localities of Rocherlea, Newnham, Invermay, Ravenswood, Waverley, St Leonards and an area around Glen Dhu all display respiratory disease rates 'significantly' higher (> 2 standard deviations) than the population mean. These localities lie in the North Esk river valley. Disease incidence around the Launceston CBD and northwest to Trevallyn is at least 2 standard deviations below the mean, as is the area around Norwood to the south.

4.4.4 Kulldorff's spatial scan statistic

SaTScan was used to implement Kulldorff's spatial scan statistic. The Asthma dataset was analyses for clustering in the entire Tamar Valley (Figure 4.10), and a subregion around George Town (Figure 4.11).

Figure 4.10 shows that analysis of the entire valley detected a very large elliptical cluster falling predominantly on the eastern side of the Tamar River. This included the Rowella region on the northwest shore also. A smaller cluster to the north of Launceston CBD at Invermay was also detected (consistent with the results of the Gi* statistic). Both of these clusters were found to be significant at $p < 0.01$. Several very small clusters were also detected though they varied in significance; most of these included only a single point, which could be considered 'outliers' in this case. Only clusters containing more than one address point were of interest here, as multiple admissions from a single, isolated address were likely either aged care facilities or some other health care facility, both of which were not considered associated with geographic or climatic processes.

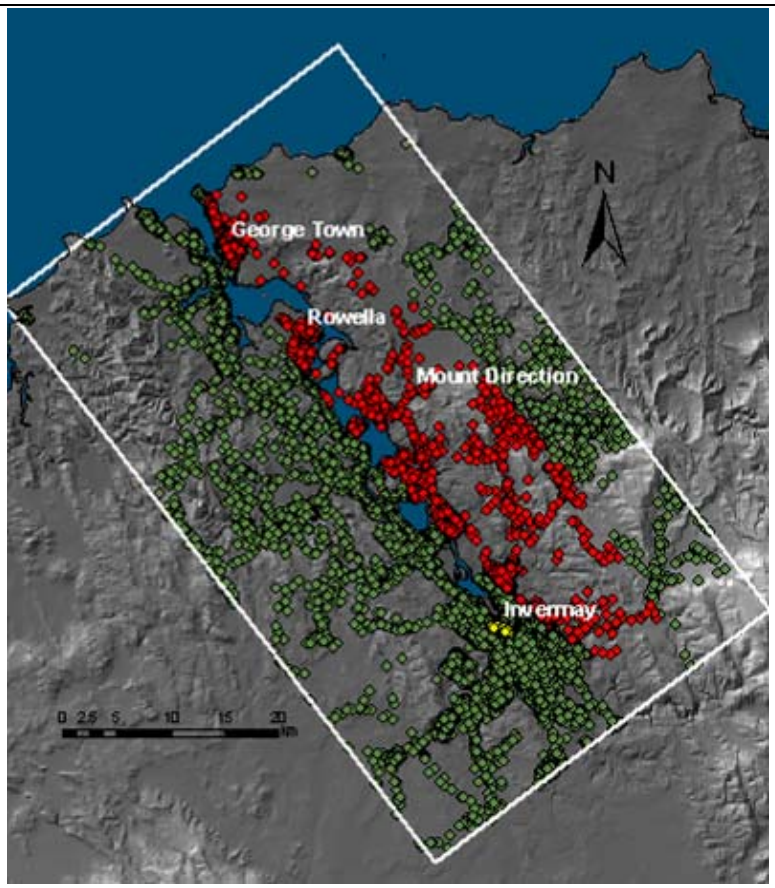


Figure 4.10 – SaTScan results for Asthma clusters in entire valley. Large red elliptical cluster centred around Mt Direction suggests that Asthma incidence is 'significantly' higher on the east side of the Tamar than the west (this cluster returned a p-value of <0.01). Smaller yellow cluster (p <0.01) is shown just north of Launceston CBD at Invermay. All other clusters vary in size and significance. Address points of the background population are displayed in green.

Figure 4.11 shows the results of the analysis of a George Town subregion of the dataset. Two elliptical clusters were detected in George Town and Beauty Point, which were both significant at $p < 0.01$. Three single point clusters were also detected, that again could be considered outliers. For comparison between the two, Figure 4.12 shows a map of the northern section of the SaTScan analysis of the entire valley (displayed in its entirety in Figure 4.10 This map covers the same extent as Figure 4.11 for ease of comparison with the George Town subset. It is seen here that SaTScan analysis of the entire dataset did not detect the 'significant' cluster at Beauty Point that can be seen in Figure 4.11 (shown in yellow). Also, the cluster in George Town seen in Figure 4.11 (displayed in red) contains fewer address points than that seen in the northern tip of the large red cluster in Figure 4.12, which stretches further east and north into Low Head. Generally, SaTScan results were 'coarser' for the entire dataset than for the George Town subset; analyses at different scale levels therefore produce different results.

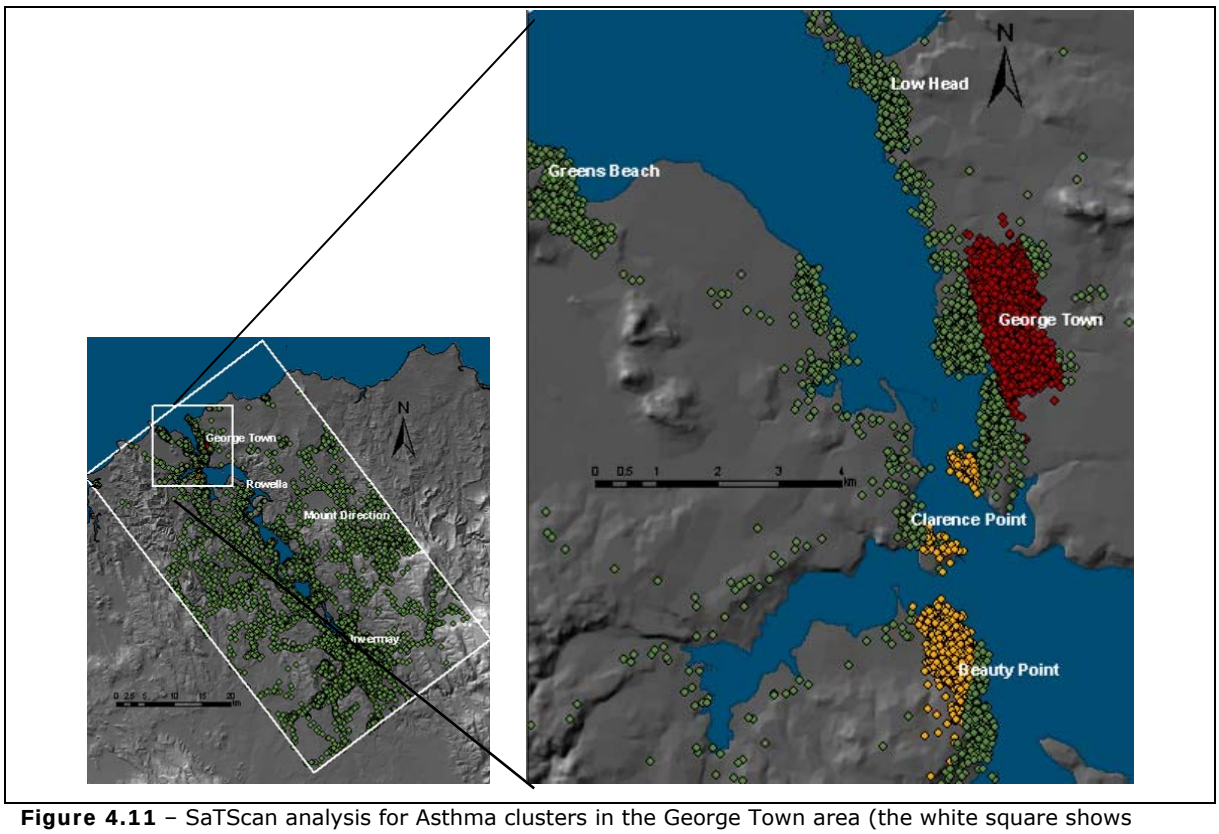


Figure 4.11 – SaTScan analysis for Asthma clusters in the George Town area (the white square shows analysis extent). Both the red and yellow clusters are significant at $p<0.001$. Green points are the scrambled address points of the APD; digital elevation model is displayed in shaded relief.

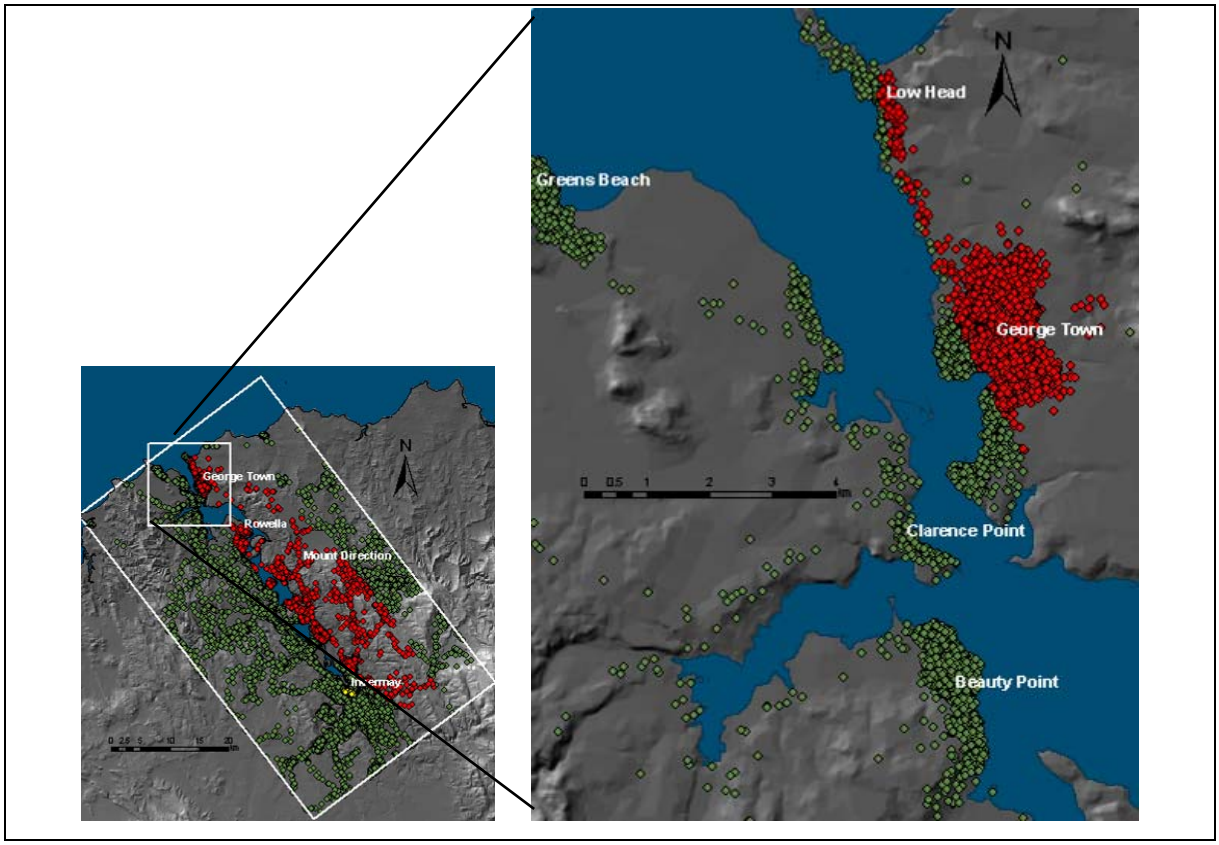


Figure 4.12 – SaTScan analysis for Asthma clusters in the entire study area, showing only the George Town region (the white square here is extent of display only). The red cluster is the northern tip of the large elliptical cluster in Figure 4.10; green points are scrambled address points of the APD. Digital elevation model is displayed in shaded relief.

However, given that the George Town subregion was analysed for Asthma incidence 'per house' and the total dataset was analysed 'per resident', these differences should be treated cautiously.

Given that SaTScan displays both cases and non-cases within a given cluster equally, the risk of re-identifying the residence of a respiratory patient is considered negligible. For this reason, isolated rural points were not removed from maps displaying SaTScan output (i.e. Figure 4.10, Figure 4.11 and Figure 4.12). Single-point clusters were, however, removed.

4.4.5 Summary statistics

Based on the results of the above cluster detection statistics, the areas around George Town, Hadspen and eastern and northern Launceston were identified as areas of elevated disease incidence. These areas were then investigated further.

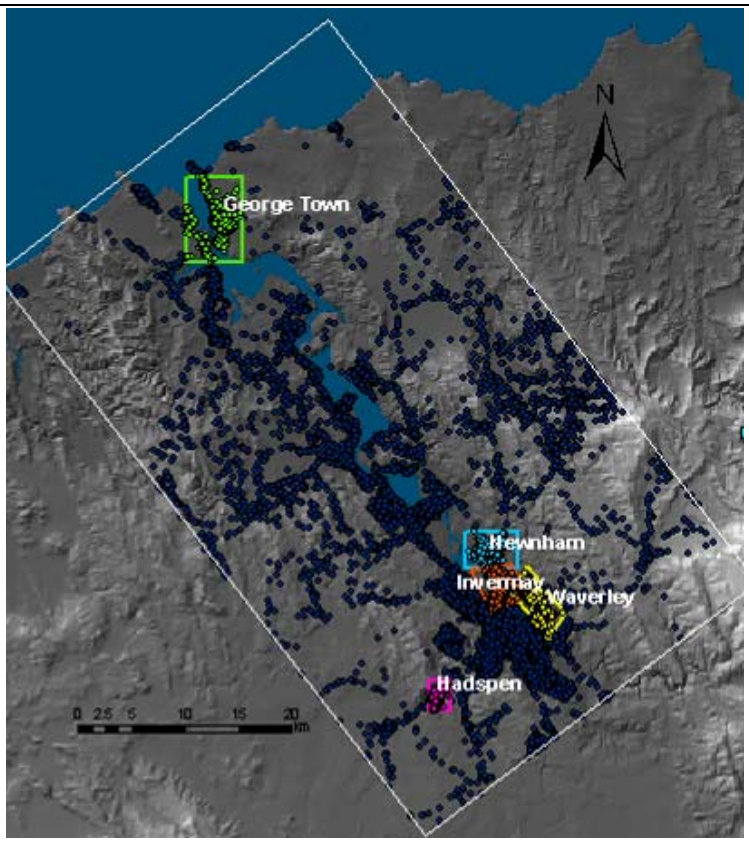


Figure 4.13 – Geographic location of digitised population subregions 'George Town' (green), 'Hadspen' (pink), 'Invermay'(orange), 'Newnham' (light blue) and 'Waverley' (yellow).

Figure 4.13 shows the geographical areas of subregions that were manually digitised in ArcGIS to incorporate areas of observed disease clustering. Disease incidence in each subregion was compared to disease incidence in the total population as a way of comparing these areas to the total study area. These results are summarised in Table 4-2 where it is seen that relative disease incidence in these local subregions is up to 65% higher than the total population.

Table 4-2 – Analysis of disease incidence in local subregions of the study area, showing up to 65% increase in disease incidence in these areas relative to the total population of the study area. Geographic locations used to define subregions are shown in Figure 4.13. (* Calculated as disease cases / population) (** Calculated as local disease incidence of subregion / disease incidence of total study area x 100)

Subregion	**Population**	**Disease cases (1992-2006)**	**Local disease incidence (%)***	**Disease incidence relative to total population (%) ** **
Total study area	94984	3730	3.92	100
'George Town'	5241	317	6.05	154
'Hadspen'	1847	83	4.49	115
'Invermay'	7506	360	4.80	122
'Newnham'	7615	383	5.03	128
'Waverley'	6277	407	6.48	165

4.5 Discussion

4.5.1 Disease clusters and statistical inference

The results of local cluster detection found in this study provide a compelling argument for the spatial analysis of health data. Distinct spatial and spatio-temporal patterns of elevated disease incidence were found, most consistently around George Town and the North Esk region east of Launceston. However, while results clearly show the presence of local disease clusters, the statistical 'significance' of these clusters should be viewed with prudence. For example, the Getis Ord Gi* statistic determines cluster significance based on a comparison between disease incidence inside the (1 km) radial search window and the whole population. The resulting 'significance' is therefore largely dependent on the chosen size of the search window, the presence of spatial autocorrelation in the dataset, and both the geographical size and spread of the entire study population (Getis and Ord, 1992). Similarly, the significance of clusters reported with Kulldorff's spatial scan statistic is highly dependent on the size of the study area. This can be clearly seen in Figures 4.11 and 4.12. As Kulldorff and Nagarwalla (1995) themselves caution, "it is not meaningful to attribute significance to a cluster without reference to the study region." Note also that disease 'clusters' identified by both the Gi* statistic and Kulldorff's spatial scan statistic included both disease cases and non-cases. That is, any given cluster (excluding those of a single point) is likely to contain a number of points with no recorded disease incidence. This requires careful consideration of the results prior to making assumptions about houses identified as being in a 'cluster'

Understanding issues of scaling is critical to the effective and accurate use of GIS in public health studies (Sexton et al., 2002, Wheeler, 2007 #135). The importance of defining the study area size is demonstrated in
Table 4-2; here, 'global' statistics have been calculated for each defined subregion. Hypothetically, if the 'Waverly' subregion represented the entire spatial extent of a study area, global statistics would return a total population disease incidence of 6.48%. It would not be known that disease incidence of this population was markedly higher than that of the population immediately surrounding the study area; and, it is possible that a cluster detection statistic would not find any local disease clusters within the study site. Similarly, the spatial extent of this study included only an area surrounding the Tamar Valley. This region has previously been reported as having the highest rate of "severe, chronic" Asthma in Tasmania (Giles, 1980). If a contemporary cluster detection study were to be done on the state of Tasmania as a whole, it is very likely that there would be different patterns of disease clustering observed within the Tamar Valley than those seen in the current study.

In addition to the disparity in statistical significance between datasets of different spatial extents, results also showed some disagreement between different cluster detection techniques. A comparison of FigureFigure 4.8 (A) andFigure 4.10 shows that the Gi* statistic found 'significant' clusters of Asthma on the western side of the Tamar that were not detected by SaTScan. Likewise, vast areas of the eastern Tamar that were included in a 'significant' Asthma cluster by SaTScan were not considered significant with the Gi* statistic. This problem has been reported in the literature. Wheeler (2007) commented that the different results returned from different techniques (SaTScan and kernel intensity function ratio) caused him to wonder "which was the most trustworthy?" Aamodt *et al* (2006) also found differences between methods and suggest that SaTScan may be more sensitive to clusters of lower relative risk than other methods tested. Researchers are generally warned here not to rely on the results of a single analysis method alone (Sahsuvaroglu and Jerrett, 2007).

The search window used in SaTScan initially centres on a point and then increases in size until (a maximum of) 50% of the population is included (Kulldorff, 1997; Kulldorff, 2006). It would seem then that the size of a resulting cluster would also be dependent on its geographical position relative to the rest of the population. Consider an elongated study area like the Tamar Valley with a population centre at either end; Launceston in the south has a far greater population than George Town in the north. Conceptually, a scan window around a point in the centre of Launceston would reach a size that incorporates 50% of the population before it reaches the population centre of George Town. As such, it would seem that a cluster in Launceston (which includes approx. 77% of the study population) will be influenced more heavily by the population immediately surrounding it than a cluster in George Town, which could have been formed by a much larger scan window, potentially reaching all the way to northern Launceston. This concept is illustrated in Figure 4.10, where it is seen that a very large cluster was detected in the sparsely populated area on the east of the Tamar, and a very small cluster was found in the city centre at Invermay. SaTScan has elsewhere been found to have difficulty detecting large clusters within large population centres, where the population distribution in the study area is not uniform (Aamodt et al., 2006). Irregularly shaped clusters are also generally more difficult to detect than circular or elliptic ones and the limitation of SaTScan to only incorporate either circular or elliptical search windows has also been noted by other researchers (Aamodt et al., 2006; Wheeler, 2007).

Another factor detracting from the statistical strength of findings in this study is the positional error inherent in the datasets. As detailed in Chapter 3, all address points were randomly moved up to 200 m from their 'true' location. Also mentioned was the

fact that the original 'true' locations of address points were already subject to some positional error in the Address Points Dataset as a whole (theLIST, 2007). For cluster detection statistics that are dependent on the *relative* location of disease cases within the population, these spatial errors may have influenced any perceived results.

4.5.2 Hospital admissions as a measure of morbidity

This study defined respiratory disease occurrence by Launceston General Hospital admissions records. This approach influenced the dataset in the following ways. Firstly, hospital admissions represent the 'top-end' of respiratory disease occurrence, as only the sickest in the population are admitted to hospital. There is a greater number of people who are treated in the Emergency Department and do not get admitted to hospital, and an even greater number of people who visit a General Practitioner or some other health care provider instead of entering the hospital system. Further still down the 'health hierarchy', many more people will increase the use of self medication (e.g. with Asthma inhalers) before seeking any medical assistance. Using only hospital admissions data therefore constricts the size of the possible study population.

Secondly, studying admissions to a public hospital necessarily precludes all respiratory patients who were admitted to private hospitals; this introduces a form of socio-economic confounding that causes the poorer in society to be over-represented (Giles, 1980). Also, it has been commented that asthmatics living at the edges of a hospital catchment tend to seek medical treatment in their local areas rather than relying on the hospital system (Giles, 1980); thus it seems that those living in remote rural areas may also be under-represented by relying solely on hospital admissions. An advantage in dealing purely with hospital admissions data, however, is that human error is minimised. ICD-9/10 codes were decided only by a small number of hospital staff from a single hospital over the study period, and therefore diagnostic discrepancies were kept to a minimum.

Fifteen years of accumulated hospital records were analysed simultaneously in this study to address the problems associated with small population studies. However, a change in the international disease coding system occurred in the study period, which had some impact on the way respiratory diseases were classified. The conversion from ICD-9-CM to ICD-10-AM between 1998 and 2000 caused some changes in respiratory disease diagnoses (AIHW, 2000; AIHW, 2001). Most notably, COPD diagnoses increased considerably as diseases that were previously labelled chronic Asthma or Emphysema are now diagnosed as COPD (Wisconsin Department of Health and Family Services, 2004). As COPD was not included in analyses of 'total disease' in this study this effect may not be great. Also, the general decline in Asthma incidence over the study period seen in Figure 4.4 does not seem totally attributable to this change in

diagnosis. Therefore, while some effects of the changeover are noted, they are not considered substantial in this study.

4.5.3 Spatial autocorrelation and confounders

The vast majority of cluster detection studies in the literature define the clustering of disease cases as a departure from Complete Spatial Randomness (CSR) (Aamodt et al., 2006; Abrams and Kleinman, 2007; Durand and Wilson, 2006; Kingham et al., 2006; Kulldorff, 2006; Kulldorff and Nagarwalla, 1995; Sabel and Loytonen, 2004; Wheeler, 2007; Wilson et al., 2004; Wilson et al., 2006; Wilson and Zawar-Reza, 2006). CSR was the underlying assumption in each of the methods employed in this section of the study. However, it has been argued that a null hypothesis of CSR is not appropriate for environmental and social studies which are dealing with heterogenous background populations, or autocorrelation (Goovaerts and Jacquez, 2004; Wakefield and Shaddick, 2006).

While spatial autocorrelation is present in most if not all public health datasets, failure to account for it will further affect the statistical validity of any reported clustering (Goovaerts and Jacquez, 2004; Jerrett and Finkelstein, 2005; Lipton et al., 2005; Maheswaran and Craglia, 2004; Maheswaran and Haining, 2004; Sabel and Loytonen, 2004). Specifically, positive global autocorrelation in a dataset will return false positive results for local clustering (Getis and Ord, 1996; Ord and Getis, 1995). A suggested alternative is to adopt a Neutral model, which takes into account spatial autocorrelation in the background population and tests the alternative hypothesis as a departure from that pattern (Goovaerts and Jacquez, 2004). Standardised Morbidity Ratios (SMR) can then be calculated based on background age structure and various other confounding factors (Goovaerts and Jacquez, 2004).

As mentioned, this study did not have access to personal or demographic data such as age, gender, socioeconomic status, smoking history and occupation. All of theses are known to influence the results of cluster detection, and most exhibit positive spatial autocorrelation (Carstairs, 2000; Hayes, 2003; Jerrett et al., 2003; Koch and Denike, 2004; Maantay, 2002; Sexton et al., 2002). “Confounding” generally has been described as “the single most important problem” in studies of cluster detection (Wartenberg and Greenberg, 1993) in: (Maheswaran and Craglia, 2004), and Rothman (1990) has gone so far as to suggest that it is of limited scientific value to conduct disease cluster analyses at all, given all the limitations of data quality and confounders. However, cluster detection studies are usually the precursor to further investigations (Besag and Newell, 1991) and even in the absence of confounding information, disease mapping and cluster detection contributes valuable information to pre-epidemiological

studies (Maheswaran and Craglia, 2004; Sabel and Loytonen, 2004). Cluster detection techniques were therefore used primarily as an exploratory tool in this study.

4.6 Conclusion

Kernel density estimation was used purely for exploratory spatial data analysis in this study. Spatial analysis of acute respiratory disease in the Tamar Valley revealed several local clusters of 'significantly' elevated disease risk using the Gets Ord Gi* statistic and Kulldorff's spatial scan statistic. Total disease incidence in local areas tested was found to be up to 65% higher than the total (global) disease rate. Variations were detected in the spatial location of each disease (Asthma, Bronchiolitis, Bronchitis and COPD), and spatio-temporal variations were found in disease distribution between years and between seasons. Disease levels generally declined over the study period (1992-2006), with George Town showing the most dramatic decline in disease cases. All diseases showed spatial patterns different to that of the background population, suggesting that disease occurrence is not randomly spread across the population.

Specifically, Launceston's eastern and northern suburbs of Rocherlea, Ravenswood, Waverley, Invermay, Newnham and St Leonards (i.e. the North Esk region) consistently showed elevated disease levels (i.e. Figure 4.9 and Figure 4.10). This was particularly true for Asthma and Bronchiolitis (e.g. Figure 4.8). The areas around Glen Dhu in southern Launceston and Hadspen to the west also showed elevated disease rates. To the north, Bronchitis was found to be especially prevalent around the northwest Tamar regions of Beauty Point and Beaconsfield, and COPD was found to be highly clustered in a small number of addresses thought to be nursing homes and aged care facilities. Disease rates (particularly Asthma) were especially high in George Town in the early years of the study period (1992 – 1995) (see Figure 4.6), and seasonal variations in spatial patterns generally were small (Figure 4.7). None of these findings could have resulted from a purely global (non-spatial) investigation of the dataset.

This chapter has provided a useful pre-epidemiological analysis of the spatial distribution of respiratory disease in the Tamar Valley. The next chapter combines these disease data with various proxies for particulate air pollution dispersion, thereby investigating possible spatial relationships between observed disease clusters and particulate air pollution exposure.

5 Exposure surfaces and disease: correlations

5.1 Research Hypothesis

- "Spatial distribution of winter disease incidence can be explained by exposure surfaces representing winter particulate air dispersion in the Tamar Valley."
- [H_{null}: "There is no observable spatial relationship between respiratory disease incidence and the exposure surfaces used."]

5.2 Introduction

This chapter explores the relationships between winter disease incidence and various exposure surfaces used to represent the spatial distribution of winter particulate air pollution in the Launceston region. The analyses were conducted on a spatial subset of Launceston rather than the entire Tamar Valley because reliable air pollution modelling data, based on known residential wood heater emissions, was only available for Launceston. Two exposure surfaces were derived from prognostic dispersion modelling of wood heater emissions in Launceston using TAPM (The Air Pollution Model) (Hurley, 2005), with surfaces representing both the average and maximum concentrations of PM_{10} over a single winter season. TAPM is currently used widely for modelling domestic and industrial pollution sources in Tasmania. A digital elevation model was also analysed to produce an additional three exposure surfaces based on terrain features known to support winter cold air drainage and cold air ponding. Statistical relationships between disease incidence and exposure surfaces were explored, and Geographically Weighted Regression was applied to test for spatial stationarity in these relationships.

5.2.1 Winter air pollution in the Tamar Valley

The Tamar Valley is a large coastal river valley characterised by stable atmospheric conditions and anti-cyclonic weather patterns in the winter months (Power, 2001). Under these conditions 'temperature inversions' are formed by nocturnal katabatic drainage transporting cold air from surrounding hill slopes down into the valley floor, displacing warmer, lighter air (Sturman and Tapper, 2006). These temperature inversions have been found to reach vertical heights of between 80 and 300 metres under stable atmospheric conditions (Lyons and expert working party, 1996; Nunez, 1991)

Temperature inversions have a two-fold effect on air pollution levels in the valley. Firstly, decreased temperature at ground level in the valley floor causes an increase in home heating. Wood combustion is a major source of domestic heating in the Tamar Valley (DEH, 2005b), and wood smoke is Launceston's primary source of particulate air pollution (Ayers et al., 1999; Todd et al., 1997). Secondly, wood smoke emissions and

other particulate pollutants become trapped within the dense air mass formed by the temperature inversion. Pollutant levels then accumulate through the night until the inversion is dissipated by solar radiation or wind, usually the following morning (Sturman and Tapper, 2006). This combination of increased emissions and poor dispersion leads to vastly elevated levels of particulate air pollution in Launceston and the Tamar Valley on winter nights.

Tethersonde balloon soundings conducted over two nights of worst case anticyclonic conditions in Launceston in 1991 modelled the formation and movement of the temperature inversion layer over Launceston (Nunez, 1991). This study reported that particulate pollution emissions were predicted to concentrate in the low lying areas of the North Esk valley and towards Glen Dhu and southern Launceston (Nunez, 1991). This pattern of pollutant accumulation in terrain depressions follows a process known as 'cold air ponding', where cold air drainage builds up over the night to fill follows in the landscape (Sturman and Tapper, 2006).

Lyons *et al.* (1996) also reported that smog/air pollution was worse in some areas of the Tamar Valley than others, and that 'smog zones' had been included on official maps of the valley dating prior to 1985.

The Tamar Valley Airshed Study (TVAS) (Power, 2001) contributed greatly to our understanding of air pollution dispersion within the Tamar Valley. Air pollution from residential and industrial sources was extensively modelled over a two year period using NUATMOS/CITPUFF diagnostic modelling. This study reported marked variations in seasonal wind flows through the valley and noted that anticyclonic conditions noticeably worsened the air pollution problem in Launceston. The frequency of anticyclonic days (defined as those with a recorded surface atmospheric pressure equal to or greater than 1020 hPa) rose from 17% in spring and summer to 30% in autumn and 40% in winter.

5.2.2 Exposure assessment and the ecological fallacy

Many researchers have commented on the difficulties of linking health effects to environmental exposure in ecological studies (Cakmak et al., 2003; Christie et al., 1992; Elliott et al., 2000a; Elliott et al., 2000b; Jerrett and Finkelstein, 2005; Mindell and Barrowcliffe, 2005; Pearce, 2002; Sexton et al., 2002; Waller and Gotway, 2004; Weng and Yang, 2006; Wilson et al., 2004; Wilson et al., 2006; Zeger et al., 2000). 'Exposure' in the context of epidemiology refers to the levels of pollution that each participant or person in a study area is exposed to over the duration of a study period. This is invariably very difficult to quantify, and in some cases also difficult to define (Briggs, 2000). Researchers very rarely have access to personal exposure data for all

participants in a study population; in all other cases, some means of approximating exposure is necessary. However, inaccuracies in air pollution measurements and exposure assessment can have significant consequences for the estimation of morbidity and mortality from pollution effects (Colvile and Briggs, 2000; Ito et al., 1995; Zeger et al., 2000). As such, a lack of reliable exposure data is widely considered to be the weakest link in epidemiology (Colvile and Briggs, 2000; Elliott et al., 2000b; Zeger et al., 2000).

The 'ecological fallacy' is a universal problem in epidemiology and spatial science and refers to the incorrect assumption that the characteristics of an individual, or household, can be directly inferred from the characteristics of the group, or aggregation, to which it belongs. A similar problem is ecological bias and this refers to false inferences of biological impact on an individual based on ecological effects within the study area (Maheswaran and Craglia, 2004). Cross-level studies then are those that draw inference from relationships between layers of different spatial resolutions (e.g. exposure surfaces and population), a category of analysis described as being "particularly susceptible to ecological bias" (Maheswaran and Craglia, 2004). The potential for ecological bias is reduced in small area studies however, as exposure is generally more consistent across the study population when the study area is small (Arnold et al., 2000; Elliott et al., 2000a).

A majority of epidemiological studies use data from a single or few air pollution monitoring sites to link air pollution exposure with health outcomes. Studies have shown, however, that urban environments can show extreme variations in pollution concentrations within distances of only tens of metres, so this methodology is now cautioned against (Brindley et al., 2004 ; Wilson et al., 2004). In spatial studies, various geostatistical interpolation methods such as kriging are commonly used to interpolate pollution levels between monitoring stations. This method also has its limitations however, as accurate kriging of air pollution in urban environments requires a dense network of monitoring sites to adequately reflect intraurban variations (Briggs, 2005).

Substantial differences in the spatial dispersion of different *size fractions* of particulate pollution have also been reported. Monn (2001) and Wilson *et al* (2004) have each reported findings from extensive literature reviews of air pollution monitoring studies which found that heterogeneity was evident in most urban areas. Heterogeneity was defined by Wilson *et al* as concentrations that varied 20% or more between sites. Notably, the larger size fraction of particulate pollution ($PM_{10-2.5}$) was generally found to be more heterogeneously spread than smaller particulates ($PM_{2.5}$), indicating that larger, heavier particles do not travel as far from their source as finer, lighter particles

(Wilson et al., 2004). Ultrafine (<0.1 µg) particles though were found to exhibit similar patterns of heterogeneity to coarse particles, which suggested that these particles are most easily transported via local turbulence (Monn, 2001; Wilson et al., 2004). These findings have implications for studies that have previously inferred $PM_{2.5}$ concentration as a direct proportion of PM_{10} levels. Future work should perhaps address the quantification of PM_{10} and $PM_{2.5}$ dispersion separately across the landscape.

5.2.3 Air pollution dispersion modelling

Air pollution dispersion models can be broadly categorised as either diagnostic or prognostic. Each requires the location and emission levels of pollution sources in the study area, though diagnostic models also utilise considerable input data from field measurements of air pollution concentration and local meteorology. Diagnostic models are therefore capable of producing detailed site-specific outputs of air pollution dispersion for a given study period, and they require limited specialised training to operate (Jerrett et al., 2005a). The time and resources involved in the collection of field data, however, can render diagnostic models less attractive than prognostic models (Jerrett et al., 2005a). Conversely, prognostic dispersion models use data sourced from synoptic scale meteorology in combination with digital terrain models and pollution emission rates of the study area, to produce outputs of air pollution concentration at any location and time in the study period for which meteorological data is available (Hurley, 1999; Hurley, 2005). Prognostic models therefore offer the great advantage of being applicable to a retrospective study. However, prognostic dispersion models require extensive training to be run effectively (Jerrett et al., 2005a). Many models employ a combination of prognostic and diagnostic elements for increased accuracy (Power, 2001).

TAPM (The Air Pollution Model) is a three-dimensional prognostic dispersion model recently developed by the CSIRO (Hurley, 1999; Hurley, 2005). TAPM uses prognostic meteorology and terrain feature information to solve fundamental fluid dynamics and vector transport equations, thereby dispensing with the need for ground based meteorological observations (Hurley, 2005; Luhar and Hurley, 2003). TAPM is currently widely used for pollution dispersion modelling of both industrial and residential pollutants in Tasmania.

While pollution dispersion models have advanced in both complexity and accessibility in recent years, dispersion modelling has only rarely been used in exposure assessment for public health studies (Briggs, 2000; Briggs, 2005; Colvile and Briggs, 2000; Elliott et al., 2000a; Jerrett et al., 2005a). This seems to be attributable both to a lack of knowledge of dispersion models among epidemiologists, and to a lack of confidence in the outputs (Briggs, 2000). A review of validation tests for dispersion modelling, for

example, found that dispersion models generally displayed 70% accuracy at best. Also the maximum spatial resolution of outputs is usually 1 km (TAPM outputs at 500 m resolution), and this is considered too coarse for many exposure assessment studies (Jerrett et al., 2005a).

5.2.4 Terrain analysis as proxy for cold air drainage

A comprehensive study of intraurban particulate concentrations in Christchurch, New Zealand, was recently conducted by Wilson *et al* (2006) using a network of monitoring sites. Christchurch has a comparable air pollution problem to Launceston. Wood and coal combustion is the principle source of particulate pollution in this city and winter temperature inversions cause regular exceedances of air pollution standards (Wilson et al., 2006). This study found the lowest measured particulate concentrations occurred on a hill; this was attributed to winter cold air drainage transporting particulates to lower elevations in the city (Spronken-Smith et al., 2002), in: (Wilson et al., 2006).

A recent micro-scale study based in Lenah Valley, Hobart, measured local fluctuations in PM_{10} across the valley landscape (Ling, 2002). Similar to the Tamar Valley, Lenah Valley is subject to temperature inversions and katabatic cold air drainage under stable winter conditions (though less pronounced and on a smaller scale than the Tamar Valley). This study recorded both increased PM_{10} and decreased temperature in the valley floor compared to hill tops on 'katabatic days' (days when katabatic winds were noted). Over a difference in elevation of around 100 m between sites, it was reported that areas of the valley floor were "regularly subjected to up to 5 times greater concentrations of PM_{10} than areas at higher elevations" (Ling, 2002).

In the absence of reliable dispersion modelling, researchers commonly use proxies for pollution exposure estimates such as proximity to pollution source, level of source activity, land use and terrain feature analysis (Corburn, 2007; Elliott et al., 2000a; van de Kassteele et al., 2006; Weng and Yang, 2006). In the Tamar Valley, it has been found that under winter anticyclonic conditions, particulate air pollution produced in Launceston drains to surrounding areas of lowest elevation (Lyons and expert working party, 1996; Nunez, 1991). This phenomenon of increasing PM_{10} concentrations with decreasing elevation has been observed by other researchers (e.g. (Ling, 2002; Wilson et al., 2006)). Terrain analysis has previously been used as a proxy for katabatic drainage along river valleys in the Hobart area (Smeal, 1998).

Terrain feature analysis was used in this study as a surrogate for cold air drainage and cold air ponding in the Tamar Valley. Concave terrain features such as valleys (at various scales) and relatively flat, low-lying areas were treated as conduits for cold air drainage and the accumulation of particulate matter through cold air ponding. These

landscape features could then be used in place of actual PM_{10} monitoring data to simulate relative particulate concentration accumulations within the landscape. Terrain proxies in this way enable examination of environmental variables at a more detailed spatial scale than the coarse (500 m) resolution of prognostic dispersion modelling allows.

5.3 Software and data requirements

A digital elevation model (DEM) with 25 m spatial resolution was provided by Land Information Services Tasmania (theLIST). This dataset was derived from 1:25,000 topographic maps and has a spatial accuracy of 12.5 m. Particulate air pollution dispersion modelling results derived using TAPM ("The Air Pollution Model") version 3.0.7 (Hurley, 2005) were provided by Dr Michael Power, Senior Scientific Officer (Air/Noise) of the Tasmanian Department of Tourism, Arts and the Environment (DTAE).

ESRI software ArcGIS version 9.1 was used for the majority of spatial data analysis and map generation in the study. LandSerf v 2.2 was used to categorise the DEM into terrain types based on quadratic surface modelling (Wood, 2005). Microsoft Excel (2003) was used primarily for graphing data interactions. The statistical package JMP (version 5.1) (SAS Institute Inc., 2003) was used for analysis of the different interactions of 'cases' and 'non-cases' with all exposure surfaces in the population, and R (Venables et al., 2006) was used to convert data information from ASCII to tabular format for use in spatial regression analysis. Geographically Weighted Regression version 3 (Charlton et al., 2003) was used to analyse spatial stationarity in relationships between disease locations and all exposure surfaces.

5.4 Methods

5.4.1 Non-spatial investigation of PM_{10} and disease

Microsoft Excel was used to explore the non-spatial relationships between hospital admissions and $_{PM10}$ concentrations, as measured at the Ti Tree Bend monitoring station at Invermay. Annual admissions for Asthma, Bronchiolitis, Bronchitis and COPD were each compared with 95th percentile PM_{10} concentrations. Relationships between PM_{10} and combined disease admissions for Asthma, Bronchiolitis, Bronchitis and COPD (ABBC), and then combined admissions for Asthma, Bronchiolitis and Bronchitis (ABB) (i.e. minus COPD) were also tested for annual trends.

Results of this investigation indicated that COPD was not correlated with PM_{10} fluctuations at the same temporal scale as the other more acute diseases (ABB). COPD admissions were therefore excluded from the dataset and relationships between PM_{10}

and ABB were explored at a finer temporal scale. Daily fluctuations in ABB hospital admissions and 24 hour average PM_{10} concentrations were plotted for each year of the study period.

5.4.2 Spatial extent of study area

The spatial extent of the study area chosen for this section of analysis was based on the area for which TAPM (The Air Pollution Model) could be reliably run. TAPM required inputs of the known proportion of residential wood heater emissions (i.e. pollution sources) for a given study period. This corresponded to an area encompassing all of Launceston in the southern Tamar Valley, seen in Figure 5.1 below. The bounding coordinates for this area (in GDA 94 UTM Zone 55S (MGA)) are, Top: 5430783, Left: 492687, Right: 527712 and Bottom: 5393283.

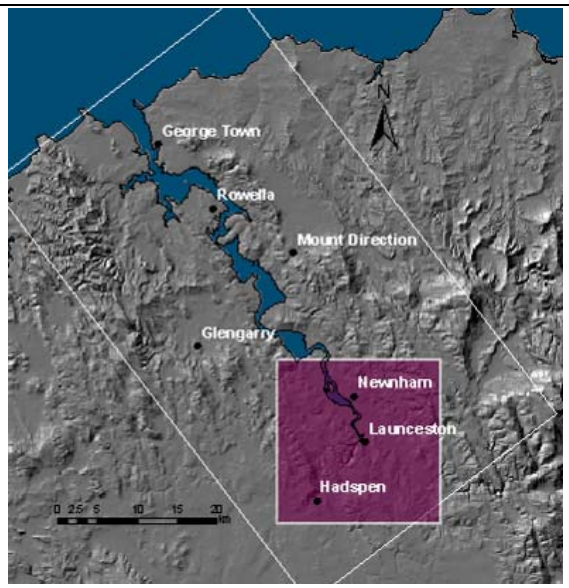


Figure 5.1 – The spatial extent of this section of analysis is shown as the pink rectangle covering Launceston, Newnham and Hadspen.

5.4.3 Digital elevation model

A digital elevation model (DEM) is a raster (grid) dataset used to store topographic information of a region. Each grid cell in a DEM represents the mean elevation of the area within that grid cell. A DEM with square grid cells of 25 m in length is said to have a 25 m resolution, and is commonly referred to as a "25 m DEM".

A 25 m DEM of Tasmania was clipped to an area of slightly larger extent than the Launceston study area seen in Figure 5.1. It was made larger rather than the same size to eliminate 'edge effects' from subsequent LandSerf analyses. All 'no data' values corresponding to water in the DEM were changed to zero with the conditional raster calculation: `(con[IsNull[DEM], 0, IsNull[DEM])`, which uses an "if…then…else" statement; this served two important functions. Firstly, as a result of the random perturbation process, some address points ended up in the Tamar River. If the grid

cells of the DEM that comprised the spatial extent of the river had remained set at 'no data', these address points would have been ignored when combined with elevation data for further analysis. By changing these values to zero these address points were assumed to lie at an elevation of 0 m, which was acceptable given the surrounding low lying river flats of their likely 'true' location. Secondly, subsequent terrain analysis in LandSerf was greatly improved by using a DEM with continuous values, as 'no data' values around the river created considerable 'edge effects' in outputs.

5.4.4 TAPM – The Air Pollution Model

TAPM version 3.0.7 was run by Dr Michael Power at the Tasmanian Department of Tourism, Arts and the Environment (DTAE). TAPM was run in single tracer mode, using residential wood heater emissions as the sole pollution source. Wood smoke from residential heating is known to contribute to the majority of particulate load in Launceston (Ayers et al., 1999; Lyons and expert working party, 1996; Power, 2001; Todd et al., 1997). Census data for 2001 was used as a measure of housing density (occupied dwellings/km^2). The spatial distribution of wood heater emissions was calculated as a simple proportion of 30% of housing density; of this 30% of wood-burning dwellings, 70% were assumed to have AS4013 compliant heaters, 23% non-compliant heaters and 7% open fires – proportional emission levels were adjusted accordingly (DEH, 2005b; Power, 2007). Meteorological data including temperature, wind speed and wind direction were input as hourly means. Emissions were specified at a mean daily temperature of 10°C; a linear scaling function doubled emission levels for a temperature of 0°C, and reduced emissions to zero at a temperature of 20°C. TAPM used a 250 m DEM for terrain height reference, and a land use reference file also of 250 m resolution; ground level meteorology is determined through interactions with terrain height and surface features (Hurley, 2005). Soil data (for ground moisture approximation) were used at 5 km resolution.

TAPM outputs used in this study were maximum ("cmax") and average ("cavg") PM_{10} concentrations for winter months (June, July and August). Only winter months were selected to capture the unique conditions of wood smoke dispersion under winter meteorological conditions, where anticyclonic weather patterns and temperature inversions predominate in the Tamar Valley (Lyons and expert working party, 1996; Nunez, 1991; Power, 2001; Sturman and Tapper, 2006). "Cmax" refers to a grid surface with *each grid cell* displaying the maximum modelled PM_{10} concentration encountered at that grid cell for the entire model run (in this case, every day of winter 2005). "Cavg" then refers to a grid surface containing the average PM_{10} concentration for each grid cell. All outputs were at 500 m resolution, which is the finest scale possible.

TAPM writes grid files in a different format to ArcGIS, so it was necessary to transpose TAPM grid files into ASCII files before importing them into ArcGIS. A short macro was developed to facilitate this process. Grids were resampled from 500 m to 25 m resolution for analysis with terrain surfaces, and later resampled again to 200 m resolution for analysis with Geographically Weighted Regression. This is explained further below.

5.4.5 LandSerf terrain analysis

LandSerf is an open source software package developed by Wood (2005) that enables the categorisation of a DEM into different terrain classes. LandSerf applies a quadratic approximation to elevation data for classification of terrain feature types. 'Quadratic approximation' refers to the fitting of a second order polynomial trend (or a trend with a single curve) to a surface. Under this model, elliptic regions are classed as 'pits' and 'peaks', parabolic regions as 'channels' and 'ridges', and hyperbolic areas as 'passes'; flat regions are classed as 'planes' (Wood, 1996). In this way, each grid cell in the resulting surface is assigned a landform class based on the grid cells in its neighbourhood. This concept is illustrated Figure 5.2 below.

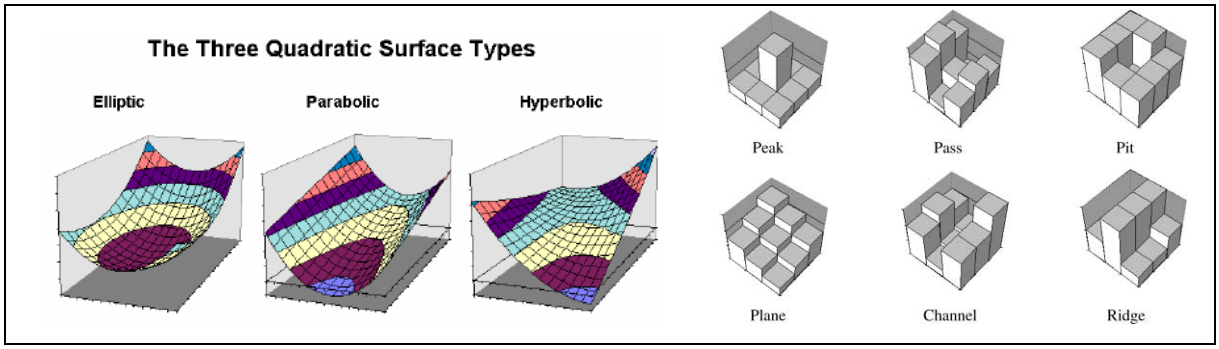


Figure 5.2 – Demonstration of the quadratic approximation function used in LandSerf to categorise digital elevation models into the six terrain classes listed on the right (peak, pass, pit, plane, channel and ridge). *(Sources: (Wood, 1996); (Fisher et al., 2004))*

LandSerf v 2.2 was used to categorise a 25 m DEM of the Launceston region into six terrain features: 'channel', plane', 'ridge', 'peak', 'pass' and 'pit'. Multi-scale fuzzy feature classification was applied, which is a concept based on fuzzy set theory (Fisher et al., 2004). Fuzzy set theory has commonly been used in the classification of 'vague' landforms, such as mountains, in geographical analysis. In fuzzy set theory, a concept, e.g. 'channel', is first defined and objects that match that concept exactly are assigned a 'membership' of 1 to the channel class; weaker matches are assigned progressively smaller membership values until 0, indicating no match at all (Fisher et al., 2004). *Multi-scale* fuzzy membership then takes into account the different classifications of the same point that can occur at different scales across the landscape. This concept is illustrated in Figure 5.3 where it can be seen that the top of a hill can variably be

assigned to the morphometric class 'channel', 'plane' and 'ridge' at different scales. The output of multi-scale fuzzy feature classification is six surfaces (one for each terrain class) with each cell in a layer having a membership to that terrain class between 0 (never) and 1 (at all scales). Only 'channels' and 'planes' were of interest in this study as these were the landforms believed to best represent the terrain depressions that support cold air drainage and ponding (Smeal, 1998; Sturman and Tapper, 2006).

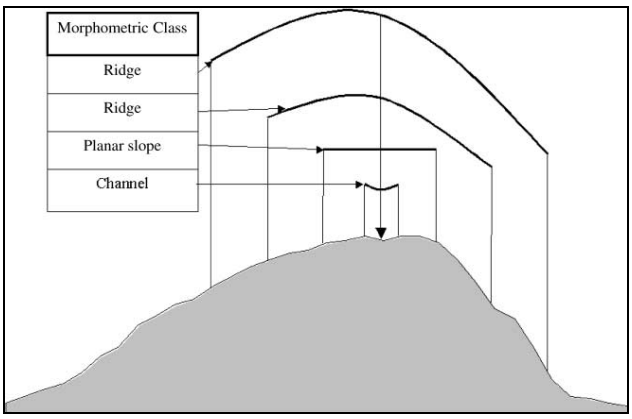


Figure 5.3 – Illustration of multi-scale fuzzy feature classification, showing that the same point on a landscape can be assigned different morphometric classes at different spatial scales. *(Source: (Fisher et al., 2004))*

An upper window size of 101 grid cells was used in this analysis (windows must have an odd number of cells to ensure values can be assigned to the centre point of the window). The upper limit of the search window was chosen intuitively by visually analysing the spatial scale of terrain features in the DEM; this showed valleys with a maximum width of around 2.5 km. Given the DEM grid resolution of 25 m, an upper window size of '101' corresponded to 101 x 101 grid cells at 25 m each, which equals a linear distance of 2.525 km. An adaptive search window ranging from 3 cells (75 m in width) up to 101 (2525 m) thus classified the terrain at each window size.

All LandSerf grid layers in the study were produced at 25 m resolution; all TAPM outputs were resampled from 500 m to 25 m resolution. All LandSerf and TAPM outputs were then 'snapped' to the DEM. Snapping is a process whereby an entire raster layer is shifted slightly so that individual grid cells align perfectly in geographic space with those of another raster layer. This is necessary in order to accurately perform spatial analyses between layers. Once all raster layers were correctly aligned in this way, the DEM and LandSerf layers were resized to the spatial extent of the TAPM layers. This was achieved by using one of the TAPM outputs as a 'mask' for each of the other raster layers, resulting in all raster layers being exactly aligned and sized.

5.4.6 Zonal statistics and data generalisation

In order to ascertain the influence of the exposure surfaces on respiratory admissions, it was necessary to first attribute the values of all underlying grid surfaces to individual address points in the study area. In this way, the local conditions giving rise to increased particulate air pollution at each address point (both cases and non-cases) could be summarised in a single table for further regression analysis. Given that all address points had previously been shifted up to 200 m from their true location it was not appropriate to simply attribute underlying grid values directly to the scrambled address point location. To remedy this, a 200 m circular buffer was created around each address point and summary statistics of the underlying exposure surfaces were then calculated for the buffer areas (i.e. zonal statistics). These values were then assigned to the corresponding *scrambled* address point in the centre of that buffer as an approximation of exposure for all possible locations of the *true* address point.

As the vast majority of 200 m buffers in the study area overlapped (address points in Launceston are predominantly less than 200 m apart), Zonal Statistics were calculated using a Spatial Ecology extension to ArcGIS which adequately coped with overlapping buffers (Beyer, 2004). The Zonal Statistics tool calculates summary statistics (i.e. minimum, maximum, mean, median, standard deviation, sum and count) of the values of a raster layer that fall within the spatial boundaries of each zone, or buffer (Beyer, 2004). Given the 25 m resolution of all grid layers in the study, each 200 m radial buffer covered approximately 200 grid cells, each with a unique value; this is illustrated in Figure 5.4. In order to determine whether or not it was appropriate to use the mean value of these grid cells in further calculations, it was first necessary to assess the distribution of data values within each zone.

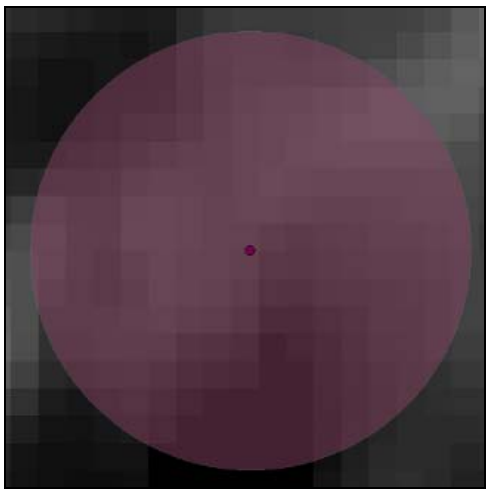


Figure 5.4 – Enlarged image of a non-overlapping 200 m buffer, or zone, illustrating the variation in underlying grid cells. Zonal statistics provide summary statistics on these grid cell values and assign these to the centroid of the zone, which in this case is the address point in its scrambled location.

Four buffers were selected at random from the study area and histograms were plotted of the spread of values in each underlying exposure surface grids. If the values of a grid layer followed a Gaussian (normal) distribution for the four sample zones, then the mean could be considered a suitable measure to assign to the address point; if values were obviously skewed, a log transformation would be necessary prior to analysis. Figure 5.5 shows the histograms produced for the three terrain grid layers (DEM, channel and planar) at each of the four sample buffers. (Given the coarse resolution (500 m) of the layers produced by TAPM it was not necessary to assess the distribution of values from these layers.) It can be seen that, while there is considerable variation in the distribution of values within all zones, there is no obvious deviation from a normal distribution. The mean grid value for each buffer was therefore considered an acceptable measure of the underlying exposure surface to attribute to the scrambled address points. Zonal means were then calculated for all buffers in the study area (n = 33791) using all five exposure surfaces (DEM, 'planar', 'channel', PM_{10} cmax and PM_{10} cavg).

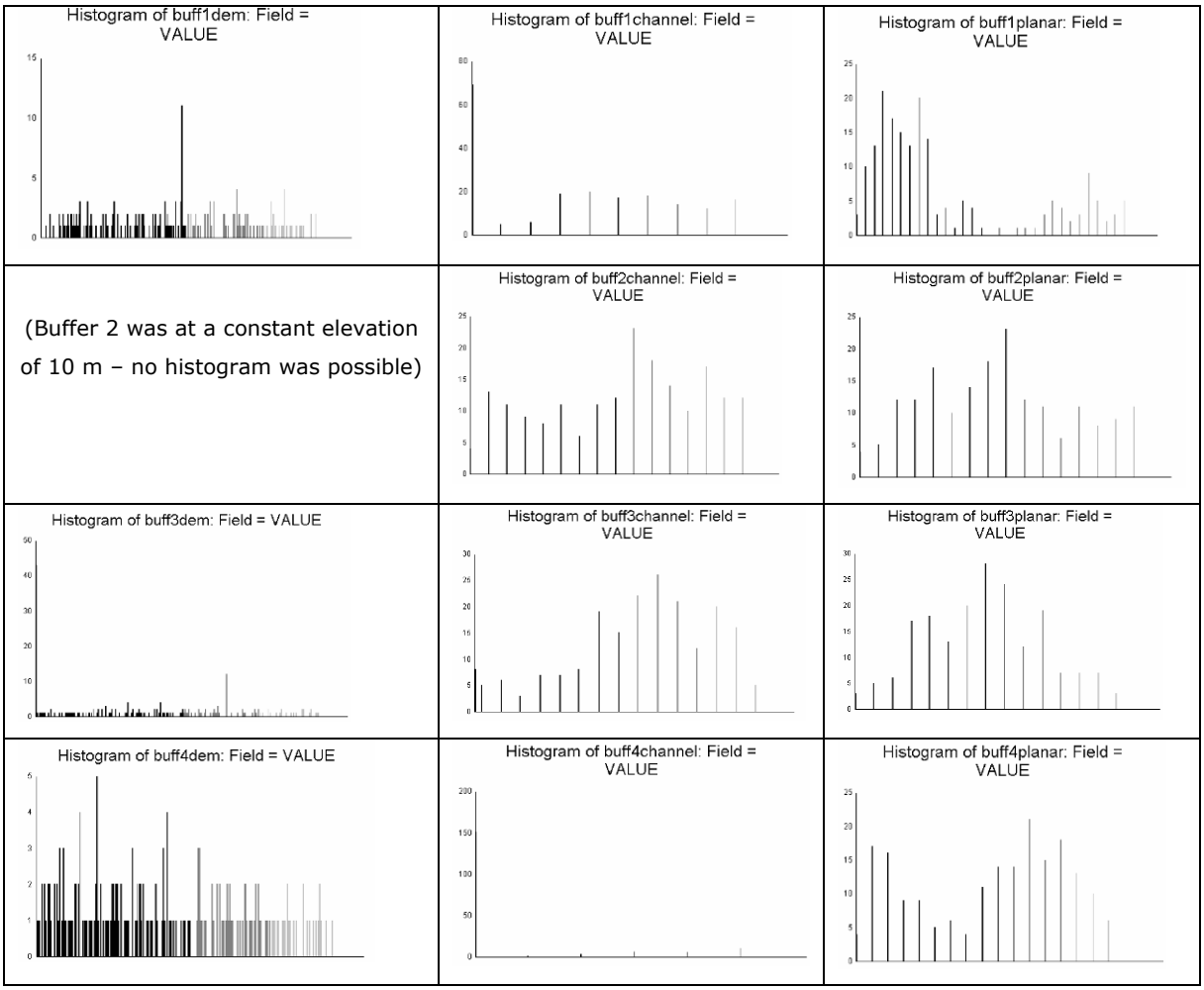


Figure 5.5 – Histograms showing the spread of raster values within four sample buffers of the DEM (left), channel (middle) and planar (right) grid layers. While distributions vary markedly between buffers, none of the histograms are obviously skewed, and so can be considered normally distributed.

The dataset produced in this way, with zonal statistics of exposure surfaces assigned to each of the 33,000+ points, unfortunately proved too cumbersome to be easily run with Geographically Weighted Regression (GWR). Specifically, a spatial logistic regression model in GWR based on these 33,000+ points would take up to 25 days to run on a desktop PC (Martin Charlton, Pers. Comm.). The dataset was therefore reduced in both spatial extent and resolution, via a process described below. The above analysis did, however, show that exposure surface data within 200 m of address points followed a reasonably normal distribution, which was important for Geographically Weighted Regression.

Asthma, Bronchiolitis and Bronchitis (ABB) cases for all winters in the study period (1992-2006), ABB cases for winter 2005, and total population were each generalised into point density layers using the 'point density' tool in the ArcGIS Spatial Analyst extension. This produced layers with a measure of density per km^2. Binary disease data (1/0; case/non-case) was used as a representation of *disease houses*, rather than disease cases per se. This was thought to give a better indication of the geography of disease incidence without the confounding influence of multiple admissions from the same address. These density layers were all produced at a resolution of 200 m. The DEM and LandSerf layers 'channel' and 'planar' were each resampled from 25 m to 200 m resolution. TAPM output layers representing 'cavg' and 'cmax' were resampled from 500 m to 200 m resolution.

These 200 m grids (eight in total) were converted to ASCII files and the dataset as a whole was then converted to tabular format using a script in the statistical package R (Venables et al., 2006). This process resulted in a table with centre X and Y coordinates and values for each corresponding raster layer assigned to each 200 m grid cell; a sample of this dataset is shown in Table 5-1. With all data in a single table, and the dataset very much reduced in size (from 33,000+ points to 6075 grid cells), Geographically Weighted Regression could be run.

Table 5-1- An section of the table used to run Geographically Weighted Regression, containing centre coordinates of each grid cell and all corresponding raster values. 'Case05' corresponds to the density surface of disease cases for winter 2005; 'wscase' is disease cases for all winters (1992-2006).

ID	X	Y	channel	cavg	cmax	dem	case05	planar	pop	wscase
1	503787	5420383	0.091	2.572	11.355	30.117	0	0.849	271.06	5.96
2	503987	5420383	0.086	2.461	10.887	20.632	0	0.898	201.92	1.98
3	504187	5420383	0.008	2.278	10.107	17.223	0	0.988	83.552	0
4	504387	5420383	0.016	2.278	10.107	12.853	0	0.983	16.415	0

5.4.7 Global statistical analysis

Binary disease data for all winters and 2005 winter months were assessed against all exposure surfaces using the statistical software package JMP (v 5.1) (SAS Institute Inc., 2003). Data for exposure surfaces were derived from the zonal statistics calculations detailed in the previous section; in this way, exposure data associated with each address location corresponded to the mean value for all possible locations of the true address location (see section 5.4.6 and Figure 5.4). Global relationships between respiratory disease and exposure surfaces for the entire dataset were then observed. Univariate t-tests were also performed to determine the significance of relationships.

5.4.8 Geographically Weighted Regression

In contrast to regression, which gives a global measure of the correlation of two variables, Geographically Weighted Regression (GWR) calculates regression coefficients between data points at all specified spatial locations across a dataset. In this way, site-specific data relationships can be assessed and spatial non-stationarity can be detected (Charlton et al., 2003). A non-stationary relationship varies across space or time, and this includes most associations observed in social and environmental science (Fotheringham et al., 2002a). GWR was used in this study to test for stationarity in the relationships between terrain features known to support cold air drainage and health outcomes for acute respiratory morbidity. In this way, the local influence exerted by each exposure surface on disease outcomes could be assessed.

The incidence of acute respiratory disease (ABB) admissions for all winters in the study period was used as the dependent variable in GWR; all terrain-based exposure surfaces (representing 'channels', 'planes' and elevation) and total population were regressed against this variable. Disease admissions for winter 2005 were analysed separately against TAPM outputs. Disease admissions data for *all* winters, rather than *2005* winter months, were used in the main regression analysis because they were approximately 15 times greater in number than the data from 2005 winter. This provided better support for the model, and the larger disease dataset was thought to give a better indication of the influence of terrain features on health outcomes.

GWR applies a kernel window to the dataset (similar to that used to illustrate Kernel Density Estimation in Chapter 3) and local regression models are fit to all points within each kernel; the choice of kernel size, or bandwidth, therefore has a considerable impact on the results obtained from GWR (Fotheringham et al., 2002a).

GWR 3 was run using a Gaussian model with an adaptive spatial kernel; the 'Akaike Information Criterion' (AIC) was used for bandwidth optimisation. This procedure allowed a single, optimal bandwidth to be determined for all regression calculations in

the dataset, rather than separate bandwidths being used for each calculation. This resulted in a single bandwidth of 312 nearest neighbours being determined for use throughout the modelling run (Fotheringham et al., 2002a). Local regression values were then determined based on 312 data points, or in this case grid cells, for each calculation. For each regression calculation, different weightings were assigned by the model to each data point according to its distance from the regression point. Note though that all data points in the analysis were evenly spaced as they corresponded to the centre co-ordinates of 200 m grid cells.

The result of GWR is a list of parameter values describing the strength of regression relationship between the dependent variable (disease density) and each independent variable (terrain surfaces and population). These results were written to the geographic centroids of each grid cell in the 200 m output grid; these points were then converted to raster files with a 200 m resolution in ArcGIS for visualisation of outputs.

5.5 Results

5.5.1 Non-spatial analysis of PM_{10} and disease relationships

Non-spatial relationships between disease admissions and PM_{10} concentrations, as measured at the Ti Tree Bend monitoring station, are shown below in Figure 5.6. Asthma and Bronchitis admissions are seen to be closely correlated to PM_{10} levels. Bronchiolitis admissions rose markedly in 1999-2000, suggesting an increase in diagnosis associated with the ICD-9 to ICD-10 changeover at this time, though when this is considered, Bronchiolitis admissions also seem reasonably correlated with annual PM_{10} fluctuations. COPD shows almost an inverse relationship with PM_{10} levels. Similar to Bronchiolitis, COPD admissions rose dramatically with the change to ICD-10 disease coding in 1999-2000, though they continued to rise sharply until 2004, indicating that the increase in admissions may not be entirely attributable to the change in diagnosis. COPD levels also rose steadily between 1992 and 1994, while PM_{10} levels fell over this time.

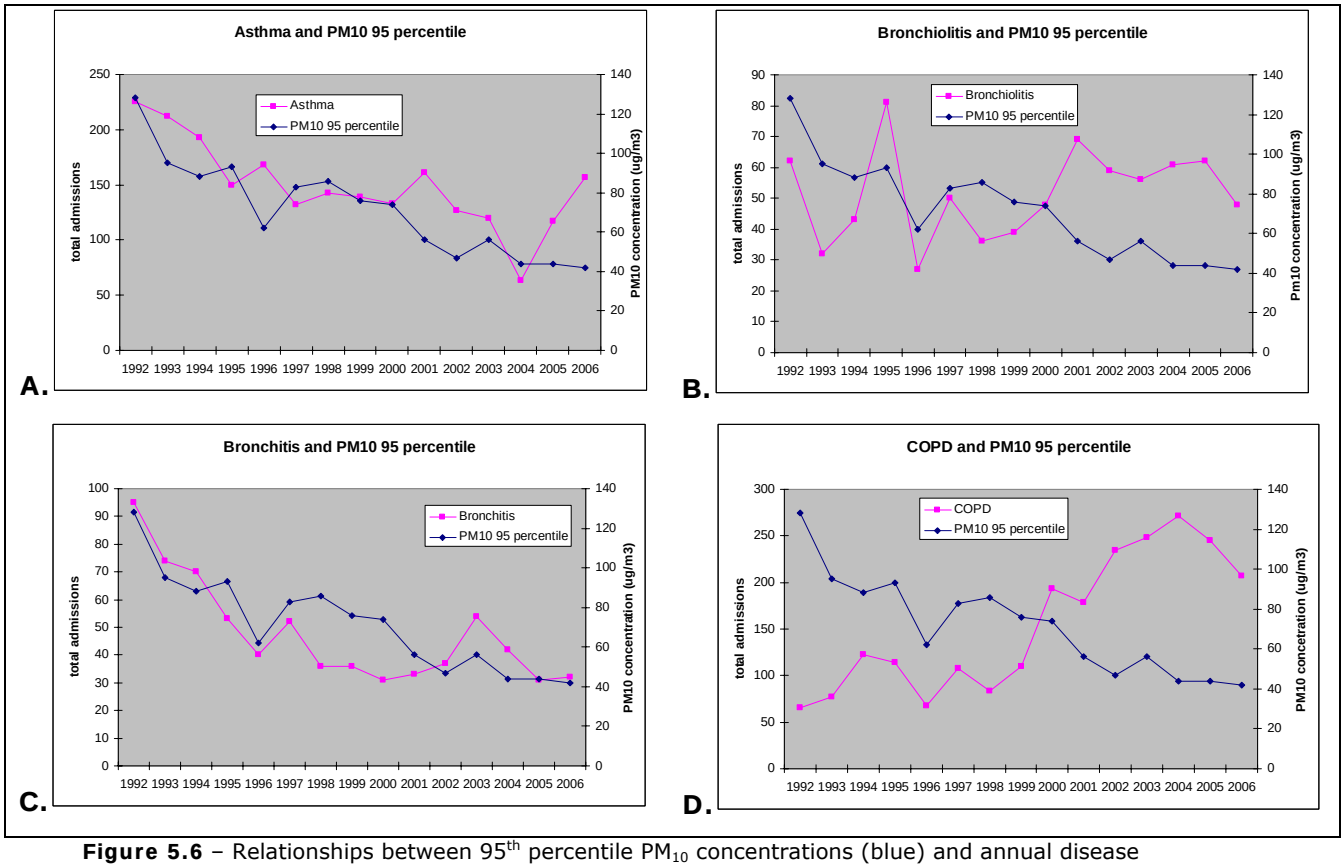


Figure 5.6 – Relationships between 95th percentile PM_{10} concentrations (blue) and annual disease admissions (pink) for (A) Asthma, (B) Bronchiolitis, (C) Bronchitis and (D) COPD for each year in the study period.

A comparison of Figure 5.7 and Figure 5.8 further illustrates the anomalous relationship between COPD and PM_{10}. These graphs show that combined Asthma, Bronchiolitis, Bronchitis and COPD (ABBC) admissions show agreement with PM_{10} only until 1999; beyond 2000 there is a strong departure from this trend. With the removal of COPD, however, combined Asthma, Bronchiolitis and Bronchitis (ABB) admissions are strongly correlated with 95th percentile PM_{10} concentrations throughout the study period. This result provides further justification for the removal of COPD from analysis in this study.

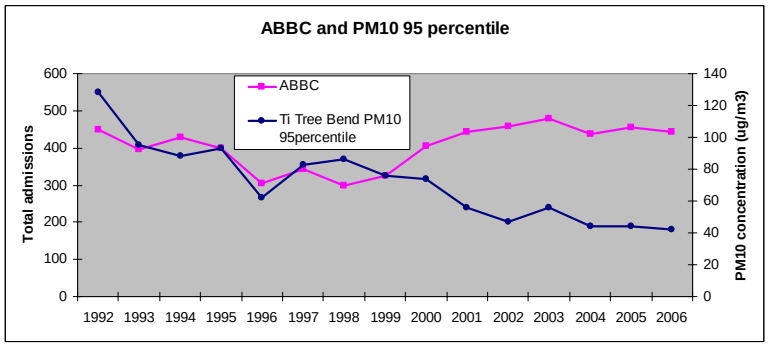


Figure 5.7 - Combined Asthma, Bronchiolitis, Bronchitis and COPD (ABBC) plotted against Ti Tree Bend PM_{10} 95th percentile values for each year in the study period, showing reasonable correlation until 1999, then a strong departure from this trend.

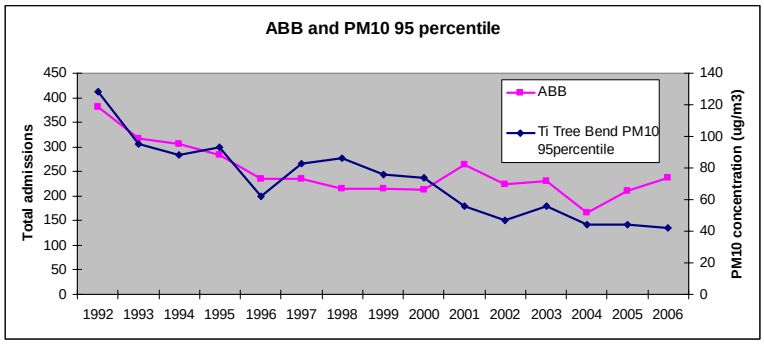


Figure 5.8 - Asthma, Bronchiolitis and Bronchitis (ABB) (i.e. minus COPD) plotted against Ti Tree Bend PM_{10} 95th percentile values for each year in the study period, showing a much more consistent correlation with the removal of COPD.

The paired graphs shown in Figure 5.9 compare daily fluctuations in hospital admissions for ABB (upper, pink graphs) and 24hr average PM_{10} concentrations recorded at Ti Tree Bend (lower, blue graphs). In the early years of the study period, PM_{10} measurements were recorded only every 6 days and predominantly in winter; numerous gaps in PM_{10} data therefore made comparisons with disease data difficult. For this reason, only data from 2003 to 2006 are displayed here. These paired graphs show strong patterns between disease admissions and 24hr average PM_{10} recordings at Ti Tree Bend. It is seen that both disease admissions and PM_{10} exhibit seasonal variations each year, with peaks in winter, though the intensity of this trend is variable. The severe peak in PM_{10} measured at Ti Tree Bend in December 2006 corresponds largely to road dust contamination, with some bush fire smoke also detected (DPIWE, 2007).

The year 2005 shows a reasonably high incidence of both disease admissions and PM_{10}, and a strong seasonal influence in both datasets that isn't entirely aligned (i.e. a temporal lag may be present). This year was therefore selected for further investigation into how disease admissions and PM_{10} concentrations interact *spatially* within the valley. Model inputs for TAPM were therefore based on 2005 meteorology and emissions data.

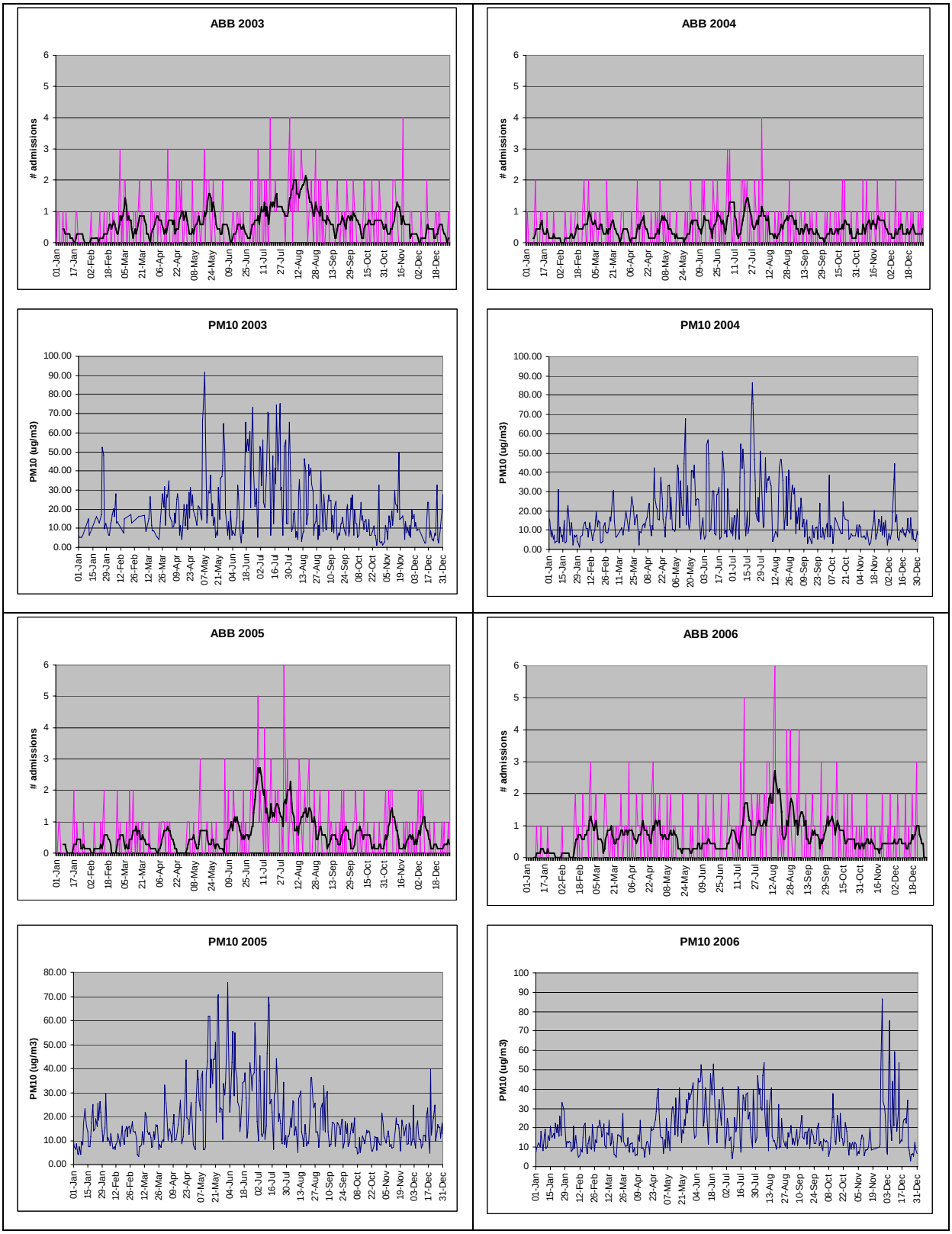
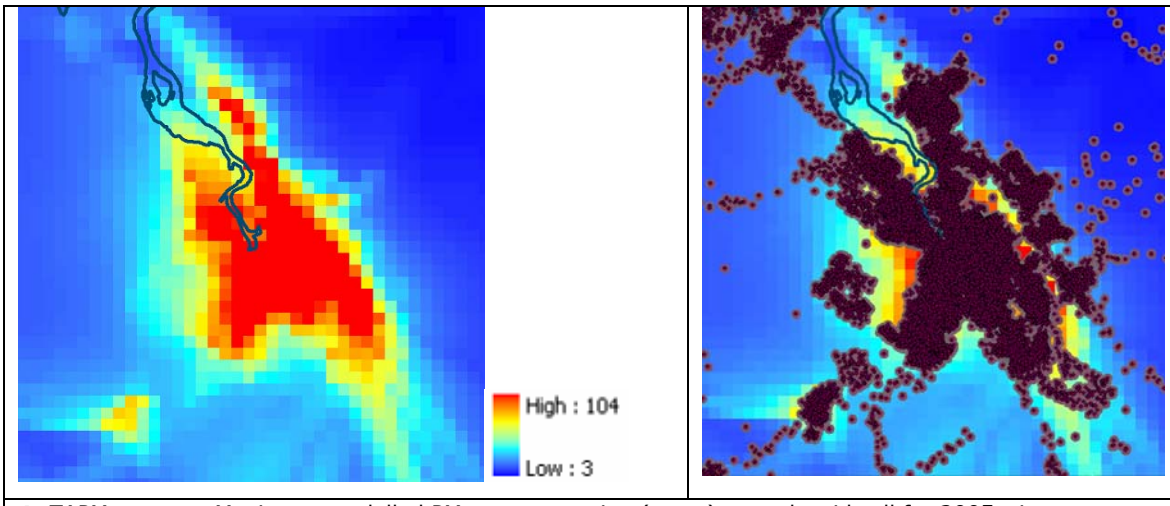


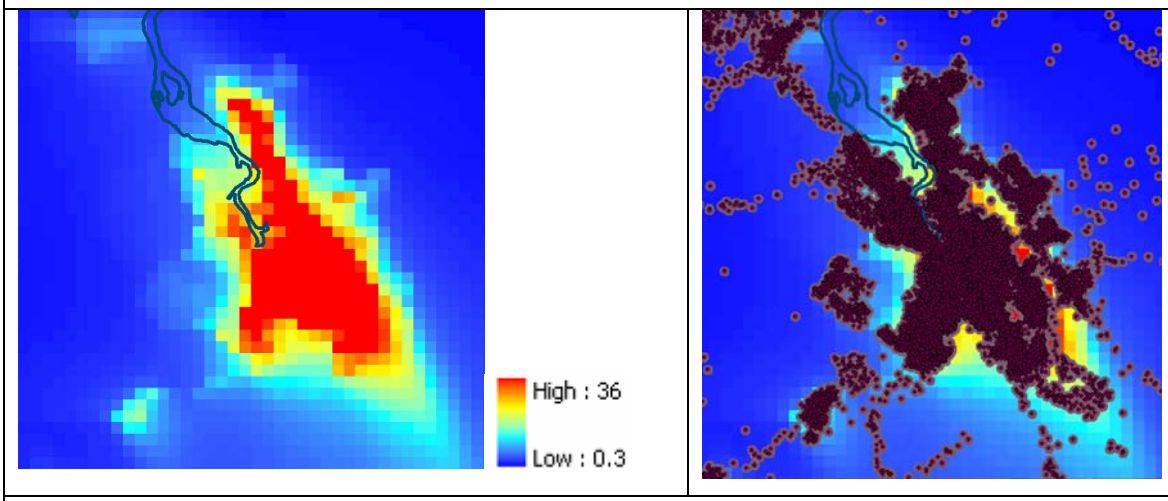
Figure 5.9 - Paired graphs of daily fluctuations in (upper, pink) combined hospital admissions for Asthma, Bronchiolitis and Bronchitis (ABB), and (lower, blue) 24hr average concentrations of PM_{10} recorded at Ti Tree Bend, for the latter years of the study period (2003-2006).

5.5.2 TAPM and LandSerf exposure surfaces

All exposure surfaces are shown in Figure 5.11 and Figure 5.11, displayed with and without address points and buffers used in zonal statistics calculations. TAPM outputs (Figure 5.11) are seen to closely follow the distribution of address points, or housing density, in the Launceston region. LandSerf 'planar' regions predominate the terrain surface, as the majority of areas at the scale tested (2.5 km search window) are 'flat' regions on the river plain or hill slopes. The LandSerf 'channel' layer is seen to follow the rivers and valleys of the DEM (seen as the white regions of high 'membership' to the channel class). Figure 5.11 also illustrates the difference in spatial resolution of the exposure surfaces, not only in grid cell size but in precision. Examination of the spatial relationships between exposure surfaces and disease are possible only at a relatively coarse resolution for the TAPM outputs, while terrain analysis shows much finer detail.

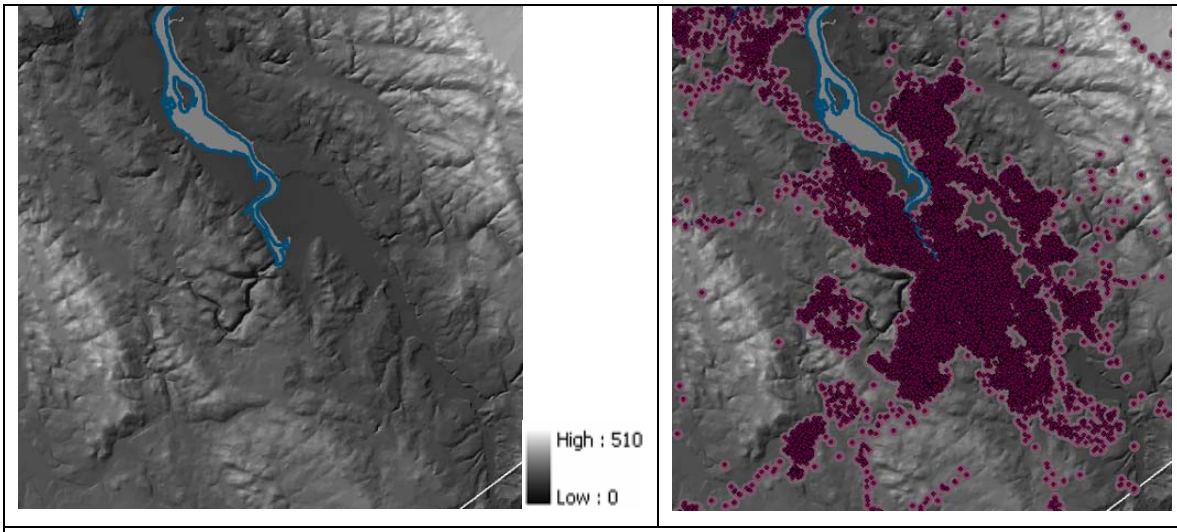


A. TAPM output – Maximum modelled PM_{10} concentration (cmax) at each grid cell for 2005 winter. Concentrations are in µg/m^3. Spatial resolution is 500 m.

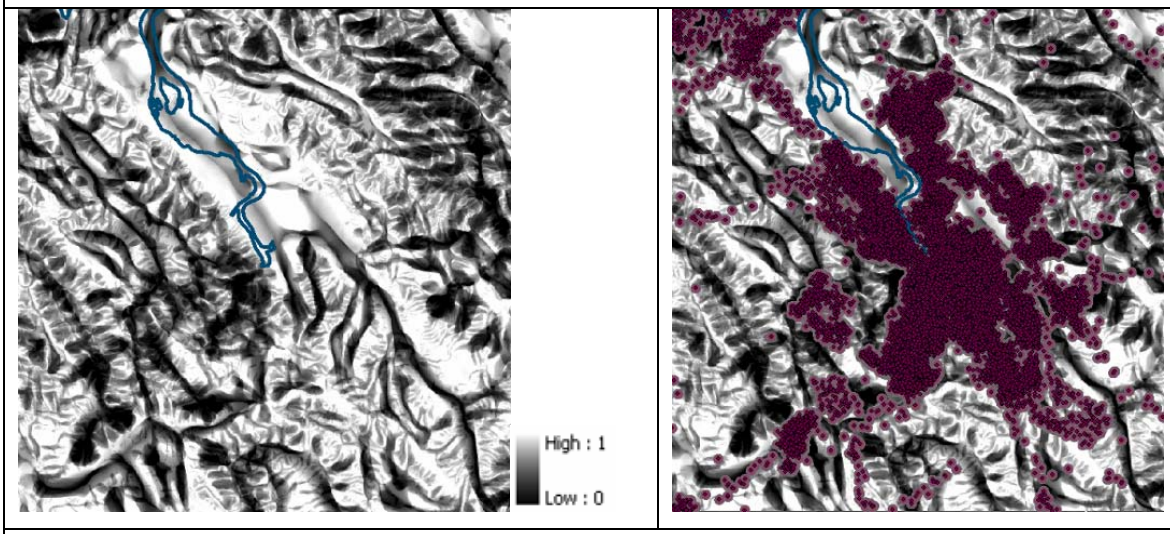


B. TAPM output – Average modelled PM_{10} concentration (cavg) at each grid cell for 2005 winter. Concentrations are in µg/m^3. Spatial resolution is 500 m.

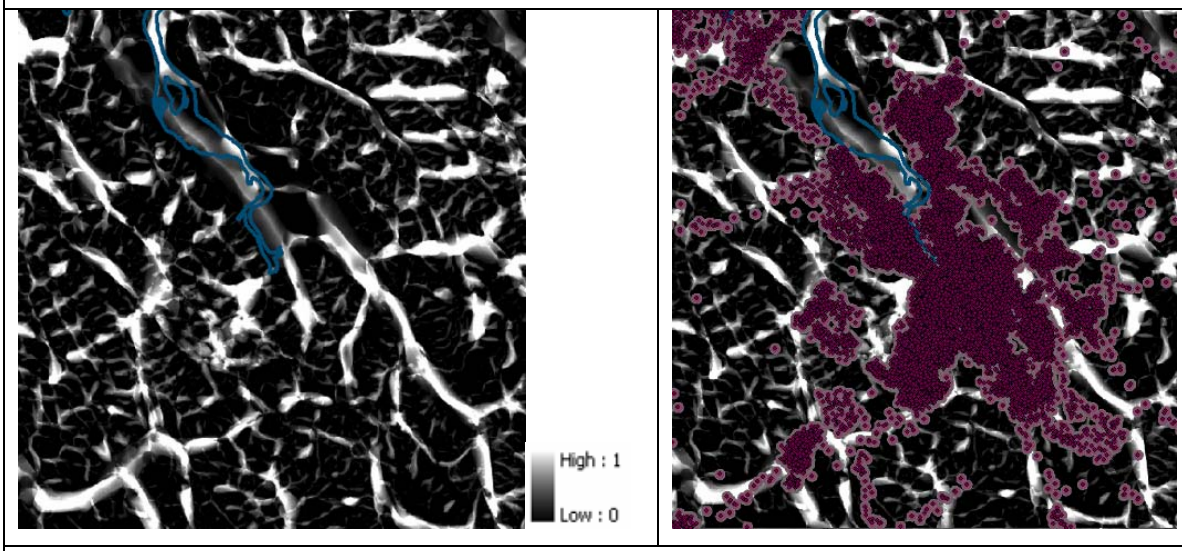
Figure 5.10 – TAPM exposure surfaces of (A) 'cmax' and (B) cavg, showing the modelled PM_{10} maximum (top) and average (bottom) concentration at each grid cell.



A. Digital elevation model, displayed with hillshade. Spatial resolution is 25 m.



B. LandSerf output – "planar". Spatial resolution is 25 m.



C. LandSerf output – "channel". Spatial resolution is 25 m.

Figure 5.11 – All terrain exposure surfaces used in the study displayed both with (right) and without (left) address points and 200 m buffers. The outline of the Tamar River is shown in blue.

TAPM cmax is also displayed in Figure 5.12 with symbology adjusted to display the areas of highest modelled particulate concentration, which is seen to be around East Launceston and Newstead.

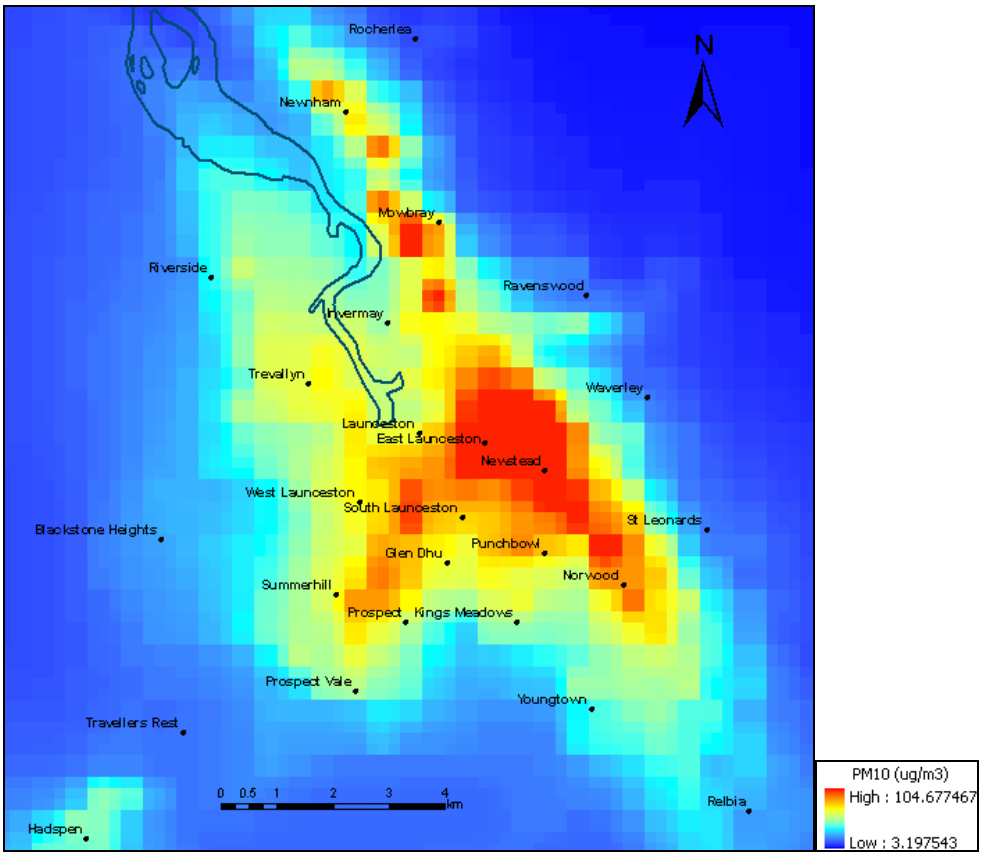


Figure 5.12 – TAPM "cmax" resampled to 200 m resolution and displayed with adjusted histogram to show areas of highest PM_{10} concentration. The areas of East Launceston and Newstead show the highest modelled particulate levels.

5.5.3 Population and disease density surfaces

Point density surfaces for total population, binary disease cases (ABB) for all winters, and binary disease cases (ABB) for winter 2005 are shown in Figure 5.13. Kernel density estimation was also applied using Hawth's spatial ecology extension for ArcGIS to generalise these patterns into 50% isopleths (i.e. the area within which 50% of the density is observed) (Beyer, 2004).

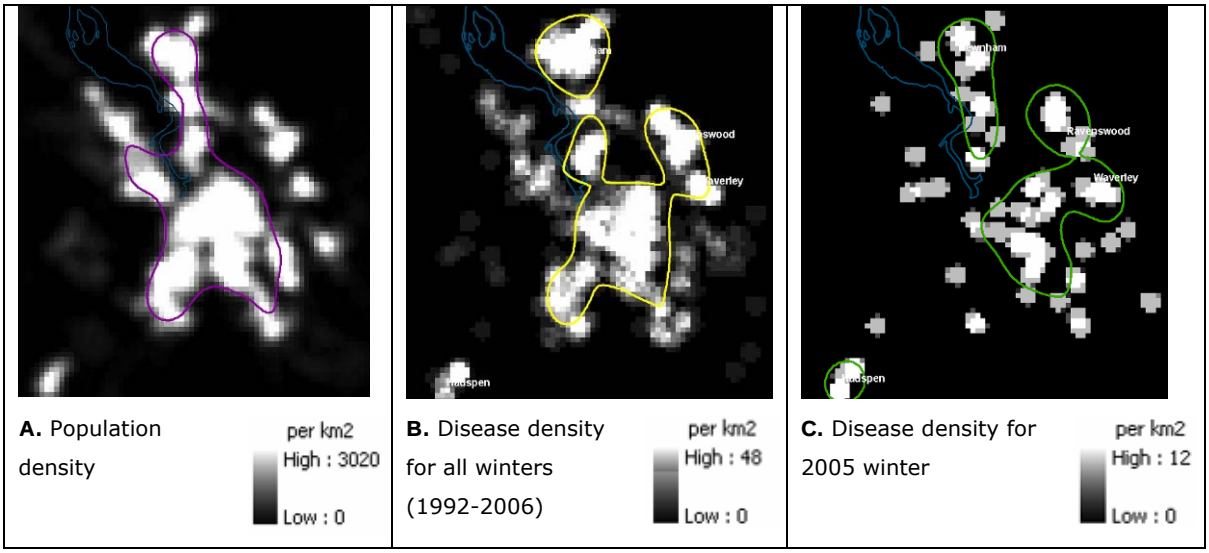


Figure 5.13 – Density distributions of (A) total population, (B) disease admissions for all winters, and (C) disease admissions for winter 2005. All are shown with 50% isopleths demarking the area within which 50% of the density distribution falls. The Tamar River outline is shown in blue. Grids are in 200 m resolution.

The 50% isopleths of total population distribution and winter 2005 disease distribution are displayed together with the TAPM cmax surface in Figure 5.14, where it is seen that the TAPM output closely matches the underlying population distribution (shown in purple). Large numbers of disease admissions were found in the areas of highest PM_{10} concentration in East Launceston and Newstead, though equally large numbers were also observed in the less populated areas around Ravenswood, Waverley and Hadspen. A three-dimensional view of this relationship between total population and 2005 winter disease admissions is displayed in Figure 5.15, where it is clearly seen that disease cases are biased towards the north eastern side of the valley from the major population centre. This image also demonstrates that the area of highest modelled PM_{10} concentration seen in Figure 5.14 – where disease incidence and population density are both high – lies at a low elevation on the valley floor.

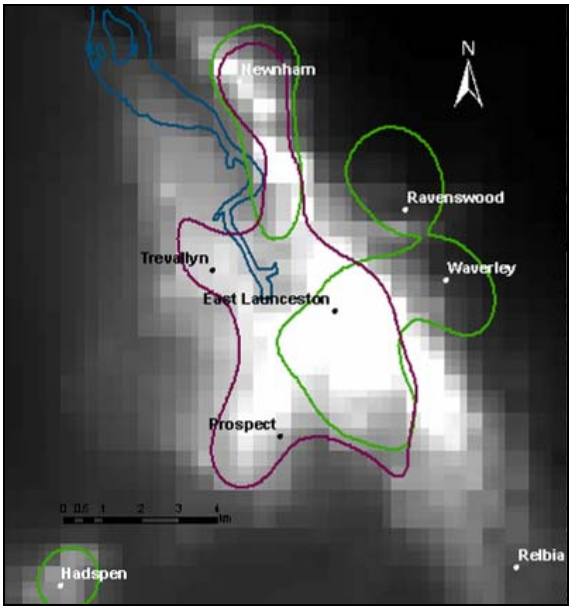


Figure 5.14 - TAPM cmax grid for 2005 winter maximum PM_{10} concentrations overlaid with 50% isopleths of 2005 winter ABB admissions (green), and total population (purple). It is seen that the TAPM output closely follows the total population distribution (upon which emission levels were based). The areas of highest PM_{10} concentration around East Launceston and Newstead (see Figure 5.12) show high disease rates, while a large proportion of disease cases also occur in the less populated areas of Ravenswood, Waverley and Hadspen. The outline of the Tamar River is shown in blue. Isopleths were derived using Hawth's kernel density estimator in Spatial Ecology extensions for ArcGIS (Beyer, 2004).

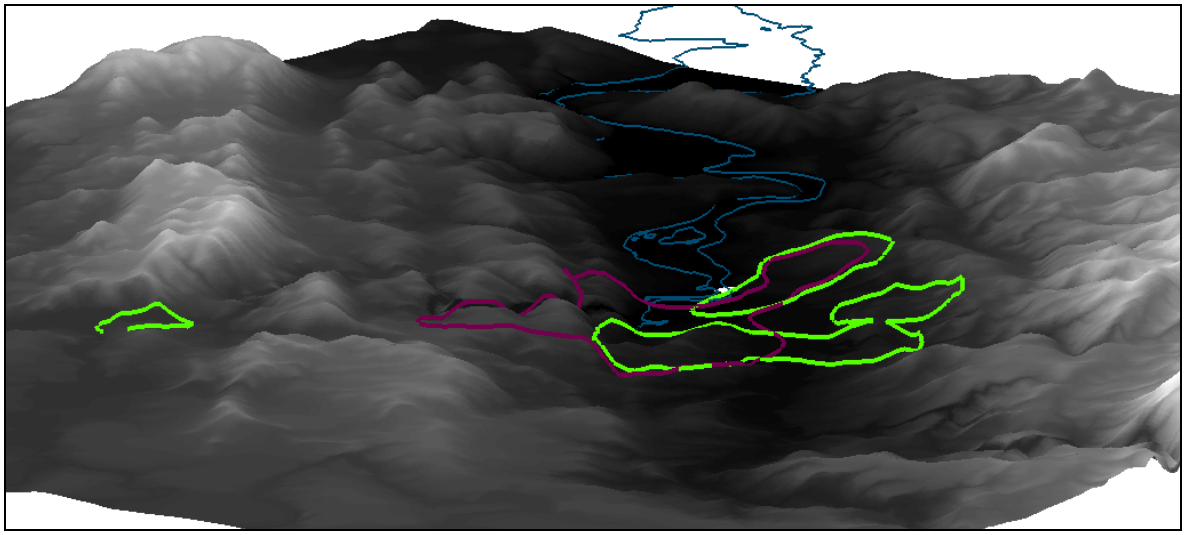


Figure 5.15- Three-dimensional view of the Launceston region seen in Figure 5.14, looking northwest through the North Esk and Tamar River valleys from Relbia. Terrain height has been exaggerated slightly for illustrative purposes; Tamar River outline is shown in blue. It is seen that a large proportion of houses with recorded winter admissions for ABB (light green) occurred in areas outside of the major population centre (purple). These were predominantly on the eastern side on the North Esk valley (on the right in this image).

5.5.4 Global statistical analysis

The distribution of address points corresponding to binary disease cases (1) and non-cases (0) were assessed against zonal statistics calculations of all exposure surfaces. The results of these global statistical analyses are summarised in Table 5-2. Results show that disease cases occurred more frequently than non-disease cases in areas classed as channels and planes; the latter relationship was significant. Disease cases also tended to be slightly more prevalent at lower elevations, though this relationship was far from significant and standard deviations were very high.

Interestingly, these analyses show that maximum and average PM_{10} concentrations predicted by TAPM (cmax) were *inversely* related to disease incidence for 2005 winter, indicating that modelled PM_{10} concentrations were higher in locations where no disease was recorded. This is consistent with the findings of the previous section that showed modelled air pollution concentrations closely followed the population distribution, while high levels of disease incidence were observed outside the major population centre. Note, however, that the standard deviations of these measurements were again very high.

Independent	Dependent	Mean	S.D.	t-test
TAPM cavg (µg/m³)	2005 winter (0)	15.71	8.54	0.175
	2005 winter (1)	13.58	9.10	
TAPM cmax (µg/m³)	2005 winter (0)	50.30	23.61	0.0027
	2005 winter (1)	42.37	25.47	
LandSerf channel	All winters (0)	0.099	0.132	0.366
	All winters (1)	0.101	0.138	
LandSerf planar	All winters (0)	0.739	0.205	0.0007
	All winters (1)	0.761	0.195	
DEM (elevation) (m ASL)	All winters (0)	75.26	55.76	0.781
	All winters (1)	73.88	50.35	
Population per house	All winters (0)	2.237	0.552	--
	All winters (1)	2.362	0.565	

Table 5-2 – Global statistical relationships between binary disease data (0 = non-cases, 1 = disease cases) for 2005 winter disease data and TAPM outputs, and for all winters disease data and all terrain-based exposure surfaces. Population per house was also measured against all winter disease data.

5.5.5 Geographically Weighted Regression

Initial outputs of Geographically Weighted Regression (GWR) provided details of the global relationships between the dependent variable (ABB admissions for all winters) and each of the independent variables. Only terrain proxies were used here (i.e. not TAPM output) because TAPM output was specific to 2005 meteorology; GWR also requires sufficient data to build a model, so disease data for *all* winters were used with terrain analysis instead of just 2005 winter data.

Variance of the global parameter estimate was compared with the variance of local estimates to give an indication of stationarity in each relationship. Specifically, if the inter-quartile range of the local parameter distribution (representing 50% of the distribution) is higher than 2 standard deviations of the global parameter distribution (which represents 68% under a normal distribution) then non-stationarity in the relationship is indicated (Fotheringham et al., 2002a). These relationships between global and local parameter estimates are summarised below in Table 5-3 where it is seen that the relationships between winter disease admissions and all independent variables exhibit non-stationarity, or local variance.

Parameter	Global (2 x standard deviations)	Local (Inter-quartile range)	Non-stationary?
Intercept	0.639	3.333	yes
LandSerf channel	0.661	2.349	yes
LandSerf planar	0.548	1.475	yes
DEM (elevation)	0.0015	0.0112	yes
Population	0.00026	0.0083	yes

Table 5-3 – Test for non-stationarity in the relationship between disease incidence for all winters and each of the independent variables tested. Non-stationarity is indicated if the inter-quartile range of the local distribution is greater than 2 standard deviations of the global distribution; all relationships in this study exhibit non-stationarity.

Graphical output of GWR analysis is displayed in Figure 5.16 where considerable local variation in relationships between winter disease incidence and exposure surfaces is demonstrated. Local variability is seen to be particularly strong in relationships between disease occurrence and all terrain features.

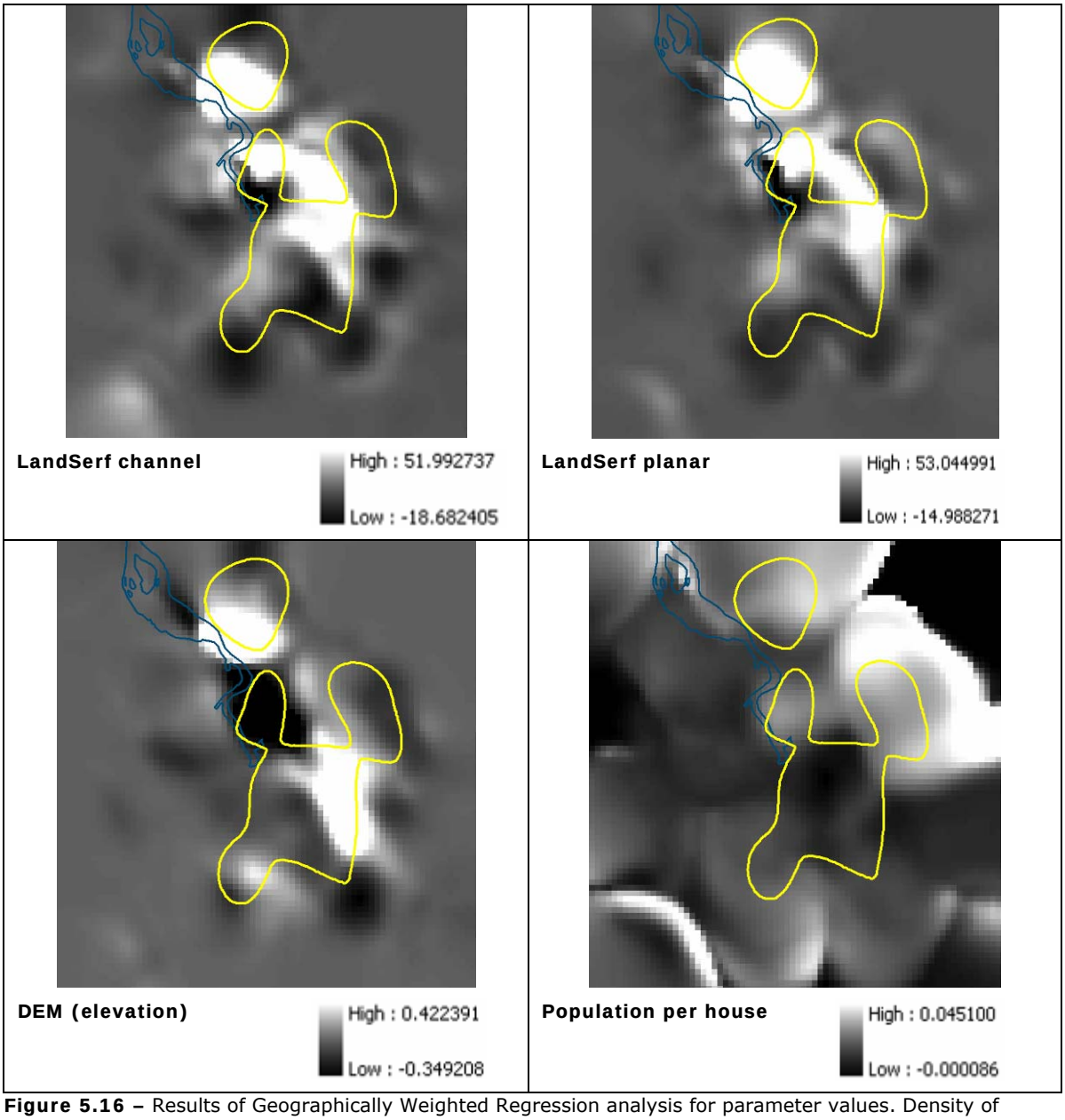


Figure 5.16 – Results of Geographically Weighted Regression analysis for parameter values. Density of disease incidence for all winters was modelled against each exposure surface, and total population. Spatial non-stationarity exists in all relationships, though most notably with the terrain analysis layers of channel, planar and elevation. Light shaded areas correspond to locations where the independent variable has a higher influence on disease incidence. 50% isopleth of all winter disease density in shown in yellow; the Tamar River outline is shown in blue.

Figure 5.16 suggests that the relative influence of terrain features (channel, planar and elevation) on disease incidence is higher in some areas than others. Most notably, the area around Ravenswood and Waverley (see Figure 5.13 for clarification) consistently showed a weak relationship (neither positive nor negative) with terrain proxies. This indicates that factors other than terrain features (and the cold air drainage these were approximating) may be contributing to the high disease rates observed in these areas.

5.6 Discussion

5.6.1 Terrain features and exposure assessment

The terrain feature types used in this study were believed to represent regions where particulate pollution would accumulate due to winter-time cold air drainage and ponding. This was based on reports that particulate air pollution drains to low-lying regions of the Tamar Valley under the stable anticyclonic conditions that predominate in winter (Lyons and expert working party, 1996; Nunez, 1991). Inverse relationships between local elevation and PM_{10} concentrations had also been observed elsewhere (Ling, 2002; Wilson, 2006). This study used 'channels' and 'planes' derived from quadratic approximation of terrain height (Wood, 1996), and elevation from a DEM as three surrogates for cold air drainage as a means of identifying areas of PM_{10} accumulations. A significant relationship was found between disease incidence and planar regions, used here to represent regions of the valley floor and hill slopes.

Given that temperature inversions in Launceston have been found to reach heights of between 80 and 300 m (Nunez, 1991), it was thought that 'channels' may show a similar pattern of correlation with disease incidence. However, the weaker (non-significant) relationship detected between disease incidence and 'channels' in this study could be explained by the findings from a study of air pollution dispersion in Christchurch, New Zealand (Wilson et al., 2006) The researchers noted that a monitoring site in their study that was *in the path of* cold air drainage experienced significant turbulence which locally dissipated the temperature inversion, thereby decreasing local pollutant levels (Wilson et al., 2006). Whether channels act as conduits for katabatic flows or reservoirs for cold air ponding may be dependent both on the scale at which they are measured and their elevation.

The very weak (non-significant) relationship detected between elevation and disease incidence in this study suggests that disease cases were only slightly more likely to occur at lower elevations. The weakness of this relationship is probably due to the fact that it is *relative* elevation, rather than elevation per se, that gives rise to conditions of cold air ponding (Sturman and Tapper, 2006). As mentioned, Ling (2002) found a strong inverse relationship between elevation and PM_{10} concentrations, though crucially, this was measured in a single, isolated valley system; the 'valley floor' in this study was 92 m ASL. Further investigations of elevation as a surrogate for cold air ponding should therefore consider a measure of relativity.

It is interesting that the observed relationships between all terrain proxies and disease incidence were weakest around the areas of Ravenswood and Waverley in the North Esk valley. This would seem to suggest that the disease incidence in these areas is not

well explained by topographic processes but that some other factor may be influencing high disease rates here. Socioeconomic confounding could be one explanation, and this point is expanded on below.

The presence of non-stationarity in relationships between disease incidence and all terrain features indicates that a more detailed spatial investigation of these relationships is warranted. The method of classifying terrain features for use as exposure surfaces in this study was reasonably subjective, and is likely to have influenced the observed relationships between terrain features and disease. As mentioned, the upper size of the search window used in LandSerf to categorise surface features was chosen somewhat arbitrarily by visually analysing the DEM. Multi-scale classification was used to minimise the impacts of choosing an inappropriate scale (Wood, 1996), though other scales could be investigated in further studies. Also, the methods used to derive the associated statistics should perhaps be reviewed in future analyses. As detailed in section 5.4.6, mean zonal statistics calculations for a 200 m radial zone around each scrambled address were used to approximate terrain conditions for all possible locations of the true address point within each zone. This may have introduced bias if the distribution of values in all zones was not Gaussian.

5.6.2 Dispersion modelling and exposure assessment

This study found a significant *inverse* relationship between disease incidence and modelled air pollution dispersion patterns derived from TAPM, indicating that modelled PM_{10} concentrations were *lower* in areas of recorded disease incidence. Three possible explanations for this finding are suggested. The most obvious, and least likely, is that there is no significant relationship between acute respiratory disease and particulate air pollution. This can quickly be discounted, however, given the voluminous body of evidence already cited to the contrary from studies worldwide, including the World Health Organisation. Included in this is the recent study by Mesaros *et al* (2007) that found a (significant) 4% increase in hospital admissions for bronchitis and bronchiolitis to the Launceston General Hospital for every 10 ug/m^3 increase in PM_{10} measured at Ti Tree Bend. And further, section 5.5.1 of this study demonstrates the strong temporal agreement between ABB admissions and measured PM_{10} concentrations from the Ti Tree Bend monitoring station.

A second possible explanation for this finding could reflect a limitation of TAPM to adequately predict pollutant concentrations in areas of low population density. As demonstrated in section 5.5.3 of this study, TAPM outputs closely matched the spatial distribution of total population density in Launceston, while some areas of disease incidence occurred outside this population centre (around Ravenswood, Waverley and Hadspen). Briggs (2005) comments that many dispersion models give sound pollution

estimates close to pollution sources but are limited in their ability to predict exposure in remote areas. TAPM assumed emissions data input as a simple proportion of housing density (i.e. 30% of population = wood burning dwellings) (Power, 2007); the findings of this study could reveal a problem with this assumption.

Dispersion modelling has rarely been used by epidemiologists for health effect assessment partly because they are poorly validated (Briggs, 2005; Jerrett et al., 2005a; Wilson, 2006). The coarse resolution of output data is also thought to be a deterrent to epidemiologists (Jerrett et al., 2005a). Those models that have been validated typically show no more than 70% agreement with measured pollution levels (Briggs, 2000). The validation and development of dispersion models for the accurate prediction of *intraurban* air pollution has therefore been identified as a priority area of research (Jerrett et al., 2005a; Wilson et al., 2004). In a preliminary validation assessment of the modelling conducted for this study, TAPM was found to significantly under-predict PM_{10} levels at Ti Tree Bend (66 μg/m^3 modelled vs. 110 μg/m^3 measured) (Power, 2007) – this shows only 60% agreement between measured and modelled concentration levels at this location. A formal validation study of TAPM also found that this model under-predicts pollutant levels in night-time stable conditions (Luhar and Hurley, 2003). Given that stable atmospheric conditions are a dominant climatic feature of the Tamar Valley in winter (Power, 2001), this issue may be a source of considerable error in this study.

Colvile and Briggs (2000) add that emissions data used in dispersion modelling are usually collected for regulatory purposes of air quality assessment and may not be suitably detailed for epidemiological applications; the validation of dispersion models is likewise usually conducted for regulatory purposes only. Also, measurements from monitoring stations, used to validate models, are usually situated in exposed positions and this may significantly underestimate pollution levels in more sheltered environments. Researchers must therefore understand the assumptions on which models are based to avoid misuse of dispersion modelling for health effect assessments. They conclude by suggesting that a closer collaboration between atmospheric scientists and epidemiologists should be encouraged (Colvile and Briggs, 2000).

In the absence of ground monitoring data from a network of monitoring sites across Launceston, it is not known whether the results found in this study are attributable to the limitations of TAPM to accurately predict pollution emissions outside of major population centres, or whether ambient pollution levels are accurately predicted by the model and some other factors are to blame for the increase in disease incidence in some areas. It is also probable that there is spatial bias in the use of wood heaters

over other home heating methods across the study area that TAPM did not account for. Given TAPM's current wide usage in Tasmania for modelling both residential and industrial pollutants, however, these results suggest that a more detailed validation study is required before TAPM can be used with confidence for health exposure estimates.

The third and final possible explanation for the findings of increased disease incidence in areas of low modelled PM_{10} concentration relates to issues of socioeconomic confounding and environmental justice, as explained below.

5.6.3 Issues of socioeconomic confounding in the Tamar Valley

The areas of Waverley and Ravenswood in the North Esk valley are among the lowest ranking socioeconomic regions in Launceston. The presence of increased disease density here in the absence of elevated (modelled) PM_{10} levels raises issues regarding possible socioeconomic confounding and environmental justice. As explained in Chapter 2, people of lower socioeconomic status may tend to experience poorer respiratory health generally due to a combination of sub-standard housing design (leading to ventilation problems), higher rates of occupational pollution exposure, higher smoking rates and generally diminished overall health (Carstairs, 2000; Hayes, 2003; Jerrett et al., 2003; Koch and Denike, 2004; Lipton et al., 2005; Maantay, 2002). In a study of air pollution and respiratory disease in Newcastle, New South Wales, Christie *et al* (1992) reported that "correlations between geographic location and respiratory admissions rates may be a manifestation of social class rather than poor air quality, although a contribution from the latter cannot be discounted".

Social and occupational factors may also offer explanations for the high levels of respiratory disease detected in George Town in the early years of the study period, and its decline in more recent years (see Chapter 4). A large number of George Town residents work in the Bell Bay industrial precinct 2 km to the south of George Town, which is a significant source of industrial pollutants. A study that accounts for occupation of participants may find that this decline in disease may be attributable to an improvement in Occupational Health and Safety standards over past decades.

On the other side of the socioeconomic divide, residents in the more affluent areas around East Launceston are likely to show lower rates of hospital admissions than might be expected for the level of ambient pollution levels in these areas. This can be attributable not only to the opposite of all factors just mentioned but, notably, because people of higher socioeconomic status tend to favour the private health care system (Giles, 1980); this study accessed only public hospital admissions records and

therefore is likely to have introduced a level of spatially influenced socioeconomic bias into the results.

The vital role of personal confounding information in studies of disease aetiology was highlighted almost a century ago. Waller and Gotway (2004) cite a study by Maxcy (1926) that found negligible spatial pattern between residential address location and disease (endemic typhus fever). Data were then correlated with occupational location and a concentration of disease was detected in the city's central business district. Maxcy then added attribute data of occupation type and found the highest disease incidence among those working in various food outlets, restaurants and green grocers, leading to a suggested link between the disease and rodent vectors via mites, fleas or lice (Waller and Gotway, 2004). Such a result argues strongly for the inclusion of personal demographic information in future studies of environmental health assessment in the Tamar Valley.

5.6.4 Methodological limitations of this study

The analysis of datasets at different spatial resolutions in this study introduced some issues of cross-level inference and ecological bias (Maheswaran and Craglia, 2004). Issues of spatial scale are central to any study linking environmental exposure to population health as the degree of perceived linkage between pollution and disease is necessarily dependent on the scale at which the study is focussed (Brindley et al., 2004; Elliott et al., 2000a; Maheswaran and Craglia, 2004; Openshaw et al., 1987; Sexton et al., 2002; Wilson et al., 2006).

The use of aggregate data instead of individual residence data introduces considerable bias to environmental exposure estimates (Zandbergen and Chakraborty, 2006) and individual level data was therefore used in this study to avoid these problems. Every effort was then taken to maintain the spatial integrity of this unique dataset by attempting to conduct all exposure analyses at the individual level. However, the 200 m random perturbation applied to individual data was accounted for by creating 200 m buffers around scrambled address points and calculating mean intersections with exposure surfaces, thereby effectively reducing the resolution of the population data to 200 m anyway. The spatial resolution of available exposure surfaces in the study also varied from 25 m for terrain surfaces to 500 m for TAPM outputs. Analysis of exposure at a coarser resolution therefore may have been warranted.

Furthermore, the results of Geographically Weighted Regression were not exploited to its full potential in this study. GRW has elsewhere been used to improve model predictions and gain a more thorough understanding of local variations in datasets (e.g. (Osborne et al., 2007). The site-specific spatial non-stationarity of relationships

between disease incidence and other factors could be incorporated into analysis of future studies.

5.7 Conclusion

The spatial investigation of interactions between winter disease incidence and exposure surfaces found a weak inverse relationship between disease incidence and elevation (i.e. disease cases were found to occur at slightly lower elevations than non-disease cases). Disease cases were also slightly more frequently associated with the terrain feature class 'channel', which was used to approximate small valley formations in the landscape. Neither of these associations was found to be significant at the global level. Winter disease incidence was, however, found to be significantly related to the 'planar' terrain features used to simulate hill slopes and the valley floor ($p = 0.0007$). Non-stationarity was detected in all relationships between disease incidence and terrain features, however, suggesting that global statistics do not adequately capture the complexity of relationships found.

A significant *inverse* relationship was found between winter disease incidence for 2005 and PM_{10} air pollution modelled using the prognostic dispersion model TAPM ($p = 0.0027$). That is, modelled PM_{10} maximum concentration levels were found to be significantly *lower* in areas of elevated disease incidence. Modelled PM_{10} average concentrations also exhibited an inverse relationship with disease incidence, though this association was not significant. Investigation of the spatial distributions of disease incidence, total population and modelled PM_{10} maximum concentrations revealed that TAPM outputs were very closely aligned with the spatial distribution of total population, while high levels of disease incidence were observed in areas of both high and low population density. An area around East Launceston and Newstead recorded the highest modelled PM_{10} concentrations and this was also an area of high disease incidence. However, the areas of Ravenswood, Waverley and Hadspen also recorded high disease incidence but very low modelled pollution levels. Validation of modelled results revealed that TAPM significantly under-predicted measured pollution levels at the Ti Tree Bend monitoring station. The failure of TAPM to adequately predict pollution dispersion in areas outside of major population densities, particularly under winter anticyclonic conditions, and various issues of socioeconomic confounding were cited as possible explanations for this result.

The findings of this section of the study further support the argument that study of relationships between environmental exposure and health outcomes requires spatial interrogation. None of these findings could have resulted from a purely temporal investigation of hospital admissions and recorded PM_{10} measurements from the Ti Tree Bend monitoring station.

6 Conclusion

6.1 Findings from this study

This investigation of respiratory disease and air pollution exposure in the Tamar Valley has revealed the spatial disparity of health risk across the valley for the first time. This study had access to a unique dataset of hospital admissions data at the individual level (i.e. point data corresponding to the de-identified residential address of patients) which enabled analysis at a very fine spatial resolution. Comparable studies have most frequently used aggregate data (i.e. address information that is grouped into some arbitrary geographic area like Census districts or postcodes). The fine resolution of individual-level data is now strongly recommended for public health investigations of this kind that seek to draw inference from the relationships between small scale environmental processes and health outcomes.

The use of this dataset raised issues of patient confidentiality and geoprivacy not normally considered in studies of coarser resolution. Confidentiality was maintained through all stages of this study. This 'de-identification' process involved the random perturbation of all address points in the Tamar Valley such that all 'houses' were shifted a random direction and distance up to 200 m from their true location prior to analysis in GIS. No confidential medical information was therefore known to the researcher (i.e. true patient address location, name, age, gender, date of birth, etc.). To further ensure the geoprivacy of individuals, all remote address locations that were believed to possibly enable the re-identification of an individual patient were removed from maps prior to publication.

Three cluster detection techniques – Kernel Density Estimation (KDE), the Getis Ord Gi* statistic and Kulldorff's spatial scan statistic – were applied to test for spatial patterns in the occurrence of Asthma, Bronchiolitis, Bronchitis and Chronic Obstructive Pulmonary Disease (COPD). KDE was used primarily as an exploratory data analysis tool, while the Gi* statistic and Kulldorff's scan statistic both gave outputs of clusters with associated statistical significance. Initial results revealed that COPD was highly clustered in a small number of address locations believed to correspond to nursing homes or aged care facilities across the valley; the location of this disease was therefore thought to be heavily biased by the location of these facilities rather than any geographic of climatic process. Subsequent analyses of 'total respiratory disease' in this study therefore included only Asthma, Bronchiolitis and Bronchitis ('ABB'). Each disease was found to display a slightly different pattern of occurrence, though many similarities were observed. George Town in the north of the Tamar Valley and areas

along the North Esk river valley in the south (i.e. Ravenswood, Waverley, Newnham and Invermay) consistently revealed areas of elevated disease risk. However, considerable variation in 'statistically significant' clusters was observed between cluster detection techniques and also between analyses conducted with the same technique at different spatial scales. Statistical inference was therefore discussed in the context of each cluster detection method.

Analysis of annual spatial variations in the occurrence of total disease (ABB) revealed that disease incidence declined generally over the study period, though most noticeably in George Town; disease rates were particularly high in George Town in the first years of the study period (between 1992 and 1995). The spatial analysis of seasonal disease rates revealed that disease patterns were relatively stable across seasons, though disease incidence generally was much greater in winter. The null hypothesis of 'complete spatial randomness' was disproved, though the relevance of this hypothesis to public health studies was also discussed.

Various exposure surfaces were created to simulate the spatial distribution of particulate air pollution in the Tamar Valley. This section of analysis was conducted on a spatial subset of the valley covering just the Launceston region, as this was the area for which air pollution modelling data was available. Two surfaces of modelled air pollution concentrations were derived from TAPM ('The Air Pollution Model'), which is a prognostic dispersion model currently in common use in Tasmania. These predictions were primarily based on estimated wood heater emission rates, synoptic scale meteorological data and terrain height for simulation of ground level wind fields. Three exposure surfaces were also created based on terrain feature classification. These surfaces simulated terrain depressions such as valleys and channels which previous studies have found give rise to increased levels of particulate air pollution in winter through the process of cold air drainage.

Spatial relationships between all exposure surfaces and winter disease incidence (ABB) were examined in detail. Only winter disease cases were used in this section of analysis to specifically investigate the relationship between disease incidence and the effects of the atmospheric temperature inversions known to predominate in the Tamar Valley over winter months. A weak relationship was found between disease location and elevation, suggesting that disease cases were slightly more likely to occur at lower elevations than non disease cases. Disease events were also found to be slightly more prevalent in the terrain features used to simulate channels and small valleys, though no significant relationship was found. A 'significant' relationship was found between disease cases and the planar regions of the valley floor and hill slopes; issues of statistical inference were again discussed. Geographically Weighted Regression was

used to investigate local variations in the strength of relationship between disease rates and terrain features. This analysis indicated significant 'non-stationarity' in all relationships, indicating that the influence of terrain features on disease was not constant across the study area.

A significant *inverse* relationship was found between disease incidence and the modelled air pollution surface of maximum PM_{10} concentration produced by TAPM, suggesting that, on average, modelled air pollution levels were *lower* in areas of high disease incidence. The areas of highest modelled air pollution (around East Launceston and Newstead) were also areas of high disease incidence, though the less populated areas of Waverley, Ravenswood and Hadspen showed high disease incidence and very low modelled PM_{10} concentrations. The failure of TAPM to adequately predict pollution concentrations, both in areas of low population density and under stable atmospheric conditions, and various issues of socioeconomic confounding were discussed as possible reasons for this finding.

The relationship between environmental air pollution and disease incidence is an intrinsically spatial relationship. None of the results of this study could have resulted from a purely global (non-spatial) investigation of hospital records and pollution levels from a single monitoring site. The findings of this study therefore argue strongly for the spatial analysis of spatial processes such as this. In addition, the inherent uncertainties associated with all public health studies were addressed in this study. Various limitations associated with disease mapping and cluster detection were explored, particularly relating to inference drawn from the varied 'statistical significance' of different cluster results. The importance of applying more than one technique was highlighted by this process, though doubt remains over the most suitable method for detecting clusters of disease. Substantial limitations were also encountered in the attempt to link disease incidence with pollution exposure surfaces. This area of study is well recognised as a source of considerable uncertainty in public health research, and several lines of reasoning were followed to explain the results found in this study. Methodological limitations of linking datasets of different spatial resolutions were also discussed.

6.2 Recommendations for further research

- There is a great need generally for guidelines on the standardised methods for both the detection and reporting of disease clusters.

- National guidelines for the ethical reporting of spatial public health information are required. This should include separate guidelines for the reporting of aggregate and individual-level information.

- This study has highlighted the need for a detailed spatial survey of *intraurban* air pollution dispersion in Launceston. This should ideally include separate measurements of PM_{10} and $PM_{2.5}$, as considerable evidence now indicates that the spatial variations between particles of different sizes changes across the urban landscape.

- This study has also identified the need to validate TAPM for public health applications generally. Epidemiologists require air pollution exposure information at a higher level of accuracy than that which generally exists for environmental monitoring and policy requirements. The potential limitations of this model to accurately predict pollution dispersion in areas of low population density are of considerable concern for public health applications. The usefulness of this model in the Tamar Valley generally should also be assessed against its ability to accurately predict pollution dispersion under stable atmospheric conditions.

- Future investigations of the relationship between respiratory disease and air pollution should include relevant demographic information to account for socioeconomic confounders.

- Air pollution in Launceston has improved in recent years as a result of a decrease in residential wood heating and an increased reliance on electric and gas heating. With electricity prices currently predicted to rise from next year, however, trends in both air pollution concentrations and respiratory disease rates should be closely monitored.

References

Aamodt G, Samuelson SO and Skrondal A (2006) A simulation study of three methods for detecting disease clusters. *International Journal of Health Geographics* **5**(15).

ABC Northern Tasmania (2004) Launceston - A dirty old town or paradise in a shroud?, in *ABC Northern Tasmania*.

Abrams AM and Kleinman KP (2007) A SaTScan macro accessory for cartography (SMAC) package implemented with SAS software. *International Journal of Health Geographics* **6**(6):?

Abramson M (2001) Occupational and environmental causes of respiratory disease. *Australasian Epidemiologist* **8**(1):32-35.

Ackermann-Liebrich U, Leuenberger P, Schwartz J, Schindler C, Monn C, Bolognini G, Bongard JP, Brandli O, Domenighetty G, Elsasser S, Grize L, Karrer W, Keller R, Keller-Wassidlo H, Kunzli N, Martin BW, Medici TC, Perruchoud AP, Schoni MH, Tschopp JM, Villiger B, Wuthrich B, Zellweger JP, Zemp E and Team S (1997) Lung function and long term exposure to air pollutants in Switzerland. *American Journal of Respiratory and Critical Care Medicine* **155**:122-129.

AIHW (2000) Autralian Hospital Statistics 1998-99, in *http://wwwaihwgovau/publications/health/ahs98-9/ahs98-9-c01pdf* (Australian Institute of Health and Welfare ed).

AIHW (2001) Australian Hospital Statistics 1999-2000, in *http://wwwaihwgovau/publications/hse/ahs99-00/ahs99-00-c01pdf* (Australian Institute of Health and Welfare ed).

Armstrong M, Rushton G and Zimmerman D (1999) Geographically masking health data to preserve confidentiality. *Statistics in Medicine* **18**(5):497-525.

Arnold RA, Diamond ID and Wakefield J (2000) The use of population data in spatial epidemiology, in *Spatial Epidemiology: Methods and Applications* (Elliott P, Wakefield J, Best NG and Briggs D eds) pp 30-50, Oxford University Press, London.

Australian Bureau of Statistics (2001) 2028.6 - Census of Population and Housing: Launceston Suburbs, 2001, Australian Bureau of Statistics.

Australian Bureau of Statistics (2007) 3218.0 Regional Population Growth.

Ayers GP, Keywood MD, Gras JL, Cohen D, Garton D and Bailey GM (1999) Chemical and physical properties of Australian fine particles: A pilot study. Report prepared for the Environment Protection Group, Environment Australia, June 1999.

Bell ML and Davis DL (2001) Reassessment of the lethal London fog of 1952: Novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environmental Health Perspectives* **109**(Supplement 3):389-394.

Besag J and Newell J (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A, Statistics in Society* **154**(1):143-155.

Beyer HL (2004) Hawth's Analysis Tools for ArcGIS. Available at http://www.spatialecology.com/htools. .

Boulos MNK, Cai Q, Padget JA and Rushton G (2006) Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. *Journal of Biomedical Informatics* **39**:160-170.

Briggs D (2000) Exposure Assessment, in *Spatial Epidemiology: Methods and Applications* (Elliott P, Wakefield J, Best N and Briggs D eds), Oxford University Press, New York.

Briggs D (2005) The role of GIS: Coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health, Part A* **68**:1243-1261.

Brindley P, Maheswaran R, Pearson T, Wise S and Haining RP (2004) Using modeled outdoor air pollution data for health surveillance, in *GIS in Public Health Practice* (Maheswaran R and Craglia M eds) pp 125-149, CRC Press, Boca Raton, Florida.

Brook JR, Graham L, Charland JP, Cheng Y, Fan X, Lu G, Li SM, Lillyman C, MacDonald P, Caravaggio G and MacPhee JA (2007) Investigation of the motor vehicle exhaust contribution to primary fine particle organic carbon in urban air. *Atmospheric Environment* **41**:119-135.

Cakmak S, Burnett RT, Jerrett M, Goldberg MS, Pope III CA and Ma R (2003) Spatial regression models for large-cohort studies linking community air pollution and health. *Journal of Toxicology and Environmental Health, Part A* **66**:1811-1823.

Carstairs V (2000) Socio-economic factors at areal level and their relationship with health, in *Spatial Epidemiology: Methods and Applications* (Elliott P, Wakefield JC, Best NG and Briggs D eds) pp 51-67, Oxford University Press, London.

Charlton M, Fotheringham A and Brunsdon C (2003) GWR 3: Software for Geographically Weighted Regression.

Chen L, Verrall K and Tong S (2006) Air particulate pollution due to bushfires and respiratory hospital admissions in Brisbane, Australia. *International Journal of Environmental Health Research* **16**(3):181-191.

Christie D, Spencer L and Senthilselvan A (1992) Air quality and respiratory disease in Newcastle, New South Wales. *Medical Journal of Australia* **156**:841-844.

Colvile R and Briggs D (2000) Dispersion modelling, in *Spatial Epidemiology: Methods and Applications* (Elliott P, Wakefield J, Best N and Briggs D eds), Oxford University Press, London.

Corburn J (2007) Urban land use, air toxics and public health: Assessing hazardous exposures at the neighbourhood scale. *Environmental Impact Assessment Review* **27**:145-160.

Curtis AJ, Mills JW and Leitner M (2006) Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics* **5**:44.

Cuzick J and Edwards R (1990) Spatial clustering for inhomogeneous populations. *journal of the Royal Statistical Society Series B, Methodological* **52**(1):73-104.

De Angelo L (2006) London smog disaster, England, (Black B ed).

DEH (2004) State of the Air: National ambient air quality status and trends report 1991-2001, Department of the Enviroment and Heritage.

DEH (2005a) National Standards for Criteria Air Pollutants in Australia: Air quality fact sheet, Department of the Environment and Water Resources.

DEH (2005b) Woodheaters in Launceston - Impacts on Air Quality, p 61, CSIRO Atmospheric Research, Aspendale, Victoria.

Diamond I (1997) Population counts in small areas, in *Geographical and environmental epidemiology* (Elliott P, Cuzick J, English D and Stern R eds), Oxford University Press, New York.

Diggle PJ (2000) Overview of statistical methods for disease mapping and its relationship to cluster detection, in *Spatial Epidemiology: Methods and Applications* (Elliott P, Wakefield J, Best NG and Briggs D eds) pp 87-103, Oxford Medical Publications, London.

Dockery DW and Pope CAI (1994) Acute respiratory effects of particulate air pollution. *Annu Rev Public Health* **15**:107-132.

DPIWE (2007) Air Moitoring Data: Ti Tree Bend monitoring station, Launceston.

Durand M and Wilson JG (2006) Spatial analysis of respiratory disease on an urbanized geothermal field. *Environmental Research* **101**:238-245.

Elliott P and Wakefield J (2000) Bias and confounding in spatial epidemiology, in *Spatial Epidemiology: Methods and Applications* (Elliott P, Wakefield J, Best NG and Briggs D eds) pp 68-84, Oxford Medical Publications, London.

Elliott P, Wakefield J, Best NG and Briggs D (2000a) Spatial epidemiology: methods and applications, in *Spatial epidemiology: methods and applications* (Elliott P, Wakefield J, Best NG and Briggs D eds) pp 1-14, Oxford Universtiy Press, London.

Elliott P, Wakefield JC, Best NG and Briggs D (2000b) *Spatial Epidemiology: Methods and Applications*. Oxford University Press, London.

Fefferman N, O'Neil E and Naumova E (2005) Confidentiality and confidence: Is data aggregation a means to achieve both? *Journal of Public Health Policy* **26**(4):430-450.

Fisher P, Wood J and Cheng T (2004) Where is Helvellyn? Fuzziness of multi-scale landscape morphology. *Transactions of the Institute of British Geographers* **29**(1):106-128.

Forastiere F (2004) Fine particles and lung cancer. *Occupational and Environmental Medicine* **61**:797-798.

Fotheringham A, Brunsdon C and Charlton M (2002a) *Geographically Weighted Regression: the analysis of spatially varying relationships*. John Wiley & Sons Ltd, West Sussex.

Fotheringham AS, Brunsdon C and Charlton M (2002b) *Quantitative Geography: Perspectives on Spatial Data Analysis*. SAGE Publications Ltd, London.

Fusco D, Forastiere F, Michelozzi P, Spadea T, Ostro B, Arca M and Perucci CA (2001) Air pollution and hospital admissions for respiratory conditions in Rome, Italy. *European Respiratory Journal* **17**:1143-1150.

Getis A and Ord JK (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis* **24**(3):189-209.

Getis A and Ord JK (1996) Local spatial statistics: an overview, in *Spatial analysis: modeling in a GIS environment* (Longley P and Batty M eds), John Wiley and Sons, Ltd, New York.

Giles G (1980) The geographical and biometeorological correlates of childhood asthma morbidity in Tasmania, Department of Geography, University of Tasmania, PhD Thesis, Hobart.

Gilliland F, Avol E, Kinney P, Jerrett M, Dvonch T, Lurmann F, Buckley T, Breysse P, Keeler G, de Villiers T and McConnell R (2005) Air pollution exposure assessment for epidemiological studies of pregnant women and children: lessons learned from the Centres for Children's Health and Disease Prevention Research. *Environmental Health Perspectives* **113**(10):1447-1454.

Goldberg MS, Burnett RT, Yale J-F, Valois M-F and Brook JR (2006) Associations between ambient air pollution and daily mortality among persons with diabetes and cardiovascular disease. *Environmental Research* **100**:255-267.

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

Greenland S and Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* **15**(3):413-419.

Hales S, Salmond C, Town GI, Kjellstrom T and Woodward A (1999) Daily mortality in relation to weather and air pollution in Christchurch, New Zealand. *Australian and New Zealand Journal of Public Health* **24**(1):89-91.

Hayes MV (2003) "Ecological confounders" in the context of a spatial analysis of the air pollution-mortality relationship. *Journal of Toxicology and Environmental Health, Part A* **66**:1779-1782.

Hurley P (1999) The Air Pollution Model (TAPM) Version 1: Technical Description and Examples, in *Technical Papers* (Research CA ed), Aspendale, Victoria.

Hurley P (2005) The Air Pollution Model (TAPM) Version 3. Part 1: Technical Description, in *Technical Paper 71* (Research CA ed), Aspendale, Victoria.

Ito K, Kinney P and Thurston G (1995) Variations in PM10 concentrations within 2 metropolitan areas and their implications for health effects analyses. *Inhalation Toxicology* **7**:735-745.

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J and Giovis C (2005a) A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology* **15**(2):185-204.

Jerrett M, Burnett RT, Ma R, Pope III CA, Krewski D, Newbold KB, Thurston G, Shi Y, Finkelstein N, Calle EE and Thun MJ (2005b) Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology* **16**(6):727-736.

Jerrett M, Burnett RT, Willis A, Krewski D, Goldberg MS, DeLuca P and Finkelstein N (2003) Spatial analysis of the air pollution-mortality relationship in the context of ecological confounders. *Journal of Toxicology and Environmental Health, Part A* **66**:1735-1777.

Jerrett M and Finkelstein M (2005) Geographies of risk in studies linking chronic air pollution exposure to health outcomes. *Journal of Toxicology and Environmental Health, Part A* **68**:1207-1242.

Kelsall JE and Diggle PJ (1998) Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics* **47**(4):559-573.

Kingham, Durand M, Aberkane, Harrison, Wilson JG and Epton (2006) Winter comparison of TEOM, MiniVol and Dust Trak PM10 monitors in a woodsmoke environment. *Atmospheric Environment* **40**(2):338-347.

Koch T and Denike K (2004) Medical mapping: The revolution in teaching - and using - maps for the analysis of medical issues. *Ther Journal of Geography* **103**(2):76-85.

Kulldorff M (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26**(6):1481-1496.

Kulldorff M (2006) SatSCan User Guide for version 7.0, p 92, http://www.satscan.org/.

Kulldorff M and Nagarwalla N (1995) Spatial disease clusters: detection and inference. *Statistics in Medicine* **14**:799-810.

Kwan M-P, Casas I and Schmitz BC (2004) Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica* **39**(2):15-28.

Leem J-H, Kaplan BM, Shim YK, Pohl HR, Gotway CA, Bullard SM, Rogers JF, Smith MM and Tylenda CA (2006) Exposures to air pollutants during pregnancy and preterm delivery. *Environmental Health Perspectives* **114**(6):905-910.

Liao D, Peuquet DJ, Duan Y, Whitsel EA, Dou J, Smith RL, Lin H-M, Chen J-C and Heiss G (2006) GIS approaches for the estimation of residential-level ambient PM concentrations. *Environmental Health Perspectives* **114**(9):1374-1380.

Ling B (2002) Woodsmoke derived particulate air pollution in Lenah Valley, in *School of Geography and Environmental Studies* p 103, University of Tasmnia, Hobart.

Lipton R, Banerjee A, Dowling KC and Treno AJ (2005) The geography of COPD hospitalization in California. *COPD: Journal of Chronic Obstructive Pulmonary Disease* **2**:435-444.

Luhar AK and Hurley PJ (2003) Evaluation of TAPM, a prognostic meteorological and air pollution model, using urban and rural point-source data. *Atmospheric Environment* **37**:2795-2810.

Lyons L and expert working party (1996) Air pollution, environmental health and respiratory diseases: Launceston and Upper Tamar Valley (1991-1994), Launceston City Council, Launceston.

Maantay J (2002) Mapping environmental injustices: pitfalls and potential of geographic information systems in assessing environmental health and equity. *Environmental Health Perspectives* **110(Supplement 2)**:161-171.

Maheswaran R and Craglia M (2004) *GIS in Public Health Practice*. CRC Press, Boca Raton, Florida.

Maheswaran R and Haining R (2004) Basic issues in geographical analysis, in *GIS in Public Health Practice* (Maheswaran R and Craglia M eds), CRC Press, Boca Raton, Florida.

McGowan JA, Hider PN, Chacko E and Town GI (2002) Particulate air pollution and hospital admissions in Christchurch, New Zealand. *Australian and New Zealand Journal of Public Health* **26**(1):23-29.

Medina S, Plasencia A, Ballester F, Mucke HG and Schwartz J (2004) Apheis: public health impact of PM10 in 19 European cities. *Journal of Epidemiology and Community Health* **58**:831-836.

Mesaros D, Wood-Baker R, FitzGerald D, Walters EH and Markos J (2007) The relationship between particle air pollution and admissions for respiratory disease in the Tamar Valley. *Respirology* **12 (Suppl 1): A38**.

Mindell J and Barrowcliffe R (2005) Linking environmental effects to health impacts: a computer modelling approach for air pollution. *Journal of Epidemiology and Community Health* **59**:1092-1098.

Monn C (2001) Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone. *Atmospheric Environment* **35**:1-32.

Moolgavkar SH (2000) Air pollution and hospital admisions for chronic obstructive pulmonary disease in three metropolitan areas in the United States. *Inhalation Toxicology* **12(Supplement 4)**:75-90.

Nunez M (1991) Tethered balloon soundings for the Launceston Region - a pilot project, Department of Geography and Environmental Studies, University of Tasmania, Hobart.

Openshaw S, Charlton M, Wymer C and Craft A (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* **1**(4):335-358.

Ord JK and Getis A (1995) Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* **27**(4):286-309.

Osborne P, Foody G and Suarez-Seoane S (2007) Non-staionarity and local approaches to modelling the distributions of wildlife. *Diversity and Distributions* **13**(3):313-323.

Oyana TJ, Rogerson P and Lwebuga-Mukasa JS (2004) Geographic clustering of adult asthma hospitalization and residential exposure to pollution at a United States-Canada border crossing. *American Journal of Public Health* **94**(7):1250-1257.

Pearce DC (2002) Spatial modelling of the relationship between respiratory admissions and ambient air pollution, in *School of Information Technology and Mathematical Sciences* p 132, University of Ballarat, Ballarat.

Peel JL, Metzger KB, Klein M, Flanders WD, Mulholland JA and Tolbert PE (2006) Ambient air pollution and cardiovasular emergency department visits in potentially sensitive groups. *American Journal of Epidemiology* **165**(6):625-633.

Pope III CA (2000) Epidemiology of fine particulate air pollution and human health: Biologic mechanisms and who's at risk? *Environmental Health Perspectives Supplements* **108**(S4):713-724.

Power M (2001) Air pollution dispersion within the Tamar Valley, in *School of Geography and Environemental Studies* p 398, University of Tasmania, Hobart.

Power M (2007) Yesterday, today and tomorrow: A modelling approach to predicting changing woodsmoke concentrations in Launceston, p 5, Environment Division, Department of Tourism, Arts and the Environment.

Quinn M (1997) Confidentiality, in *Geographical and environmental epidemiology: Methods for small-area studies* (Elliott P, Cuzick J, English D and Stern R eds), Oxford University Press, New York.

Sabel C and Loytonen (2004) Clustering of Disease, in *GIS in Public Health Practice* (Maheswaran R and Craglia M eds), CRC Press, Boca Raton, Florida.

Sahsuvaroglu T and Jerrett M (2007) Sources of uncertainty in calculating mortality and morbidity attributable to air pollution. *Journal of Toxicology and Environmental Health, Part A* **70**:243-260.

Salvaggio JE (1994) Inhaled particles and respiratory disease. *Journal of Allergy and Clinical Immunology* **94**:304-309.

SAS Institute Inc. (2003) JMP User's Guide, Cary, NC, USA.

Schwartz J, Spix C, Touloumi G, Bacharova L, Barumamdzadeh T, le Tertre A, Piekarksi T, Ponce de Leon A, Ponka A, Rossi G, Saez M and Schouten JP (1996) Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health* **50**(Suppl 1):S3-S11.

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

Sexton K, Waller LA, McMaster RB, Maldonado G and Adgate JL (2002) The importance of spatial effects for environmental health policy and research. *Human and Ecological Risk Assessment* **8**(1):109-125.

Sheppard L, Levy D, Norris G, Larson TV and Koenig JQ (1999) Effects of ambient air pollution on nonelderly asthma hospital admissions in Seattle, Washington, 1987-1994. *Epidemiology* **10**(1):23-30.

Silverman B (1986) *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Simpson RW, Williams G, Petroeschevsky A, Morgan G and Rutherford S (1997) Associations between outdoor air pollution and daily mortality in Brisbane, Australia. *Archives of Environmental Health* **52**(6):442-454.

Smeal A (1998) Katabatic winds and particulate concentrations in Glenorchy, in *School of Geography and Environmental Studies* p 105, University of Tasmania, Hobart.

Snow J (1854) On the mode of communication of cholera.

Spronken-Smith RA, Sturman AP and Wilton EV (2002) The air pollution problem in Christchurch, New Zealand - progress and prospects. *Clean Air* **36**(1):23-29.

Stedman JR, A J Kent, S Grice, T J Bush, R G Derwent (2007) A consistent method for modelling PM_{10} and $PM_{2.5}$ concentrations across the United Kingdom in 2004 for air quality assessment. *Atmospheric Environment* **41**:161-172.

Sturman A and Tapper N (2006) *The weather and climate of Australia and New Zealand*. Oxford University Press, Melbourne.

theLIST (2007) Address Points Dataset, TasMap, Department of Primary Industries and Water, Hobart, Tasmania.

Todd JJ, Saxby W, Prasad D, Wilson C and Kinrade P (1997) Residential and local sources of air pollution in Australia, in *Inquiry into Urban Air Pollution in Australia* (Engineering TGotAAoTSa ed), Carlton South, Vic.

Ulirsch GV, Ball LM, Kaye W, Shy CM, Lee CV, Crawford-Brown D, Symons M and Holloway T (2007) Effect of particulate matter air pollution on hospital admissions and medical visits for lung and heart disease in two southeast Idaho cities. *Journal of Exposure Science and Environmental Epidemiology*(2007):1-10.

van de Kassteele J, Koelemeijer R, Dekkers A, Schaap M, Homan C and Stein A (2006) Statistical mapping of PM10 concentrations over Western Europe using secondary information from dispersion modelling and MODIS satellite observations. *Stochastic Environmental Research and Risk Assessment* **21**:183-194.

Venables W, Smith D and the R Development Core Team (2006) An Inrtoduction to R: Notes on R: A programming environment for data analysis and graphics Version 2.5.1 (2007-06-27).

Wakefield J and Shaddick G (2006) Heath-exposure modeling and the ecological fallacy. *Biostatistics* **7**(3):438-455.

Waller L and Gotway C (2004) *Applied Spatial Satistics for Public Health Data*. Wiley-Interscience, New Jersey.

Weng Q and Yang S (2006) Urban air pollution patterns, land use, and thermal landscape: an examination of the linkage using GIS. *Environmental Modeling and Assessment* **117**:463-489.

Wheeler DC (2007) A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. *International Journal of Health Geographics* **6**(13).

WHO (2005) World Health Organisation air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: Global update 2005: Summary of risk assessment, pp 1-21, WHO, Geneva, Switzerland.

WHO (2006a) Use of the air quality guidelines in protecting public health: a global update, in *Fact Sheet Number 313*.

WHO (2006b) World Health Organisation challenges world to improve air quality: Stricter air pollution standars could reduce deaths in polluted cities by 15%.

WHO (2007) About the Public Health Mapping and GIS programme, in: http://www.who.int/health_mapping/about/en/, (World Health Organization ed).

Wilson GJ (2006) Spatial variability of intraurban particulate air pollution : Epidemiological implications and applications. PhD thesis., University of Canterbury, New Zealand.

Wilson JG, Kingham S, Pearce and Sturman A (2004) A review of intraurban variations in particulate air pollution: Implications for epidemiological resesarch. *Atmospheric Environment* **39**(34):6444-6462.

Wilson JG, Kingham S and Sturman A (2006) Intraurban variations of PM10 air pollution in Christchurch, New Zealand: Implications for epidemiological studies. *Science of the Total Environment* **367**(2-3):559-572.

Wilson JG and Zawar-Reza P (2006) Intraurban-scale dispersion modelling of particulate matter concentrations: applications for exposure estimates in cohort studies. *Atmospheric Environment* **40**(6):1053-1063.

Wisconsin Department of Health and Family Services (2004) Comparing causes of death between years: Accounting for the change from ICD-9 to ICD-10, Wisconsin Department of Health and Family Services.

Wood J (1996) The Geomorphological Characterisation of Digital Elevation Models, PhD Thesis, University of Leicester, UK, http://www.soi.city.ac.uk/~jwo/phd, London.

Wood J (2005) LandSerf v 2.2, at http://www.landserf.org Department of Information Science, City University London., London.

Wordley J, Walters S and Ayres JG (1997) Short term variations in hospital admissions and mortality and particulate air pollution *Occupational and Environmental Medicine* **54**:108-116.

Yunesian M, Asghari F, Vash JH, Forouzanfar MH and Farhud D (2006) Acute Symptoms related to air pollution in urban areas: a study protocol. *BMC Public Health* **6**:218-222.

Zandbergen PA and Chakraborty J (2006) Improving environmental exposure analysis using cumulative distribution functions and individual geocoding. *International Journal of Health Geographics* **5**:23-37.

Zeger SL, Thomas D, Dominici F, Samet J, Schwartz J, Dockery DW and Cohen A (2000) Exposure measurement error in time-series studies of air pollution: Concepts and consequences. *Environmental Health Perspectives* **108**(5):419-426.