Spatial Modelling of Woodsmoke Concentrations and Health Risk Associated with Residential Wood Burning

by

Christy Lightowlers
B.Sc., University of Victoria, 2000

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Geography

Supervisory Committee

Spatial Modelling of Woodsmoke Concentrations and Health Risk Associated with Residential Wood Burning

by

Christy Lightowlers
B.Sc., University of Victoria, 2000

Supervisory Committee

Dr. C. Peter Keller, Department of Geography
Supervisor

Dr. Trisalyn Nelson, Department of Geography
Departmental Member

Dr. Andrew Kmetic, Department of Aboriginal Health
Outside Member

Abstract

Supervisory Committee
Dr. C. Peter Keller, Department of Geography
Supervisor
Dr. Trisalyn Nelson, Department of Geography
Departmental Member
Dr. Andrew Kmetic, Department of Aboriginal Health
Outside Member

Within the context of global climate change and soaring energy prices, people are searching for inexpensive and renewable sources of energy; therefore, burning wood for home heating is increasing. Woodsmoke contains substances known to harm human health and is a major contributor to air pollution in many parts of the world; yet there is limited research into the health effects of woodsmoke and existing research suffers from methodological constraints. As a result, there is interest in producing robust woodsmoke exposure estimates for health research and air quality management purposes. Studying health and the environment is inherently spatial; however, research related to air pollution and health tends to be aspatial. As investigators begin to understand the influence of spatial processes on research findings, the importance of adopting a spatial approach to modelling exposure and health risk is becoming apparent. This thesis describes a spatially explicit model for predicting fine particulate matter ($PM_{2.5}$) attributable to woodsmoke from residential heating in Victoria, British Columbia, Canada. Spatially resolved measurements of $PM_{2.5}$ were collected for 32 evenings during the winter heating seasons of 2004/05, 2005/06, 2006/07 using a nephelometer installed in a passenger vehicle. Positional data were collected concurrently using a Global Positioning System (GPS). Levoglucosan, a chemical unique to woodsmoke, was measured to confirm the presence of woodsmoke in the measured $PM_{2.5}$. The spatial scale for the analysis of woodsmoke data was determined using semivariograms to identify the maximum distance of spatial dependence in the data which typically occurred near 2700m. Different spatial approaches for modelling woodsmoke concentrations were evaluated both qualitatively in terms of transferability, meeting statistical assumptions, and potential for exposure misclassification; and quantitatively to assess the association between the model's

predicted $PM_{2.5}$ concentrations and observed $PM_{2.5}$. The baseline model characterized exposure based on the $PM_{2.5}$ value from the closest fixed monitor (R=0.51, α=0.05). The Krigged model produced a seasonal average surface based on nephelometer measurements and showed the weakest performance (R=0.25, α=0.05). The regression models predicted concentrations of woodsmoke based on predictor variables available from census data, typically used in health research, and spatial property assessment data (SPAD), an underused data source at a finer spatial resolution. Different approaches to regression modelling were investigated. A regression model already developed for Victoria performed the best quantitatively (R=0.84, α=0.05); however, qualitative considerations precluded it from being selected as an appropriate model. A quantitatively (R=0.62, α=0.05) and qualitatively robust regression model was developed using SPAD (M6). SPAD improved the spatial resolution and model performance over census data. Removing spatial and temporal autocorrelation in the data prior to modelling produced the most robust model as opposed to modelling spatial effects post regression. A Bayesian approach to M6 was applied; however, model performance remained unchanged (R=0.62, α=0.05). The spatial distribution of susceptibility to health problems associated with woodsmoke was derived from census data relating to population, age and income. Intersecting the exposure model with population susceptibility in a Geographic Information System (GIS) identified areas at high risk for health effects attributable to woodsmoke.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank the members of the Border Air Quality Study (BAQS) team that I had the opportunity to work with and learn from – it was great to be a small part of a greater project. Thank you to my supervisor Peter Keller for taking me on and coming through when I was in need of support: I always felt I was a priority amongst your rather hectic schedule and responsibilities. To my first supervisor, Michael Buzzelli, I (and my family), thank you for your commitment to family values which made my time at the University of British Columbia manageable while I was pregnant and made my transition to the University of Victoria so smooth.

I had a fantastic committee. I would like to thank Dr. Nelson and Trisalyn separately even though she is one in the same: Dr. Nelson for her professionalism and academic advice; and Trisalyn for being my friend. I am so glad you were a part of my experience – you kept me sane, kept me on track, kept me inspired, and I was always excited to see your office door open. Thank you to eagle-eye-Andrew Kmetic for your attention to detail when editing my thesis and your patience when trying to teach me the basics of Bayes. I realize now how much a supervisor and committee influence your experience as a graduate student and how fortunate I have been to have had the opportunity to work with such a great group of people.

Spatial statistical modelling with a Bayesian flair would not have been so much fun without my local BAQS team: Eleanor, you are an amazing support, role model and woman in general. I cannot overstate how much I have learned from working with you. Karla, thank you so much for all your hard work and weather data. Last but not least, thanks to Perry for being readily available for coffee - also very important.

This research was supported in part by Health Canada via an agreement with the British Columbia Centre for Disease Control to the BAQS as well as the Vancouver Island Health Authority, the Capital Regional District and the BC Ministry of Environment. In addition to financial support, staff from these organizations provided instrumentation as well as a wealth of expertise to go along with those instruments. From the Ministry of Environment: Mark Graham, Ruth-Ann Devos, Poul Christensen, Jon Sutherland and Warren McCormick were more than generous with their data, expertise, support and technical equipment.

Dedication

This thesis is dedicated to my husband Dave and my daughter Kaiya who gave me the gifts of perspective (Kaiya) and support (Dave) that I needed to see this through. I am also dedicating this thesis to my parents who are **always** there for me, unconditionally.

Chapter 1: Introduction

Outside of the Lower Fraser Valley, woodsmoke is the largest source of fine particulate matter ($PM_{2.5}$) [1] air pollution in British Columbia (BC) (Lepage and Boulton 2000). Burning wood for residential heating is increasing due to rising energy prices for electricity, gas and oil, the availability of wood as a renewable resource, and the promotion of wood as a greenhouse gas neutral energy source (Larson and Koenig 1994; Zelikoff et al. 2002; Boman et al. 2006; Naeher et al. 2007). BC's climate, topography and settlement patterns exacerbate the potential air quality and health impacts of woodsmoke in many communities where emissions from wood stoves are trapped in valley bottoms during the winter heating season due to atmospheric inversions (Ministry of Environment 2007). Despite the importance of woodsmoke as a contributor to poor air quality in BC and in many parts of the world, there is little understanding of the degree of risk, the effects of long-term exposure and the biological mechanisms linking woodsmoke to adverse health outcomes (Zelikoff et al. 2002; Naeher et al. 2007). Given the increase in wood burning for residential heating, there is a need to conduct research into the health effects of woodsmoke in order to support strategies to reduce air pollution resulting from residential heating.

Exposure assessment is an epidemiological tool used to advance the understanding of the relationship between woodsmoke and health. Exposure refers to contact between a human and an element in the environment and is a function of concentration and time (Cupitt et al. 1994; Nuckols et al. 2004). Studying the interaction between humans and the environment is inherently spatial (Nuckols et al. 2004); however, studies examining the relationship between air pollution and health typically characterize exposure for a population using measurements from a few sparsely located air quality monitoring stations, and often only one. As air pollution varies at the local scale due to the location of emission sources, local topography and weather conditions, the inability to incorporate spatial variability within urban areas is cited as a major deficiency in the field of exposure assessment (Briggs et al. 2000; Hoek et al. 2002;

[1] PM less than 2.5 μm in diameter. Anthropogenic sources of $PM_{2.5}$ include motor vehicles, power plants, industry, and home heating through the use of fireplaces and wood stoves.

Brauer et al. 2003). In addition, health research in general is coming under increasing scrutiny for conducting investigations with little regard for spatial process (Elliott et al. 2000; Ricketts 2003; Cockings et al. 2004).

There are several examples of epidemiological studies being re-investigated or modified to investigate the effect of spatial processes (i.e., Cakmak et al. 2003; Jerrett et al. 2003b; Reif et al. 2003; Buckeridge et al. 2005; Jerrett et al. 2005b). For example, Jerrett et al. (2005b) analyzed health effects due to air pollution by characterizing exposure at the local census scale and at the city-wide scale in Los Angeles. Authors found negative health outcomes to be three times greater using exposure estimates at the local scale when compared to the city-wide scale. These and other similar findings (i.e., Miller et al. 2007), suggest that air pollution and health research conducted with little regard for spatial variation in air pollution levels may be drawing misleading conclusions.

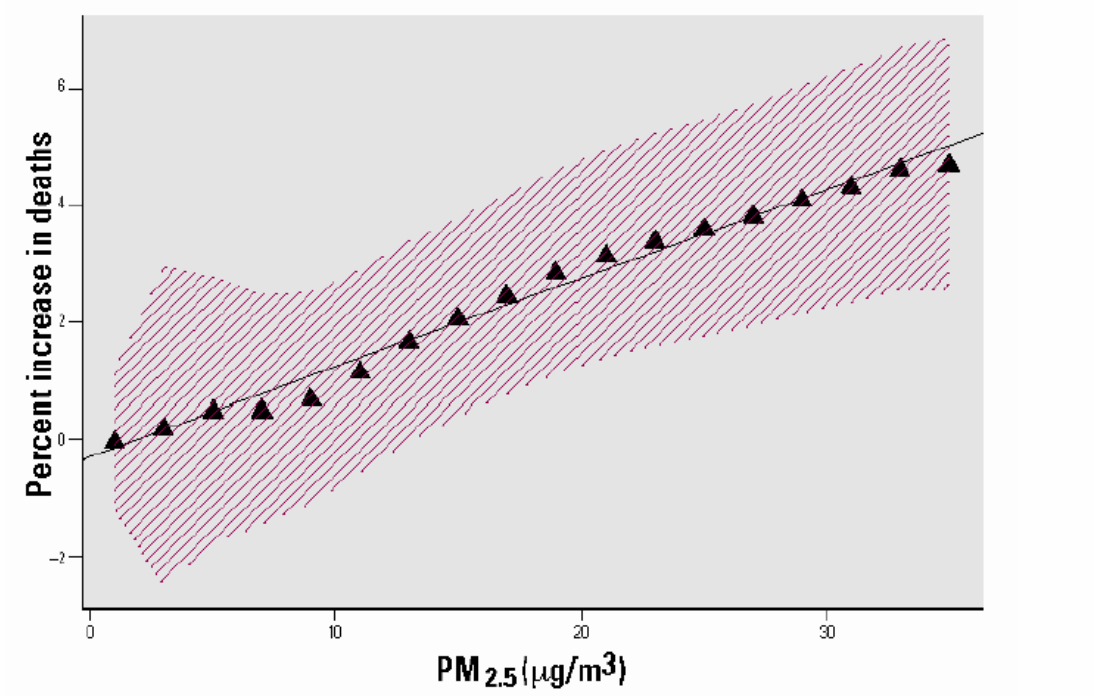
Since research results are impacted by spatial scale, researchers need to examine patterns at an informed spatial scale (Jelinski and Wu 1996) so that causal explanations, variables and generalizations match the scale of patterns observed (Clark 1985). In addition to obscuring the relationship between air pollution and health, aspatial analysis also has implications for air quality management as pollution hot spots go undetected by a sparse monitoring network. Due to the importance of incorporating spatial processes into health research and air quality management, there is increasing interest in producing spatially explicit estimates of air pollution exposure.

1.1 Problem Statement

In the Capital Regional District (CRD)[2], burning wood for residential heating has been identified by local environment and health authorities as a potentially important source of human exposure to $PM_{2.5}$. Several studies document an association between $PM_{2.5}$ and negative health effects (Lippmann and Schlesinger 2000; Brauer 2002; Brunekreef and Holgate 2002; Schwartz et al. 2002; Brauer et al. 2003). $PM_{2.5}$ particles are small enough to penetrate the gas exchange region of the lungs and has been associated with increases in mortality (Figure 1), hospital admissions, and respiratory and cardiovascular disease (Brunekreef and Holgate 2002). Figure 1 shows no safe threshold

[2] For study area details, see Chapter 3.

for $PM_{2.5}$: any increase in $PM_{2.5}$ is associated with an increase in mortality (Schwartz et al. 2002). As mentioned previously, recent research characterizing air pollution exposure at a more local level indicates that the $PM_{2.5}$ and mortality relationship in Figure 1 is potentially underestimated (Jerrett et al. 2005b; Miller et al. 2007).



(Data from the Harvard Six Cities Study showing daily deaths plotted against $PM_{2.5}$ values)

Figure 1. Percent increase in mortality as a function of exposure to $PM_{2.5}$ (from Schwartz et al. 2002)

During the winter heating season in the CRD (October through March), $PM_{2.5}$ associated with wood burning exceeds levels during commute periods (Figure 2), with this pattern holding regardless of the day of week. Although the 24 hour average for $PM_{2.5}$ (5.5 µg/m^3) is below the national health reference level[3] of 15µg/m^3 over a 24 hour period, Figure 3 demonstrates that during the evening, levels exceeding 15µg/m^3 are not uncommon. In addition, wood burning stoves and fireplaces can create indoor pollution as high as 820 µg/m^3 for a 24 hour period with approximately 70% of woodsmoke from

[3] Although Canada has a national health reference level, there is no safe threshold for $PM_{2.5}$ where no negative health effects are observed (Brunekreef and Holgate 2002; Schwartz et al. 2002).

chimneys re-entering the home and neighbouring residences (Zelikoff et al. 2002). Since individuals spend approximately 60-70% of their time outside of work at home (Zelikoff et al. 2002) during periods when air pollution attributable to woodsmoke is at its peak, and there is no safe threshold for $PM_{2.5}$, the frequent occurrence of extremes demonstrated in Figure 3 gives sufficient cause for health concern.

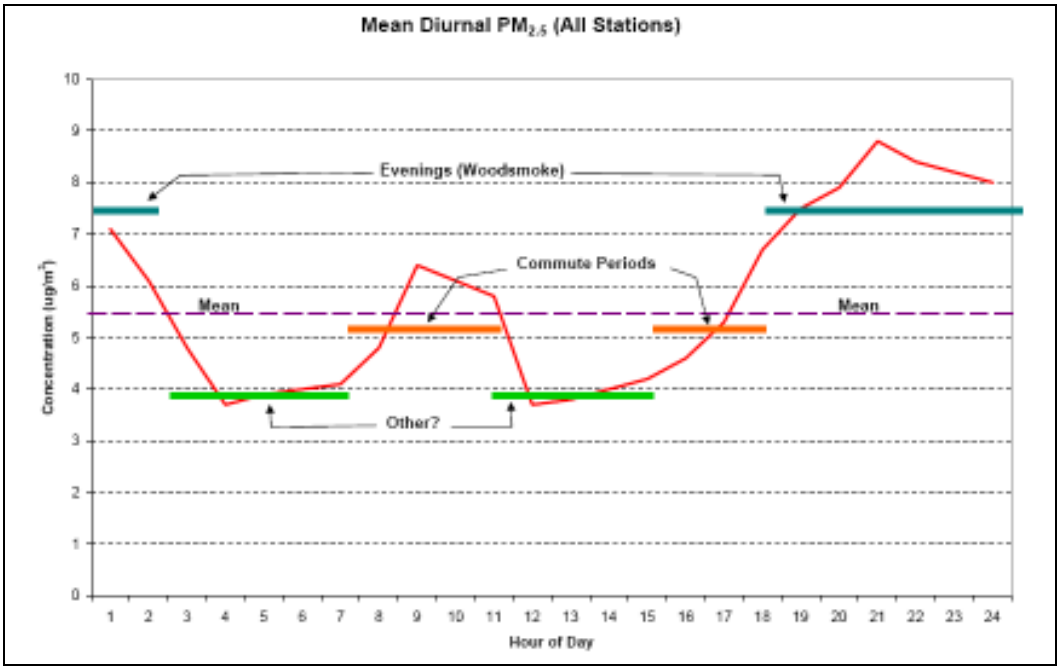


Figure 2. Diurnal $PM_{2.5}$ concentration during the winter heating season in the Capital Regional District (data are from the BC Ministry of Environment fixed monitoring network and represent the hourly averages for 2003/5, 2004/5, and 2005/6)

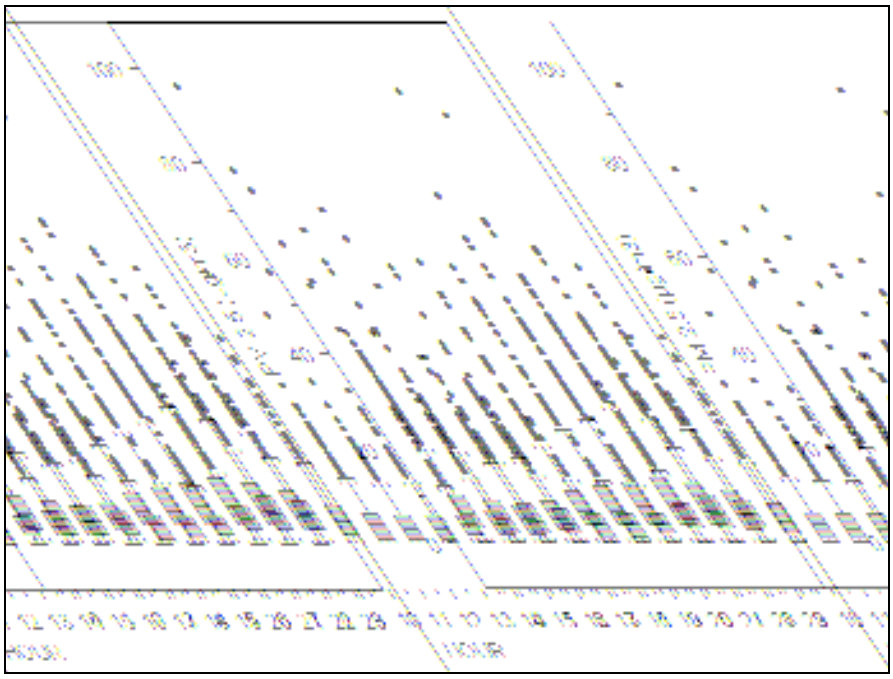


Figure 3. Boxplot of hourly $PM_{2.5}$ concentrations during the winter heating season in the Capital Regional District (data are from the BC Ministry of Environment fixed monitoring network and represent the hourly averages for 2003/5, 2004/5, and 2005/6)

1.2 Research Goal and Questions

This thesis aims to contribute to the fields of health geomatics[4], air quality management, woodsmoke and health research, as well as to the field of exposure assessment by creating a model that characterizes the spatial distribution of woodsmoke concentrations associated with residential wood burning during the winter heating season throughout the CRD. Spatially explicit measurements of $PM_{2.5}$ were collected using a novel mobile monitoring method to support modelling. Several spatial approaches to modelling $PM_{2.5}$ attributable to woodsmoke are developed and evaluated, including a land use regression (LUR) model developed by Larson et al. (2007). This thesis compares these models and addresses the following research questions:

1. Is the current fixed monitoring network representative of the spatial distribution of woodsmoke throughout the CRD?

[4] Geomatics refers to the science and technology involved in the spatial analysis of geo-referenced data (Boulos et al. 2001).

2. How does the observed spatial distribution of woodsmoke differ from that predicted by the Larson et al. (2007) woodsmoke model?
3. Does exposure to woodsmoke vary with weather conditions?
4. What is the spatial distribution of health risk attributable to woodsmoke throughout the CRD?

1.3 Thesis structure

This thesis is organized as follows: Chapter 2 is a review of the literature relating to woodsmoke and health, as well as current practices and limitations in spatial approaches to exposure assessment and risk characterization. The chapter concludes with an assessment of the literature and places this research in the context of identified gaps in woodsmoke and health research, and spatial approaches to exposure assessment and risk characterization.

The remainder of the thesis is organized around the spatial analytical framework proposed by Anselin (2006) that includes three steps: Exploratory spatial data analysis (ESDA), visualization, and spatial modelling. ESDA is the search for patterns in data (Chapter 3). Visualization depicts these patterns using spatial interpolation and spatial modelling is an attempt to explain and predict these patterns (Chapter 4).

Chapter 3 begins with an overview of the study area and data. Figure 4 is an overview of the data collection, data development and modelling process. The first section of the schematic refers to the data collection procedures and subsequent data development for modelling. To support spatial modelling, a mobile monitoring campaign collected spatially representative measurements of $PM_{2.5}$ data (Section 3.2) with an examination of the spatial structure and scale of the data. $PM_{2.5}$ data were analyzed for levoglucosan, a tracer unique to wood burning, to confirm woodsmoke is contributing to $PM_{2.5}$ pollution (Section 3.3). Spatially referenced data were collected from a variety of sources, at a variety of spatial resolutions, as independent variables and include data on topography, meteorology, social status, economic status, demographics, housing characteristics and $PM_{2.5}$ data from local fixed monitoring sites (Section 3.4). Independent variables were selected based on a theoretical understanding of air pollution dynamics and the determinants of wood burning for residential heating. Associations

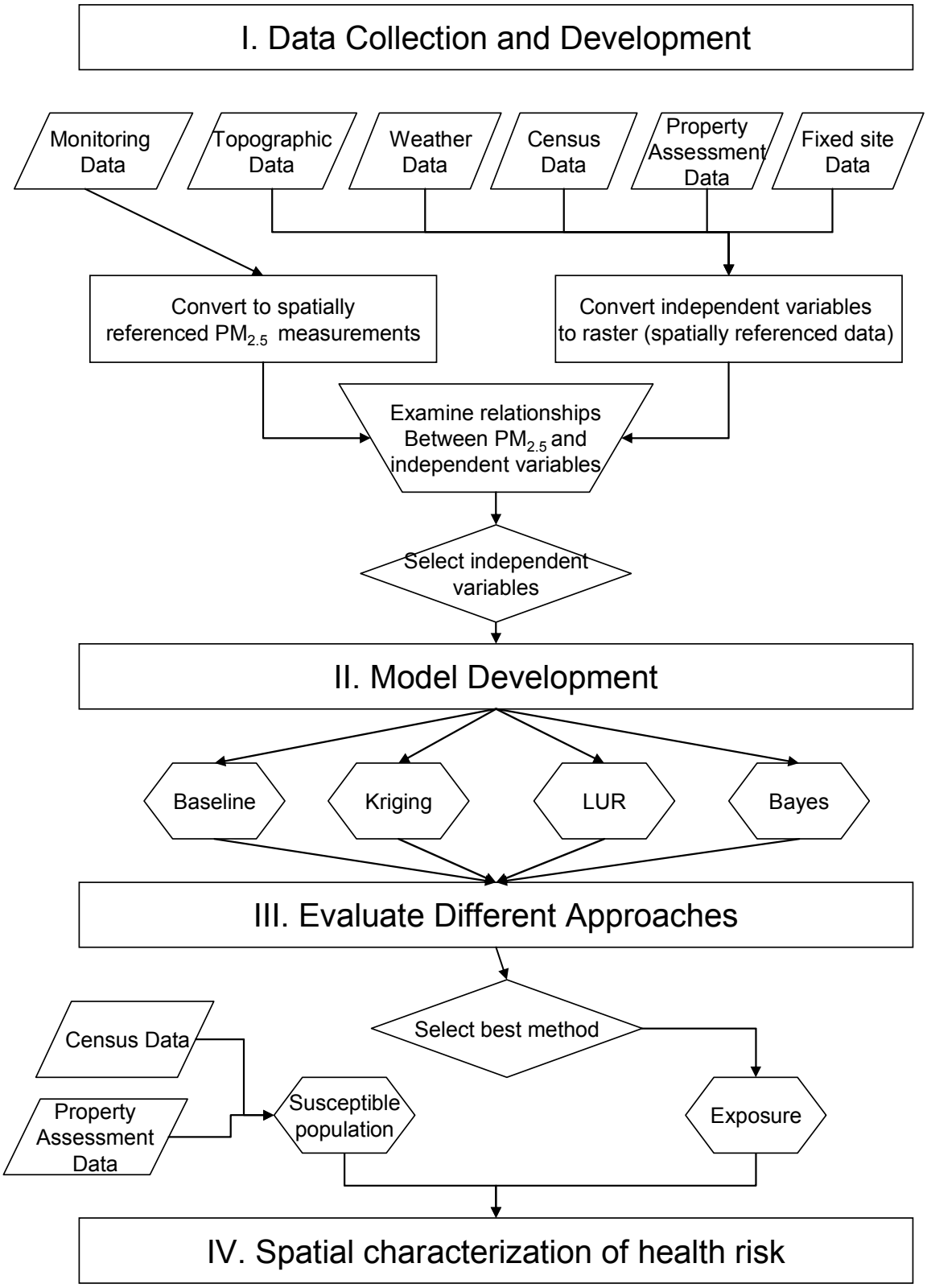


Figure 4. Overview of data collection and model development strategy

between independent and dependent variables were analysed to select variables for the subsequent modelling process (Section 3.5).

Chapter 4 describes the second and third sections of Figure 4 where different approaches for modelling exposure are examined. The different approaches include a baseline scenario, kriging, the Larson et al. (2007) woodsmoke model and a new approaches LUR modellling. The baseline scenario refers to a typical characterization of exposure using measurements from the nearest fixed site monitor. Kriging is a geostatistical technique that interpolates values at unmeasured locations based on monitoring data and the underlying spatial structure of that data. LUR uses ordinary least squares regression (OLS) to make predictions at unmeasured locations based on predictive variables such as land use. Once a regression model is developed, the model coefficients are applied to independent variable Geographic Information Systems (GIS) layers (also referred to as raster or grid data in this thesis). Arithmetic grid operations in ESRI's ArcMap's 'Spatial Analyst' extension builds pollution surfaces based on the regression models. The Larson et al. (2007) model, an ecological approach to modelling exposure, is evaluated using measured data. Then a multi-level approach is developed to retain the fine spatial resolution of the dependent and independent variable data. Finally, a Bayesian approach applies Bayes's theorem to LUR method. This chapter concludes with a discussion and comparison of the results from the different approaches and recommendations for a woodsmoke model.

Chapter 5 applies the new approach for woodsmoke modelling to a practical example by investigating the spatial distribution of health risk associated with residential wood burning. The final chapter, Chapter 6, provides conclusions and discusses the implications of this research in the broader context of spatial analysis, health research and policy; and concludes with recommendations for future research.

Chapter 2: Literature Review

This literature review summarizes the research relating to the health impacts of woodsmoke with a discussion of limitations in this field (Section 2.1). The health risk assessment process is then introduced with a focus on spatial approaches to exposure assessment and risk characterization (Section 2.2). Methods for characterizing health risk associated with air pollution are covered (Section 2.3) including the importance of scale in modelling air pollution (Section 2.4). The chapter summary positions this thesis in the context of advancing the understanding of the woodsmoke and health while contributing to the field of exposure assessment through spatial analysis (Section 2.5).

2.1 Woodsmoke and Health

The review of woodsmoke and health literature begins with a discussion of the toxicology and biology of woodsmoke followed by a review of the epidemiological evidence for negative health effects and concludes by identifying research gaps.

Woodsmoke is made up of many substances; however, it is the release of $PM_{2.5}$, volatile organic compounds, and inorganic gases that are of health concern. Although several constituents of woodsmoke such as carbon monoxide, nitrogen dioxide, benzene and PM have well documented adverse health effects, the toxicity of woodsmoke requires separate evaluation from its constituents, as has been done with tobacco smoke, because the health effects are not necessarily additive (Naeher et al. 2007) and exposure to woodsmoke is not as well studied as its constituents (Zelikoff et al. 2002). Of the woodsmoke toxic constituents, PM shows the most significant relationship with health effects (Boman et al. 2006). While some woodsmoke and health research use PM less than 10 μm (PM_{10}) as the exposure metric, $PM_{2.5}$ is preferred because larger particles tend to be removed through gravitational processes or are filtered out by the nose. Smaller particles (i.e. $PM_{2.5}$) can penetrate the gas exchange region of the lungs, having a greater impact on health (Lippmann and Schlesinger 2000; Boudet et al. 2001; Brauer 2002), and PM associated with woodsmoke is small in size, with most particles being smaller than 1 μm (Larson and Koenig 1994). In both toxicological and epidemiological studies, health effects of $PM_{2.5}$ are well documented (Lippmann and Schlesinger 2000;

Brauer 2002; Brunekreef and Holgate 2002; Schwartz et al. 2002; Zelikoff et al. 2002; Brauer et al. 2003), but most of these studies relate to $PM_{2.5}$ from fossil fuel combustion and may not reflect the same toxicity per unit mass as $PM_{2.5}$ attributable to woodsmoke (Naeher et al. 2007).

The mechanisms linking woodsmoke to biological responses are not well documented. Zelikoff et al. (2002) reviewed animal studies related to woodsmoke exposure to propose possible mechanisms for toxicity. The biological responses observed in laboratory tests on animals exposed to woodsmoke include suppressed immune system, increased incidence of cancer, decreased ventilation frequency, increased macrovascular permeability, pulmonary edema, and necrotizing tracheobronchial epithelial cell injury. Autopsies revealed the immune system as a target of woodsmoke toxicity. Zelikoff et al. (2002) hypothesize that carbon particles carry toxins into the lungs affecting macrophages, a primary defence of the respiratory system. The compromised respiratory immune system then produces secondary effects by increasing vulnerability to infection. Although extrapolating results of laboratory tests on animals to humans is considered controversial, it provides the biologic plausibility that similar mechanisms are at work in humans.

There is one controlled study of human exposure to woodsmoke. The number of subjects was small (n=13); however, exposure to woodsmoke, at 200-300$\mu g/m^3$, showed systematic inflammatory effects in the respiratory system (Naeher et al. 2007).

In a review of epidemiological studies in areas prone to high woodsmoke concentrations, Larson and Koenig (1994) found associations between PM and non-cancerous adverse respiratory effects, especially in children. Most of the studies related to children living in homes with wood stoves; however, four studies examining outdoor ambient levels showed associations with negative health outcomes such as emergency room visits for asthma. Limitations of these outdoor air pollution studies include neglecting to confirm woodsmoke as the source of PM and characterizing exposure based on one fixed site air quality monitor which could subject the study results to exposure misclassification (exposure misclassification is discussed in Section 2.2).

Boman et al. (2003) summarize the results from 9 studies where woodsmoke is identified as a major source of ambient air pollution (there is some overlap with the

studies summarized by Larson and Koenig above). All studies showed significant associations with adverse health outcomes, especially for children. Adverse health outcomes included asthma, respiratory symptoms, increased daily mortality and reduced lung function. The major limitation cited by Boman et al. (2003) was neglecting to confirm woodsmoke as the source of air pollution. In addition, 5 studies used measurements from only one monitoring site to characterize exposure where 4 studies used more than 2 monitors. All of the studies basing exposure on measurements from one monitor used PM_{10} as the exposure metric subjecting results to measurement error and exposure misclassification (Suk 1997; Nuckols et al. 2004). The 4 studies measuring PM_1 and $PM_{2.5}$ at more than one station showed increased relative risk for asthma and a reduction in lung function.

The most recent review of woodsmoke and health literature comes from Naeher et al. (2007). The goal of the review is to determine if woodsmoke merits management separately from its constituents and if there is a difference in health risk in comparison to particles of a similar size from other sources such as vehicles. The authors review 20 studies, 10 of which are not covered in Boman et al. (2003) or Larson and Koenig (1994). These studies examine woodstove and fireplace use in the home as well as health associations with outdoor concentrations attributable to woodsmoke. These studies showed similar findings to the reviews above with humans exhibiting respiratory symptoms (i.e., coughing, wheezing, chest tightness), respiratory infection, headaches, asthma, otitis media (middle ear infection), and sore throat.

Again, similar limitations exist in these studies: there are no spatially resolved or personal measurements of woodsmoke, and woodsmoke is unconfirmed as the source of PM. Limitations aside, all of the studies cover a range of exposure levels in areas known to be impacted by woodsmoke.

Naeher et al. (2007) cite some contradictory findings including a study observing no relationship between otitis media and woodstove/fireplace use in a large study of 904 infants; although, woodstoves use was associated with coughing in this study. There are also two studies showing little association between woodstove use and respiratory health and thus far, research does not show any association with heart disease.

Naeher et al. (2007) conclude that although evidence is limited, it is sufficient to argue a causal relationship exists between respiratory health and woodsmoke. This declaration is corroborated by studies of biomass burning in developing countries; although, the authors did not have enough evidence to evaluate the degree of risk, cardiovascular or cancer effects.

None of the reviews discuss a study in Boise, Idaho that identifies contributors to atmospheric carcinogens as either coming from woodsmoke or mobile sources (i.e., cars).[5] Cupitt et al. (1994) employed a more sophisticated exposure modelling approach than any of the above studies by using time activity diaries to estimate time spent in microenvironments, penetration factors, and inhalation rates. Limitations include the small number of subjects (n=43) making it unrepresentative of the population, and researchers estimated outdoor concentrations from 2 fixed air quality monitors.

The study analyzed two air filter samples from ambient monitoring stations, one from a woodsmoke impacted area and one from an area dominated by mobile sources, to estimate tumour potency and human cancer risk from extractable organic material adsorbed to ambient aerosols. Cancer risk was extrapolated based on dose-response studies in rodents, a method that has shown a high correlation with human cancer risk. Results revealed that although residential wood combustion constituted the largest portion (78%) of exposure to extractable organic material, it only accounted for 20% of the cancer risk. Mobile sources made up 11% of ambient samples but accounted for 80% of the risk. Figure 5 shows where the cancer risk attributed to the woodsmoke (WSC) and the woodsmoke/mobile sources (WSMSC) falls in relation to cancer risk from cigarette smoke and emissions from coke ovens.

[5] One hypothesis as to why it was not included is because it was based on a risk assessment model as opposed to the study of biological responses in humans or animals.

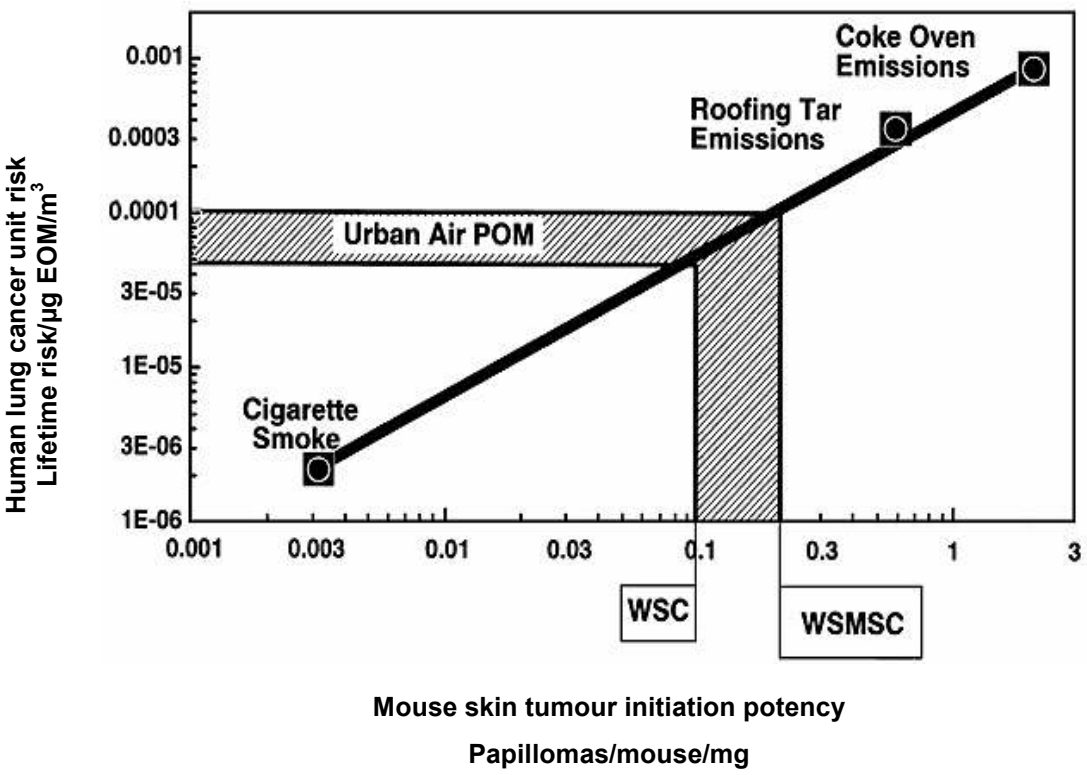


Figure 5. Human lung cancer risk extrapolated from tumour potency in rodents exposed to cigarette smoke, woodsmoke (WSC), woodsmoke and mobile sources (WSMSC), roofing tar, and coke oven emissions (from Cupitt et al. 1994)

Figure 5 shows cancer risk from woodsmoke being 30 times greater than cigarettes (Naeher et al. 2007).[6] The Cupitt et al. (1994) study supports Naeher et al. (2007) suggestion that PM from woodsmoke may not have the equivalent toxicity per unit mass as PM from other sources, and therefore requires separate management strategies. In contrast to the findings from Cupitt et al. (1994) and Naeher et al. (2007), Boman et al. (2006) state that studies from woodsmoke impacted areas showed stronger health effects than other sources of PM. This contradiction highlights the importance of confirming the source of PM as well as the requirement to evaluate the risks of woodsmoke separately from PM or its other toxic constituents.

[6] While cancer risk is much greater woodsmoke the intake fraction of woodsmoke is lower because it is diluted by ambient air. The intake fraction refers to the amount of a pollutant that is breathed in as a fraction of the amount that is emitted (Marshall et al. 2006). Cigarette smoke, on the other hand, is inhaled directly into the lungs resulting in a greater intake, and potentially greater health risk.

Although contradictory findings exist, there is coherence between the majority of epidemiological and human studies reviewed, combined with animal toxicological research suggesting a causal relationship between woodsmoke and adverse respiratory health effects, particularly in children. To confirm causality, areas to address in woodsmoke and health research include understanding the biological mechanisms causing health effects, the effects of long term exposure, and assessing the carcinogenic and cardiovascular effects to address the issue of plausibility raised by the toxicological studies.

The bulk of evidence for negative health effects associated with woodsmoke come from epidemiological studies that suffer from similar limitations. Aside from the Cupitt et al. (1994) study, no studies apportion the amount of PM that is attributable to woodsmoke and the studies cited above are based on a few fixed monitoring sites – and often only one – to characterize exposure for an entire population in a given area. In addition, studies using PM_{10} as the exposure metric risk are potentially obscuring the relationship with woodsmoke particles, through the inclusion of larger particles. Therefore, most results are subject to inaccuracies which can impede political action to address woodsmoke air pollution (Jerrett and Finkelstein 2005). Improving estimates of exposure is critical for addressing the gaps in the understanding of woodsmoke as a health risk. The following section of the literature review addresses current practices advancing the field of exposure and health risk assessment using spatial approaches.

2.2 Health Risk Assessment

Health risk assessment refers to characterizing potential adverse health outcomes resulting from human exposure to an element in the environment (Paustenbach 2002c). The steps followed in a risk assessment include hazard identification, dose-response assessment, exposure assessment and risk characterization (Figure 6). The results of the risk assessment process are then used to inform the subsequent risk management process.

The first step of the risk assessment process is hazard identification which involves assessing the toxicity, or potential for toxicity, of the element under investigation and potential health effects, or health endpoints (Gochfeld & Burger 1997). The dose-response assessment quantitatively defines the relationship between the chemical dose of the element taken in by an organism and the associated health endpoints

(Moore 2002). In most health risk assessments, an element of uncertainty arises from the dose-response assessment due to the extrapolation of results from high doses tested on animals to human exposure at much lower concentrations found in the environment (Paustenbach 2002c).

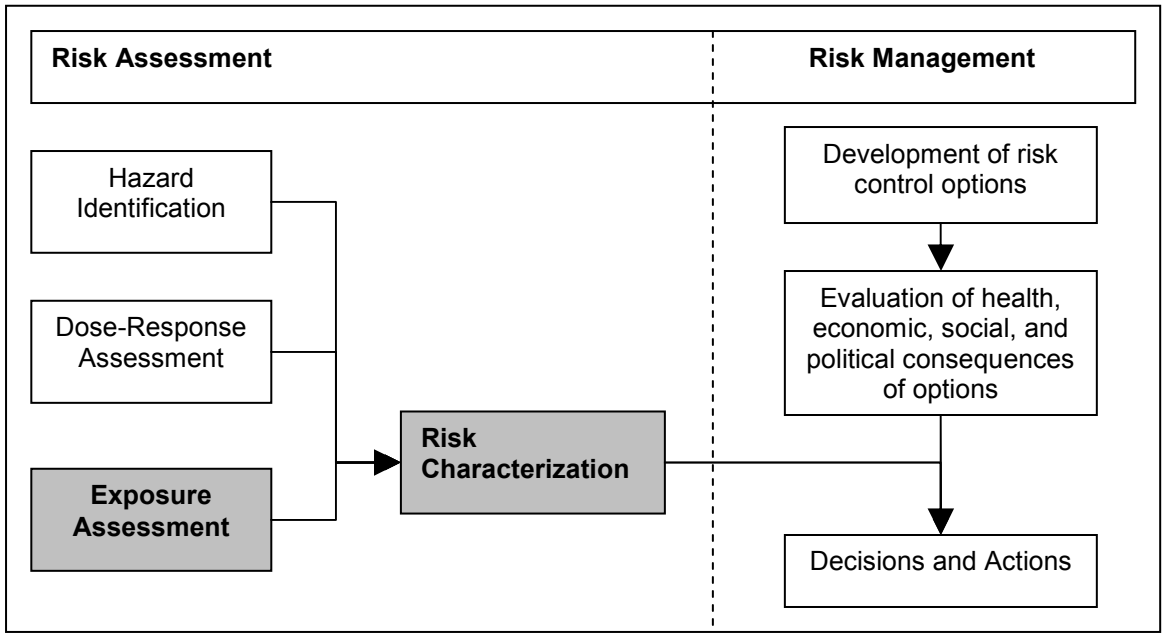


Figure 6. Risk assessment and risk management processes (adapted from Paustenbach 2002)

Hazard identification, dose-response assessment and the risk management process are beyond the scope of this thesis; therefore, the remainder of the literature review focuses on methods and practices in spatial approaches to exposure assessment and risk characterization (Figure 6).

2.2.1 Spatial Approaches to Exposure Assessment

Exposure refers to contact between a human and an element in the environment and is a function of concentration and time (Cupitt et al. 1994; Nuckols et al. 2004). Exposure assessment is a tool for estimating individual or population exposure by examining the concentration, amount, route and duration of exposure to an element in the environment (Moore 2002; Paustenbach 2002b).

Exposure can be characterized in different ways and ranges from individual measurements to the use of proxies for exposure such as distance to an emission source. Individual measurements using personal monitors provide the most accurate measure because of the ability to resolve individual variability in activity patterns, exposure to indoor sources of an element, and penetration rates of an element from outdoor air. Nevertheless, many epidemiological studies use outdoor ambient concentrations as a proxy for personal exposure because personal exposure studies can be onerous, resource intensive and, as a result, often involve small numbers of subjects (Boudet et al. 2001; Jerrett et al. 2003a).

Using outdoor ambient concentrations as a proxy for exposure can subject findings to exposure misclassification and measurement error (Suk 1997). Exposure misclassification occurs when a study population or subject is assigned to an incorrect exposure class. For example, exposure misclassification occurs when a person experiencing a high level of exposure is classified as having low or moderate exposure in an epidemiological analysis. In the CRD, the potential to underestimate risk as a result of classifying exposure based on fixed monitors exists because one of the regulatory monitors is stationed on the coast and is meant to provide a background measure for air quality management purposes because it is relatively un-impacted by anthropogenic emission sources. This reduces the $PM_{2.5}$ average concentration for the CRD, particularly for the area surrounding this monitor which is heavily impacted by residential wood burning. Because air pollution can vary within hundreds of metres, these data do not incorporate spatial and temporal variability.

Measurement error arises when the measurement used to characterize exposure is a poor surrogate for actual exposure (Nuckols et al. 2004; Jerrett and Finkelstein 2005). For example, measurement error occurs if a study uses measurements to characterize exposure that contain systematic numerical errors. All measurements are subject to error; however, it is the degree to which it occurs that jeopardizes results.

Although the effects of exposure misclassification cannot be predicted, two studies examining the effect of aggregating air pollution data to a regional level and looking for associated health effects found that exposure estimates at a more local scale showed stronger associations with negative health effects than examining the association

at a broader regional level (Jerrett et al. 2006; Miller et al. 2007). This implies that studies examining the relationship between air pollution and health subject to exposure misclassification are potentially underestimating health risk.

The effects of measurement errors are more predictable than exposure misclassification. Measurement error is thought to bias regression coefficients towards zero and reduce statistical significance (Jerrett and Finkelstein 2005) meaning research results are less likely to show an association between exposure and health effects.

Since individual exposure is rarely monitored due to time and resource constraints, modelled simulations are increasingly employed to supplement available data. A combination of modelling and measurement provide a practical solution to estimating exposure in health research.

Different options exist for modelling exposure, and those that incorporate spatial variability are considered more robust (Miranda & Dolinoy 2005; Nuckols et al. 2004). Jerrett et al. (2005a) reviewed the most common approaches to spatial modelling of exposure to air pollution within cities. The authors examined over fifty published models divided into six categories: proximity models, geostatistical models, LUR models, dispersion models, integrated meteorological-emission models and hybrid models which combine personal or household monitoring with a previously mentioned method. Table 1 summarizes some of the advantages and disadvantages associated with each method. In general, moving down the list from proximity to hybrid models represents an increase in precision and accuracy, as well as an increase in the requirements for resources such as specialized software, expertise and funding.

Proximity models are based on the distance to an emission source and are relatively easy to implement. Nevertheless, the potential for exposure misclassification is high and therefore, they are considered too simplistic. Geostatistical techniques, such as kriging, use monitoring data to estimate levels at unmeasured locations. The advantage of kriging is that it provides an error estimate for unmeasured locations; however, the quality of results is dependent on the density and distribution of monitoring stations. In addition, results cannot be extrapolated beyond the distance of spatial dependence in the measurements used to build the krigged surface (See section 3.2.2 for a discussion of spatial dependence).

Table 1. A comparison of approaches to air pollution exposure modeling (adapted from Jerrett et al. 2005).

Model	Example	Theory-concept match	Potential limitations to health studies	Data requirements	Software/expertise	Transferability	Overall implementation costs
Proximity based	Distance to road	Low	Crude exposure estimate	Traffic volumes Distance	GIS* Statistics	Low	Equipment: Low Software: Low Personnel: Med
Geostatistical	Kriging	Medium	Depends on monitoring network density	Monitor data	GIS Spatial statistics	Low	Equipment: Med Software: Med Personnel: Low
Land use regression	Larson et al. (2007)	Medium	Depends on density of observations	Traffic volumes Land use Monitor data Topography	GIS Statistics Monitor experts	Medium	Equipment: Med Software: Med Personnel: Med
Dispersion	CALINE	Medium	Simplistic assumptions about pollutant transport Extensive inputs	Traffic volumes Emissions Meteorology Monitor data Topography	Statistics Monitor experts Dispersion software	High	Equipment: High Software: High Personnel: Med
Integrated meteorological emission models	CALPUFF MODELS-3	Medium	Coarse resolution	Emissions Meteorology Monitor data Topography	Monitor experts Special software	Medium	Equipment: High Software: High Personnel: High
Hybrid (personal monitoring & preceding method)	Depends on combination	High	Small and biased sample Depends on combination	Questionnaire Personal monitoring data Other	Monitor experts Survey design Depends on combination	Depends on combination	Depends on combination Generally high

LUR modelling uses ordinary least squares (OLS) regression to make predictions of air pollutant concentrations based on predictive variables such as land use (Setton, Hystad et al. 2005; Briggs 1997). LUR modelling is gaining popularity as a tool for modelling exposure because it tends to perform well in comparison to more complicated and resource intensive approaches and it has the potential to model at a high spatial resolution. Like geostatistical approaches, LUR requires measured data; however, the spatial resolution of the model can be improved where predictor variable data exist at a higher resolution (for a more detailed discussion of LUR modelling see Section 2.4). Unlike geostatistical models, once a robust model is built, a LUR model can be applied to times or areas where little or no measured data exist; however, a small measurement campaign must be employed to evaluate performance in the new area. In addition, LUR models, as well as Dispersion and Integrated Meteorological Emissions models, can aid in targeting policy interventions by identifying key contributors to air pollution.

Dispersion models are based on Gaussian plume dynamics and use emissions, weather and topographical data to create spatial estimates of exposure. Although they are able to produce high spatial resolution estimates depending on the data sources, they are costly and often apply to a narrow area surrounding an emission source.

Integrated models use meteorological and chemical modules to simulate pollutant processes. They have the potential for real-time modelling with the capacity to incorporate time-activity patterns. Nevertheless, they are expensive and resource intensive; and as a result, have limited application in epidemiological analysis until they are further refined (Jerrett et al. 2005a). One application of Calpuff, and integrated model, to a woodsmoke impacted areas of BC found this type of modelling was unable to capture short term burn events characteristic of the area (Meyn 2006). A study of air pollution characterized by an integrated model and mortality is reviewed in Section 2.2.2.

Other approaches that are not used extensively in modelling air pollution exposure at the intraurban scale but display potential include neural network modelling, Bayesian approaches and vulnerability mapping. Neural network modelling is widely used in short term forecasting of hourly air pollutant concentrations (Comrie 1997; McKendry 2002; Grivas and Chaloulakou 2006). Neural network models are based on the concept of a biological neuron and include a system of interconnected nodes with the

potential for weighted links between them. The input nodes represent the input variables which feed into a 'black box' consisting of one or more hidden nodes, and results are output as a number of output nodes. Building a neural network model involves selecting the input variables, selecting the number of hidden nodes, selecting the number of output nodes, and selecting connection weights between them. The model is 'trained' using existing data, and weights are adjusted to minimize error (Comrie 1997; Marven 2006). The model is validated using existing data that were not used to train the model.

Strengths of neural network models include the lack of assumptions regarding the relationships between the input variables and underlying data distributions, which are weaknesses of more traditional statistical approaches (Comrie 1997; Marven 2006). Weaknesses include the black box effect, the complicated nature of the model, and the potential for overfitting the model (Slini et al. 2003).

Comrie (1997) compared multivariate regression models with neural network models to predict daily ozone levels in 8 American cities. For ease of comparison, both approaches used the same input variables which included the previous day's ozone levels, temperature, wind speed, ultra violet radiation and atmospheric moisture. Comrie found that neural network techniques performed marginally better than regression; however, the complicated nature of neural network model left regression modelling as a viable option for forecasting pollution levels. No examples applying the neural network approach to spatial distribution of air pollution were found.

Another approach, with limited application in spatial modelling of exposure, is the use of Bayesian inference. A Bayesian approach is based on Bayes's theorem which relates conditional probabilities to make probabilistic statements regarding unobserved information (Fotheringham et al. 2000). A Bayesian approach can be applied at any step in the risk assessment process and is well suited to deal with unknown parameters and missing observations. In situations where several studies exist related to the phenomena under study, such as clinical trials, Bayesian analysis provides a method for incorporating pre-existing data or models to inform the current investigation.

Bayesian approaches are often employed in studies examining the relationship between exposure and health effects, particularly in dose-response assessment and in risk management which are plagued by uncertainty and missing data.

An example of a Bayesian approach to exposure assessment is given by Shaddick and Wakefield (2002). The authors used a Bayesian approach for spatial-temporal modelling of exposure in an epidemiological analysis of acute health effects associated with high air pollution concentration events. Pollutant, temporal and spatial dependence were exploited to estimate missing data values that occurred over a three year period (1994-1997) at 8 fixed monitoring sites in London, United Kingdom.

Model results showed that dependence in the data could be used to estimate missing data values. For example, if data were missing for one pollutant at a monitoring site, it could be estimated from non-missing values of other pollutants measured at that site. Other strengths cited by the authors included its simplicity and the ability to provide measures of uncertainty for estimates at unmeasured sites. The potential to model health and exposure jointly using this approach also exists; however, this potentially results in health data 'informing' exposure, thereby compromising the validity of the exposure and health relationship modelled.

The spatial component of the Shaddick and Wakefield (2002) model was a weakness. The model is not transferable to more topographically complex areas due to the limiting assumptions regarding spatial stationarity (there is no trend in air pollution levels over a region, only local variations from the average). London is relatively flat with stable meteorological conditions rendering that assumption plausible. In addition, the model required continuous monitoring data for capturing acute exposure events limiting the spatial resolution of the study to the location of the fixed monitors.

The Bayesian approach has several advantages to consider for woodsmoke modelling:

- It deals well with uncertainty and missing data;
- It is suitable for studies with a small number of observations;
- It can incorporate pre-existing knowledge; and,
- Industry standard software (WinBUGS) is available for download free of charge.

The first three points are not an issue associated with the data collected for this research. High resolution spatial data are available for predictor variables and there are several thousands of observations for woodsmoke. In addition, there is little pre-existing

evidence to incorporate into the model. Bayesian analysis also has the same assumptions regarding colinearity as the regression approach.

Another technique meriting discussion is a vulnerability mapping technique developed by Mavroulidou et al. (2004) to identify areas susceptible to poor air quality as a result of traffic in two districts south of London, UK. Authors developed an interaction matrix with six variables considered influential to air pollution from mobile sources. The variables include traffic, wind speed, stability, surface roughness, topography and buildings. The influence of each variable on the other was quantified by expert opinion and corroborated by numerical model (ADMS-urban). The resulting weighted values were assigned to each variable. These weights were combined with raster datasets of each variable in a GIS to identify areas vulnerable to poor air quality.

This technique is similar to the LUR approach where individual 'weights' (or model coefficients in the regression model) are applied to each model variable; however, it is how the weights are derived that differs. The regression coefficients are derived based on OLS regression between the dependent and predictor variables whereas the weights for the vulnerability mapping technique are derived from expert opinion and numerical model. Both employ similar mapping techniques, where model coefficients or weights are applied to raster data sets of each variable and combined in a GIS to create maps of air pollution or areas vulnerable to air pollution.

The vulnerability mapping results were not evaluated with measured data; however, the vulnerability map highlighted areas of high risk that were substantiated by numerical modelling. Strengths of the model include the capacity for unlimited matrix size and the transferability of the model to other areas. This approach can also be used to identify areas with adverse conditions for pollutant dispersal presenting a useful policy and urban planning tool. That variables and their influence must be known is a drawback not seen in other approaches such as LUR. Although variables must also be known in LUR modelling, a model demonstrating low performance will indicate variables are missing and an examination of the spatial distribution of residuals can provide some insight into what those might be. In addition, the influence of each variable using the OLS technique is determined by the variable's statistical relationship to the dependent variable as opposed to expert opinion.

Given the strengths and weaknesses of the different methods for the spatial modelling of exposure, LUR modelling is selected as the principle approach for investigation in this thesis research. LUR demonstrates more accurate predictions than proximity and geostatistical models, and it is considered an appropriate alternative to the more complex and resource intensive approaches such as dispersion or integrated models (Comrie 1997; Briggs et al. 2000; Elliott et al. 2000; Cyrys et al. 2005; Jerrett et al. 2005a; Ross et al. 2006). There are several examples demonstrating the advantages of LUR modelling over other approaches (Jerrett et al. 2005a; Briggs et al. 2000; Elliott et al. 2000; Cyrys et al. 2005). Table 2 shows a comparison of three different approaches to modelling traffic-related air pollution. LUR shows a higher coefficient of determination (R^2) as well as a smaller standard error. In addition, defining a set of input variables to the LUR model can identify the most significant contributors to an air pollution problem, provided the predictor variables are accurately identified, having additional use for targeting policy.

Table 2. Coefficient of determination (R^2) and standard error (μg/m^3) shown in brackets for different spatial methods of modelling air pollution (adapted from Briggs et al. 2000).

City	Dispersion (CALINE-3)	Geostatistical (Kriging)	Land Use Regression
Huddersfield	**0.63** (5.25)	**0.44** (6.45)	**0.82** (3.69)
Prague		**0.34** (10.66)	**0.87** (4.67)

Weaknesses of this approach include the requirement for primary data collection, the assumption of a normal distribution for variables (environmental variables tend to be lognormally distributed), and the assumption of independence between variables. LUR has not performed well in acute exposure scenarios; however, it performs well for chronic exposure scenarios given the data and expertise required. Additionally, LUR is being used increasingly to predict traffic-related pollutants; however, it has not performed as well predicting $PM_{2.5}$ attributable to mobile sources (Brauer et al. 2003).

The strength of LUR models increases with more measurements, a condition satisfied by the woodsmoke sampling design (see Chapter 3). A possible refinement for the LUR approach is Geographic Weighted Regression (GWR) which produces local estimates of regression coefficients (i.e., estimates for each neighbourhood unit). GWR measures the relationships inherent in the model around each point *i*. Data from observations close to *i* are weighted more than data further away. Therefore, estimated parameters will be functions of the weighting scheme. It is assumed that observed data near to point *i* have more influence in the estimation of the values than data located further from *i.* Local estimates are mapped to show the spatial variation in the measured relationship which can be thought of as the spatial distribution of measured variance. This can be used to examine the assumption of stationarity in the global regression analysis. GWR also provides a spatially varying R^2 statistic (Fotheringham et al. 2000) which is under scrutiny for being artificially inflated (Wheeler 2007). Although GWR presents a methodological development worth investigation, the criticisms surrounding its performance, its limited transferability beyond the study area, and time constraints preclude this technique from being explored.

The next section reviews the only two examples of woodsmoke modelling, both of which employ a spatial LUR approach.

2.2.1.1 Modelling Woodsmoke

Tian et al. (2004) modelled the spatial distribution of potential residential wood burning (RWB) in Central California. The authors define the potential for RWB as the number of households with wood burning activity. The researchers found elevation to be the most influential variable, due to cold, windy climates at high elevations, followed by forest accessibility, degree of urbanization and temperature in explaining the variation in households with RWB activity. Model results were verified via telephone survey and model estimates were used to calculate $PM_{2.5}$ emissions from RWB. $PM_{2.5}$ emissions were estimated as a function of fuel-wood consumption, the number of households with RWB activity and combustion efficiency of wood burning appliances.

Although Tian et al. (2004) produced the first effort to model the spatial distribution of woodsmoke, it models the *potential* for wood burning activity. This method relies on survey methods for model validation which is costly and results are

rapidly outdated (Tian et al. 2004). Furthermore, results have not been validated with ambient monitoring data (Larson et al. 2007).

Using a mobilized nephelometer to measure light scatter off particles smaller than 2.5 μm, Larson et al. (2007) provide the first measurement-based LUR model to predict spatial variation in woodsmoke during the winter heating season in the Greater Vancouver Regional District (GVRD) and the CRD.

Larson et al. (2007) use a hydrological catchment (or watershed) approach for defining spatial units of the woodsmoke model (Figure 7). Hydrological catchments were defined by elevation as the area of land draining into a valley. The theory behind the catchment approach is that on meteorologically stable evenings, surface wind is influenced by hydrological drainage and flows downhill; therefore, a given location is impacted by uphill sources of woodsmoke. This model assumes woodsmoke exposure is negligible for unstable meteorological conditions (i.e., windy evenings) during the winter heating season.

The $PM_{2.5}$ measurements within a catchment were averaged and became the dependent variable in the regression model. An algorithm searched uphill from each catchment centroid (the yellow centroid in Figure 7) to select the uphill catchment areas draining into the catchment of interest. If the centroid of an upstream catchment was within a specified search distance[7] (the large circle in Figure 7) it is included in the calculation of the predictor variables for the catchment of interest. The catchments shaded orange in Figure 7 fit these criteria, and form what is called the catchment buffer area. Predictor variables were aggregated to the catchment buffer area level and are regressed on the dependent variable. Potential predictor variables, 28 in total, were derived from census data and SPAD and related to income, building age, demographics and emissions. Predictor variables were selected for inclusion in the model based on the correlation with the dependent variable (i.e. variables with the highest correlation were selected) and significance in the model. Predictor variables for the CRD include the average number of fireplaces, the percent of the population that is low income and the number of immigrants in the catchment buffer area. Figure 8 shows the predicted $PM_{2.5}$ values based on the CRD model.

[7] For the CRD, upstream search distance is 4km

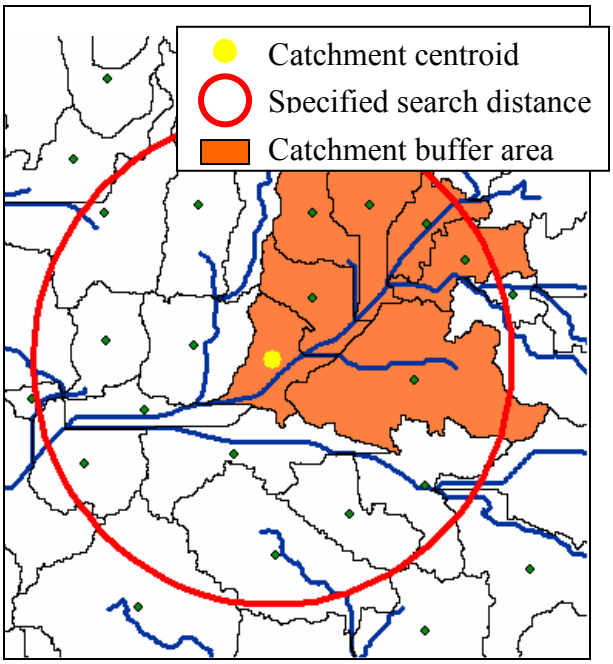


Figure 7. Hypothetical hydrological catchment basins, catchment basin centroids, search radius and catchment buffer area (from Larson et al. 2007)

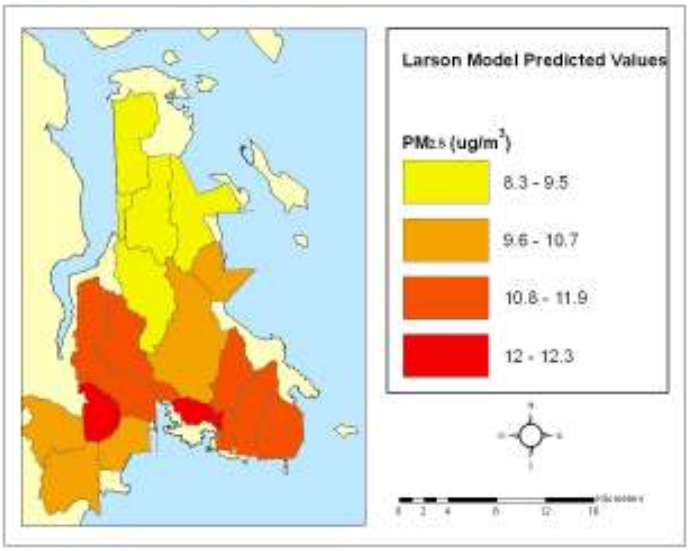


Figure 8. Larson et al. (2007) predicted woodsmoke concentrations for 9 km^2 catchment basins in the Capital Regional District

Through this approach, researchers identified areas experiencing elevated woodsmoke concentrations not measured by the regulatory fixed monitoring network. This study addressed the gap of confirming the presence of woodsmoke through the measurement of levoglucosan levels, a unique tracer of woodsmoke (this was completed for the GVRD, not the CRD). This model illustrates the ability to enhance monitoring data through the use of readily available data and the model performs well (R^2=0.73, p<0.00).[8]

Weaknesses of the Larson Model include:

- The ecological approach since data were aggregated to the catchment basin and catchment buffer area creating potential for the ecological fallacy as well as the Modifiable Areal Unit Problem (MAUP)[9];
- Coastal basins are assumed to be unexposed;
- Lack of standardizing predictor variables, some predictor variables are expressed as a percentage whereas others are raw numbers;
- Colinearity between variables violating regression modelling assumptions;
- Lack of theory underpinning the inclusion of immigrants as predictors of woodsmoke; and,
- The model predicts a seasonal average; however, it is built on data collected on cool, clear and calm evenings which may not represent seasonal average concentrations.

This concludes the review of spatial approaches to exposure assessment and woodsmoke modelling. The next section examines literature related to risk characterizations, the next step in the health risk assessment process (Figure 6).

2.2.2 Risk Characterization

Definitions of risk characterization vary and depend on the goal of the risk assessment. Risk characterizations in the literature ranged from a simple discussion of extrapolating results from an epidemiological analysis to another geographic area

[8] In Larson et al., (2007), the R^2 for the CRD model is reported as 0.84 (a typo) which is the R value, not the R^2.

[9] The ecological fallacy arises when the statistical relationship existing at an aggregated scale of analysis is assumed to hold at a more detailed scale. It is related to the MAUP where changes in the size and shape of spatial units produce varied results (O'Sullivan and Unwin 2003).

(Greenberg 1997), to the complicated modelling of population health effects by the United States Environmental Protection Agency (US EPA) reviewed below.

The risk characterization process draws upon the three previous steps in the risk assessment process to estimate the likelihood and severity of health effects due to exposure (Pierson et al. 1991; Paustenbach 2002a). While risk characterization depends on quantitative results from the previous assessments, it is qualitative in how those results are interpreted due to the subjectivity involved in defining acceptable risk (Paustenbach 2002c). For instance, there is no threshold for exposure to $PM_{2.5}$ below which no negative health effects occur; however, obtaining a negligible ambient concentration of $PM_{2.5}$ is unrealistic and undesirable. Therefore, risk assessors and managers must choose and defend an acceptable level.

Risk characterization is the stage where the sources of uncertainty inherent in the risk assessment process are communicated (Williams and Paustenbach 2002; Faustman and Omenn 1996). The aim of risk characterization is to facilitate informed decision-making by risk managers by summarizing and communicating key findings in the risk assessment process (Paustenbach 2002c). Table 3 displays a framework for the risk assessment process together with the information in columns A though G that the risk characterization draws upon. Ideally the exposure assessment identifies the pollution

Table 3. The components of a health risk assessment (adapted from Pierson et al. 1991).

*Risk Assessment									
Hazard Identification									
Exposure Assessment	**Dose Response**			**Exposure Assessment**			**Risk Characterization**		
A	B	C	D	E	F	G	H	I	J
Pollution Source	**Dosimetry Factors**	**Dose**	**Response Factor**	**Pollutant Concentration**	**Exposure Duration/Setting**	**Exposure**	**Individual Health Effect**	**Exposed Population**	**Population Health Effects**
Ambient Air	Contact rate	Cumulative and/or peak	Non-carcinogenic potency	Peak	Short Duration	Integrated from direct measure	Adverse effect	Infants	Incidence
Earth or Ground	Absorption rate		Non-carcinogenic threshold	Constant	Long Duration		Multiple organs	Children Home School	
Gas stoves	Regional surface area of lungs	Pollutant mass per body weight per time	Dose-response function	Variable	Microenvironments	Calculated from columns E and F	Multiple symptoms	Adults Male Female	
Tobacco Smoking	Body weight		Multiple organs	Averaging time	Outdoors		Severity of symptoms/effect	Worker Homemaker Smoker Nonsmoker	
Vehicle exhaust	Other factors	Pollutant mass per surface area per time	Multiple symptoms		Indoors		Impairment of function	Seniors Other	
Woodstoves			Synergistic effects		Time-Activity Patterns			Susceptibility	

*See also Figure 6.

sources, pollutant concentrations, and duration and intensity of exposures based on time spent in different environments (columns A and E-G). This is combined with available dose-response information from columns B-D to arrive at estimates of health effects (columns H-J). Cancer risks are usually characterized by the number of cases attributed to exposure. Non-cancer health effects, such as those associated with woodsmoke, are more complicated to characterize since there are multiple health effects and susceptibility varies within a population. Table 3 represents an ideal – typically these data are not available and risk assessments are more limited in scope (Pierson et al. 1991).

The US EPA risk characterization of air toxics provides an example of a classic risk characterization; however, it is not a spatial approach. The US EPA modelled pollutant concentrations using the ASPEN dispersion model and estimated exposure using an inhalation exposure model called HAPEM4. The HAPEM4 model incorporates census data, human activity patterns, ambient air quality levels, meteorological data, and indoor/outdoor concentration relationships to estimate an inhalation exposure distribution for groups of individuals. Dose-response information came from government health databases (EPA 1995; EPA 2005). Authors combined dose-response and exposure information using the risk characterization program called TRIM.Risk, a collection of risk characterization tools used to combine human health and environmental risk data using probabilistic and non-probabilistic functions specified by the user (EPA 2005).

The results were limited by data constraints as data quality varied significantly across the country. Nevertheless, this risk assessment demonstrated that millions of people in the US are exposed to significant health risks from air pollution. The risk assessment did not identify high risk geographic areas; rather it presented ranges of risk across the country. The models used to construct the assessment were not based on measured data and modelled estimates tended to be lower than the few measurement validations performed.

The US EPA risk characterization provides an example where all the steps in a risk assessment are combined to produce estimates of various health effects. It also demonstrates the resource intensity, the complexity of combining a diversity of data sets and the limitations of the risk assessment process. In spite of all of the modelling and data, it is only considered a screening level assessment (EPA 2005).

A spatial approach to a risk characterization more limited in scope than the US EPA risk characterization is provided by Scoggins et al. (2004) in their spatial analysis of exposure to nitrogen dioxide (NO_2) and mortality in Auckland, New Zealand. Controlling for several confounding variables (i.e., socio-economic status, sex, age, ethnicity, external cause mortality), Scoggins et al. (2004) modelled exposure using dispersion modelling (CALGRID) at the census unit level. They developed two models: one using logistic regression to predict excess deaths in high air pollution census units (census units with greater than 13 $\mu g/m^3$ annual average NO_2) and one using odds ratios to estimate the percent increase in mortality per 1% increase in annual average NO_2. Authors estimated that 472 excess deaths per year were due to air pollution; a rate three times greater than deaths attributable to automobile accidents.

Figure 9 shows the number of people exposed to various annual average NO_2 levels. The map demonstrates more than half the population of Auckland being exposed to levels of some concern and where those census units are located, providing a useful policy tool. Weaknesses of this approach include its limited scope in that it only includes ambient pollutant concentrations and one population health effect (columns E and J from the risk assessment framework in Table 3). There is potential exposure misclassification and the ecological fallacy because it assumes a uniform distribution of exposure and population characteristics within each census unit and that the concentration at a person's home address provides an adequate measure of a person's exposure to NO_2. The categorization of air quality (i.e., Excellent, Good) is given little explanation and yet it is a large determinant of the perception of risk. In addition, the dispersion model was based on data limited by the emissions inventory and climate data.

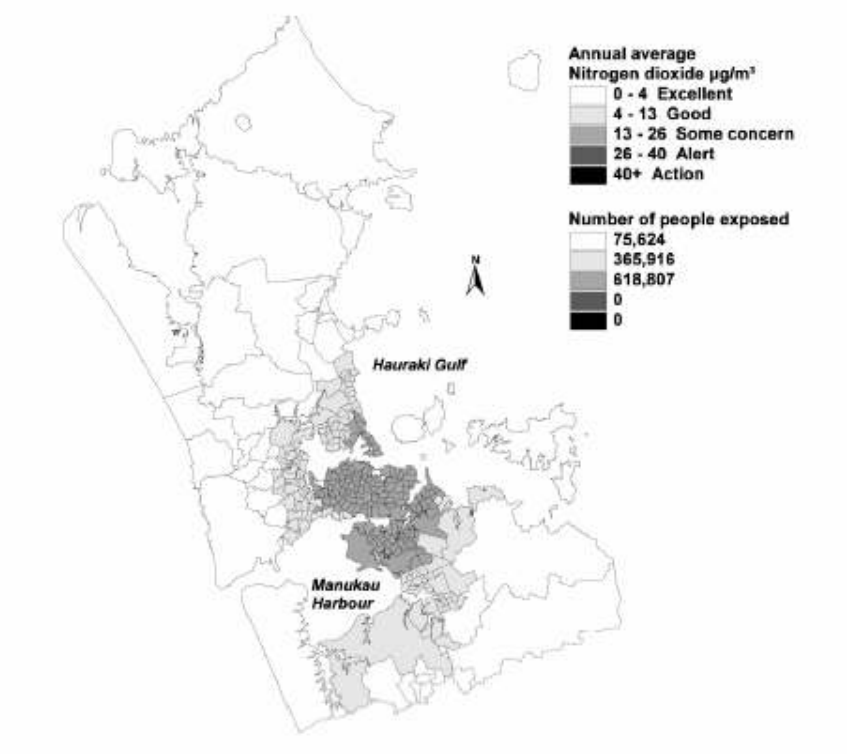


Figure 9. Annual average NO_2 and number of people exposed to different concentration levels in Auckland, New Zealand (from Scoggins et al. 2004).

2.2.3 Assessing Health Risk Associated with Woodsmoke

Any literature resembling a risk characterization for woodsmoke come from Cupitt et al. (1994), Larson and Koenig (1994), Zelikoff et al. (2002), Boman et al. (2006), and Naeher et al. (2007) that are summarized in Section 2.1. These papers are not an analysis of the spatial distribution of health risk associated with woodsmoke; however, they summarize the evidence indicating woodsmoke is a health concern. These papers resemble the hazard identification stage of the risk assessment process as opposed to a risk characterization. A full scale risk characterization is beyond the scope of this thesis; however, there are smaller scale approaches such as the technique shown in Figure 9 from Scoggins et al. (2004) or outlined in Mavroulidou et al. (2004) that provide methods for a limited characterization of the spatial distribution of health risk associated with woodsmoke as part of this thesis.

There are few spatial approaches to characterizing health risk from air pollution. Jerrett and Finkelstein (2005) call for comparisons between spatial and non-spatial approaches to assess the effects of spatial processes on modelling risk estimates. Jerrett

and Finkelstein (2005) provide a simple framework for identifying the Geography of Risk defined as the spatial distribution of health risk associated with exposure (Figure 10). This approach consists of identifying the area of overlap between the Geography of Exposure and the Geography of Susceptibility with an understanding of the limitations when working with data of different temporal and spatial scales.

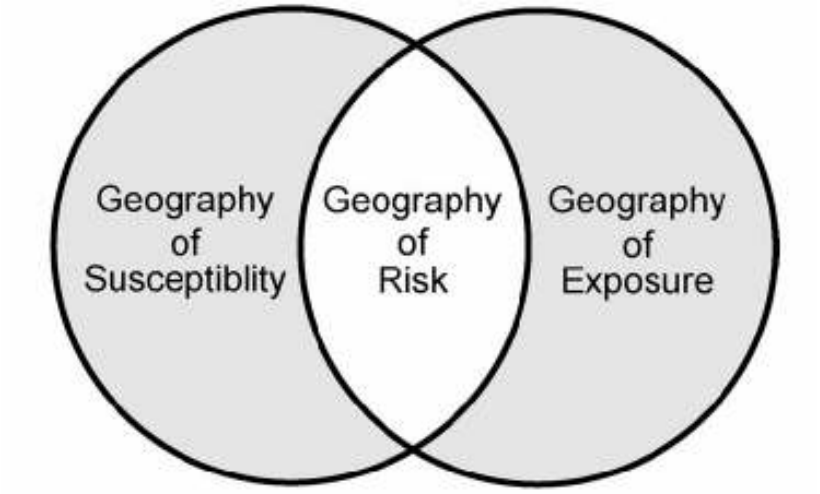


Figure 10. Analytic framework for identifying the Geography of Risk (from Jerret and Finkelstein 2005)

The Geography of Exposure is characterized by defensible metrics of exposure posing as proxies for individual exposure (Jerrett and Finkelstein 2005). The Geography of Susceptibility consists of two components: compositional and contextual susceptibility. Compositional susceptibility refers to individual variations in susceptibility such as smoking habits or age. Compositional susceptibility is difficult to model because individual level data are often difficult, time consuming, and expensive to obtain. Contextual susceptibility refers to the place or environmental factors that influence health such as the socio-economic status of a neighbourhood. Contextual factors are considered influential to health producing odds ratios as high as 2 (Jerrett and Finkelstein 2005). For comparison purposes, health effects attributable to air pollution produce odds ratios in the range of 1.1-1.2.

Although the approach appears simple, spatial mismatches often occur between data used to characterize exposure and susceptibility because they are not collected at the same scale. Where the overlap between exposure and susceptibility is incomplete or

incorrect due to incongruence is data scales, unreliable results are produced (Jerrett and Finkelstein 2005). Jerrett and Finkelstein suggest spatial analysis of residuals to determine if spatial mismatch is a problem in modelling the Geography of Risk since it produces spatial autocorrelation in model residuals.

The next section explores the importance of scale in more detail for spatial modelling of air pollution.

2.3 Spatial Scale

Scale is one of the most important considerations for creating and analyzing geographic data for exposure assessment and epidemiology (Nuckols et al. 2004) yet little is known about the scale at which air pollutants vary (Jerrett and Finkelstein 2005). Diem (2003) and Nuckols et al. (2004) outline different aspects of scale including cartographic scale, geographic scale, measurement scale and operational scale. The cartographic or map scale refers to the ratio between the size of a feature on a map to the size of feature on the ground (Diem 2003; Nuckols et al. 2004). The geographic scale is the spatial extent of the study area and is usually determined by data availability, or in health research, it is determined by the population under study (Nuckols et al. 2004). The spatial extent has implications for spatial analysis referred to as 'edge effects' which arise when an artificial boundary is imposed on a study area that does not necessarily reflect the extent of the process in question. Problems arise when observations in the centre of the region contain neighbours in all directions whereas observations at the edges of the study area do not (O'Sullivan and Unwin 2003). The spatial extent can be modified by partitioning the data to investigate stationarity, an underlying assumption in most spatial interpolation and modelling procedures.

The measurement scale, or sampling interval, determines the smallest unit distinguishable in a study (i.e., the spatial resolution). Spatial resolution is of particular interest because different patterns emerge with investigations of varying resolutions (Wiens 1989). For instance, field sparrows demonstrate aggregation at the local scale. In contrast, at a larger regional scale, the local pattern of aggregation is replaced by segregation (Davis et al. 2000). A small sampling interval facilitates subsequent interpolation and is essential for determining the operational scale (Diem 2003).

The operational scale refers to the area at which a spatial process occurs (Diem 2003; Nuckols et al. 2004). The operational scale is identified by the presence of spatial dependence: the notion that spatial data at nearby locations are more related than data located further away. Processes are scale dependent meaning they may be detected at a certain resolution or extent but not another (Nuckols et al. 2004) which necessitates a small sampling interval to accurately determine the operational scale (Diem 2003).

Another aspect of scale not addressed by Nuckols et al. (2004) or Diem (2003) is the scale of analysis which refers to the size of search radius or neighbourhood size used in spatial analysis such as hot spot detection, kriging or spatial modelling. Neighbourhoods used in spatial analysis are often defined by distance (O'Sullivan and Unwin 2003) and need to be informed by the operational scale to avoid some of the problems associated with spatial data such as the Ecological Fallacy and the MAUP.

This discussion of scale highlights the importance of understanding scale in spatial modelling. Pollutant dispersal processes will operate at specific scales; therefore it is important to understand the operational scale to model it. Since observed spatial patterns are dependent on the scale of an investigation, researchers need to examine patterns at an informed spatial scale (Jelinski and Wu 1996; Wiens, 1989) to produce reliable results. Air pollution has a strong spatial component, yet little regard is given to spatial scale and the selection of spatial units for modelling (Jerrett and Finkelstein 2005); a gap in the exposure assessment literature that this research seeks to address.

2.4 Literature Review Summary

The literature review sets the context for this thesis. The goal is to provide an epidemiological tool for advancing the understanding of the woodsmoke and health relationship and to contribute to methodological developments in spatial approaches to exposure and risk assessment through the application of geomatics.

Research investigating the relationship between woodsmoke and health is limited and results are often based on one monitoring station resulting in exposure misclassification. Spatial scale goes unacknowledged in any literature related to woodsmoke and health, meaning current research may be failing to capture underlying spatial processes that are linked to health (Jerrett and Finkelstein 2005) which potentially underestimates health risk.

Increasing the potential for exposure misclassification and measurement error in existing health research is the failure to confirm the source of PM as woodsmoke, especially where coarse particulates (i.e., PM_{10}) are the metric under investigation. Woodsmoke consists of particles smaller than 1μm; therefore, the relationship between smoke and health may be obscured by larger particles. Uncertainty surrounding research results will stall political efforts to address this form of air pollution, especially since heating homes with wood is ubiquitous and is perceived as a relatively innocuous activity. Therefore, using a smaller particulate measurement at a fine spatial resolution, as well as confirming the source of particulate matter, is imperative to advancing research in woodsmoke and health.

Although woodsmoke research suffers from methodological constraints there is sufficient evidence to merit a health concern. There is coherence between the epidemiological studies, human controlled and animal toxicological evidence, suggesting a causal relationship between woodsmoke and adverse respiratory health effects. To date, no research has addressed long term exposure to woodsmoke or the carcinogenic effects rendered plausible given carcinogenic effects seen in laboratory animals exposed to woodsmoke.

The literature related to spatial approaches to exposure assessment demonstrate that where measurement data are limited, air pollution levels can be predicted by readily available data on topography, climate, demography and socio-economics to provide exposure estimates at a fine spatial scale. The examples in this review reveal how a spatial approach can refine measures of exposure, in turn, improving our understanding of the health effects of woodsmoke and the spatial distribution of health risk associated with it.

This thesis takes current limitations and future directions into account to produce a model characterizing the spatial distribution of woodsmoke at a fine spatial resolution. The literature review examined several options for spatial modelling of exposure to air pollution and the advantages and disadvantages of each. Jerrett et al. (2005a) propose hybrid models as the ideal approach because the use of personal monitors for measurement validation of another approach increases the capacity to achieve population representativeness while capturing the role of individual variability. Since the hybrid

approach is beyond the time and resource constraints for this research, LUR is selected as the modelling approach with a comparison to the baseline scenario (exposure based on measurements from fixed monitors), kriging and a Bayesian modification of LUR.

There have been no spatial approaches to risk characterizations for woodsmoke specifically; however experience can be drawn from risk characterizations for air pollutants and from the risk assessment literature in general. Marven (2006) outlines considerations for analyzing environmental risk from oil spills which can be extended to health which includes predicting exposure levels locally for a region, integrating diverse datasets to calculate meaningful outcomes, and integrating sub-models and analyses into a larger risk model.

Defining acceptable levels of risk also requires consideration due to the subjectivity involved and its importance on the perception of the results of a risk assessment. Risk is not synonymous with exposure due to population variability in time activity patterns and susceptibility. In rare instances where population distributions of exposure and doses are available, these data sets can be used to calculate a distribution of risk (EPA 1995). Nonetheless, defining population susceptibility is a challenging task, and as a result, exposure is often used as a proxy for risk representing a significant limitation in most risk characterizations. When Gulliver and Briggs (2005) modelled exposure profiles for individuals based on mobility and time spent in different microenvironments, they found it did not lead to significantly different exposure estimates when compared with traditional approaches of modelling exposure at one's primary residence or school. Nonetheless, the effect of mobility and infiltration can only be assessed through a personal monitoring campaign to evaluate the model.

This thesis builds and expands upon the methods discussed to produce robust estimates of woodsmoke exposure needed to address existing research gaps with respect to carcinogenic and cardiovascular effects, long term effects, intake fraction and risk characterization for woodsmoke. Current limitations and directions in woodsmoke research are taken into consideration to produce a statistically robust model characterizing the spatial distribution of woodsmoke with special attention to spatial scale. The model is developed with the hopes that further refinements such as the inclusion of time-activity patterns, indoor infiltration, personal monitoring, intake

fraction and the definition of susceptible populations could be incorporated in future research.

Chapter 3: Study Area and Data

This Chapter summarizes the study area the data required to support spatial modelling of exposure and health risk associated with woodsmoke.

3.1 Study Area

The study area encompasses the CRD which is located on Vancouver Island, BC (Figure 11). The area is located on the southern tip of the island at ~ 48°N latitude and ~123°W longitude.

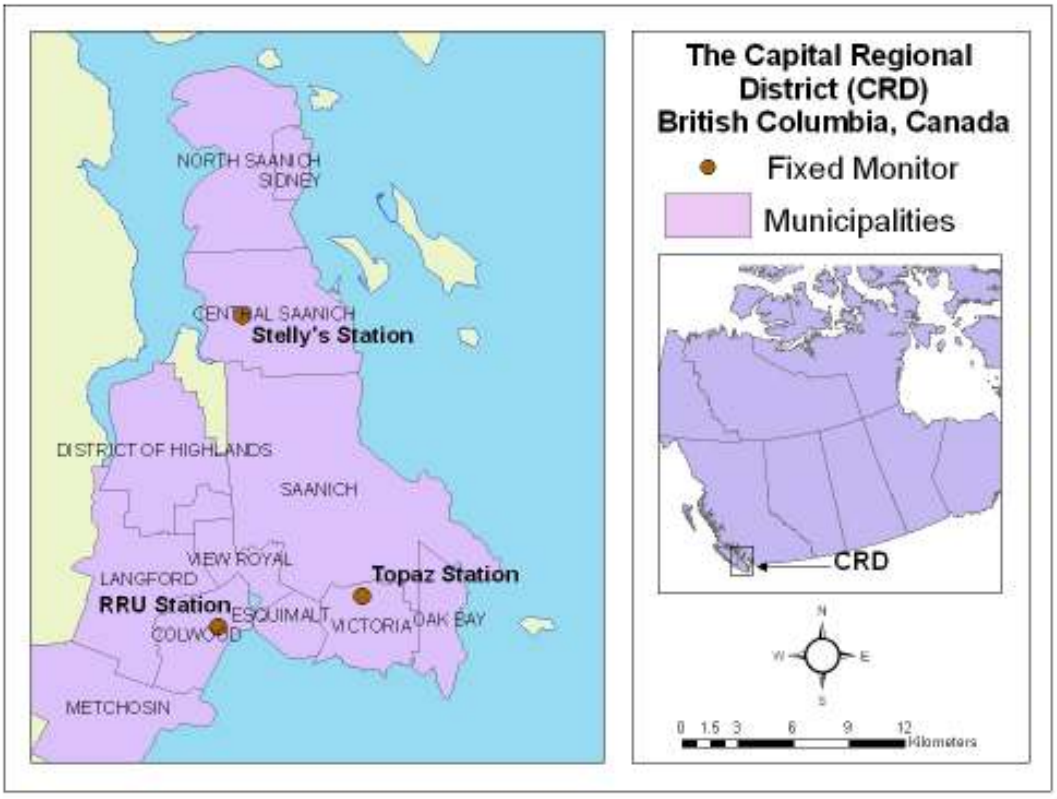


Figure 11. The Capital Regional District, British Columbia, Canada, and its municipalities

3.2 Woodsmoke Data and Summary of Field Work

A Radiance Research M903 nephelometer (Figure 12) measured light scatter of particles less than 2.5 μm, and a DeLorme Bluelogger© (WAAS enabled) Global Positioning System (GPS) recorded the geographic coordinates of each measurement. Both instruments were installed in a passenger vehicle and a funnel at the end of flexible

copper tubing extended outside the vehicle as the air intake for the nephelometer. The air intake extended out the window on the opposite side of the exhaust pipe as a precautionary measure to avoid vehicle self pollution. Outside air entering the nephelometer through the air intake passed through a heater to maintain consistent temperature and relative humidity before measurement.

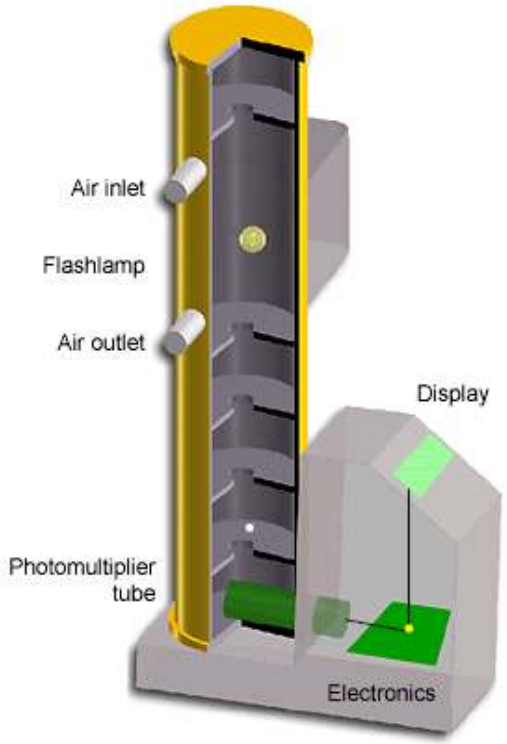


Figure 12. Radiance Research M903 nephelometer

The nephelometer logged average light scatter values every 15 seconds at 0, 15, 30 and 45 seconds for each clock minute. The GPS logged latitude, longitude and elevation every 15 seconds, beginning with the first second of satellite acquisition, resulting in logged values at, for example, every 9, 24, 39, and 54 seconds for each clock minute. GPS and nephelometer data were combined based on the closest logged time (i.e., GPS data logged at 9, 24, 39, and 54 seconds matched nephelometer data logged at 15, 30, 45, and 0 seconds). The largest error in matching time therefore is 7.5 seconds. Given travelling speeds in the range of 40 to 60 km/h (11 to 17 m/s), the logged GPS position is a maximum of 128 metres from the actual position of the nephelometer. In

addition, GPS signal quality varies with positional errors of up to 30m. Data were not corrected for positional errors, as an accuracy of +/- 160 m is considered acceptable for this study.

The nephelometer logged the average light scatter values measured in the preceding 15 seconds. With traveling speeds of 40 to 60 km/h, the logged value represents an average light scatter value measured over 165 to 255 metres.

At the end of each sample evening, nephelometer and GPS data were downloaded as comma separated values (CSV) files. The CSV files were imported into ArcMap and converted to point shapefiles with $PM_{2.5}$ as attribute data (conversion of light scatter to $PM_{2.5}$ is discussed below). The sample evenings were cleaned based on a minimum distance rule of 30m between points to avoid measurements taken at stop lights where vehicle emissions could potentially skew data.

Nephelometer and GPS data were collected during evenings, between 9 and 11pm during 3 consecutive winter heating seasons (November through March of 2004/2005, 2005/2006, 2007/2008) using a route covering the area expected to be impacted by smoke. The route was not scientifically derived using a location-allocation algorithm, rather, the route was chosen based on local and expert knowledge including the Vancouver Island Health Authority, the CRD, the University of Victoria's Spatial Sciences laboratory staff and Dr. Timothy Larson (University of Washington). Since mobile sampling is limited to roads and areas of high population density[10], sampling is not evenly distributed throughout the region. Data were collected for 32 sample evenings, over the 3 heating seasons, for a total of 14,196 individual measurements (Figure 13). 16 of the sample evenings cover the same route for model development. The other 16 sample evenings were collected to use for model validation. The sample evenings included weekends, weekdays and holidays to investigate possible differences. Figure 14 shows the winter heating season diurnal pattern of $PM_{2.5}$ by day of the week. This corroborates the finding from monitoring data that day of the week and holidays do not impact evening $PM_{2.5}$ levels.

[10] Since the ultimate goal of modeling woodsmoke is to estimate exposure, only populated areas were sampled.

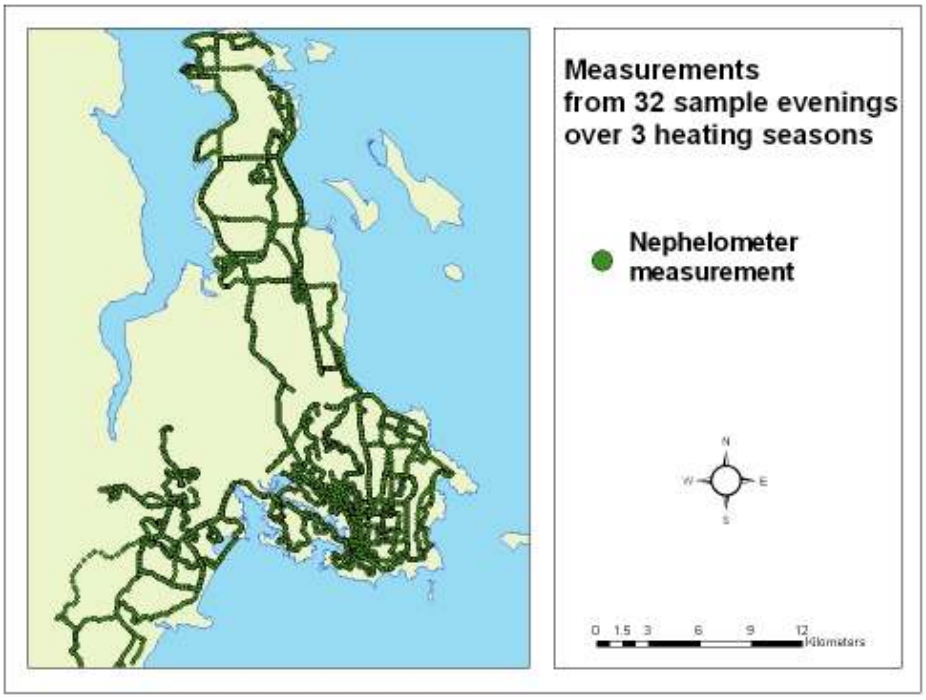


Figure 13. Nephelometer measurement routes from 3 winter heating seasons in the Capital Regional District

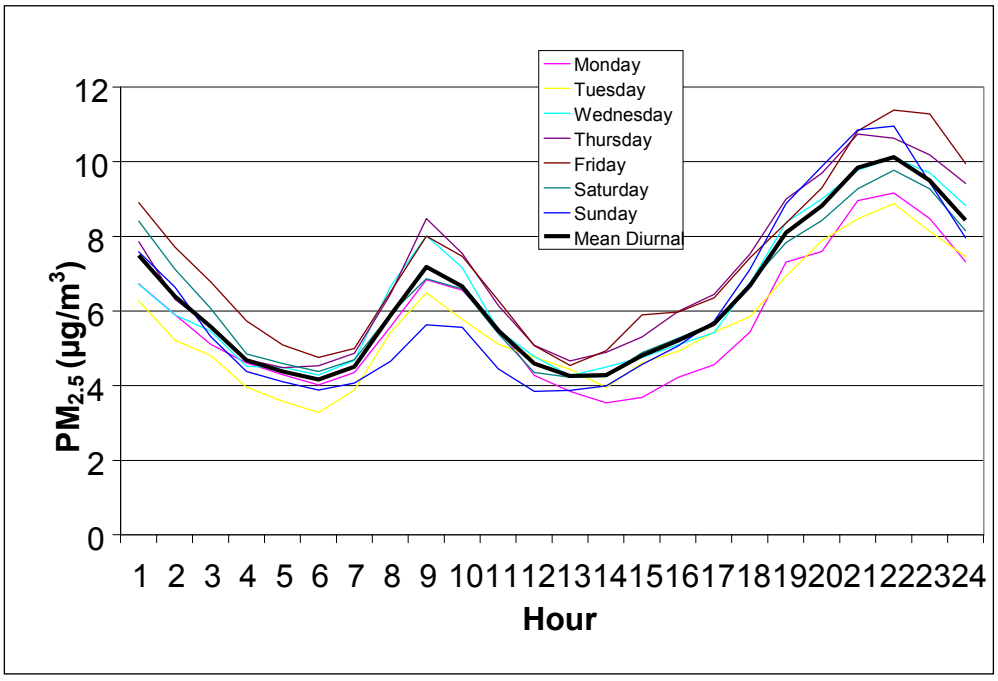


Figure 14 . Diurnal pattern of $PM_{2.5}$ in the Capital Regional District during the winter heating season by day of the week (average of 3 fixed monitors over 3 heating seasons)

The nephelometer measures light scatter (b_{sp}) off particles less than 2.5 µm which requires conversion to a $PM_{2.5}$ concentration using the formula from Allen et al. (2003):

$$PM_{2.5} = ((b_{sp} * 100000) - 0.01)/0.28 \quad (1)$$

The validity of this conversion for Victoria was investigated using data from a nephelometer stationed alongside the TEOM monitor[11] located at Topaz station for one week. Figure 15 shows the results of the nephelometer and fixed monitor measurements for the week. The nephelometer measurements were regressed against the fixed site measurements to evaluate the validity of equation 1, the results of which are shown in Tables 4 and 5.

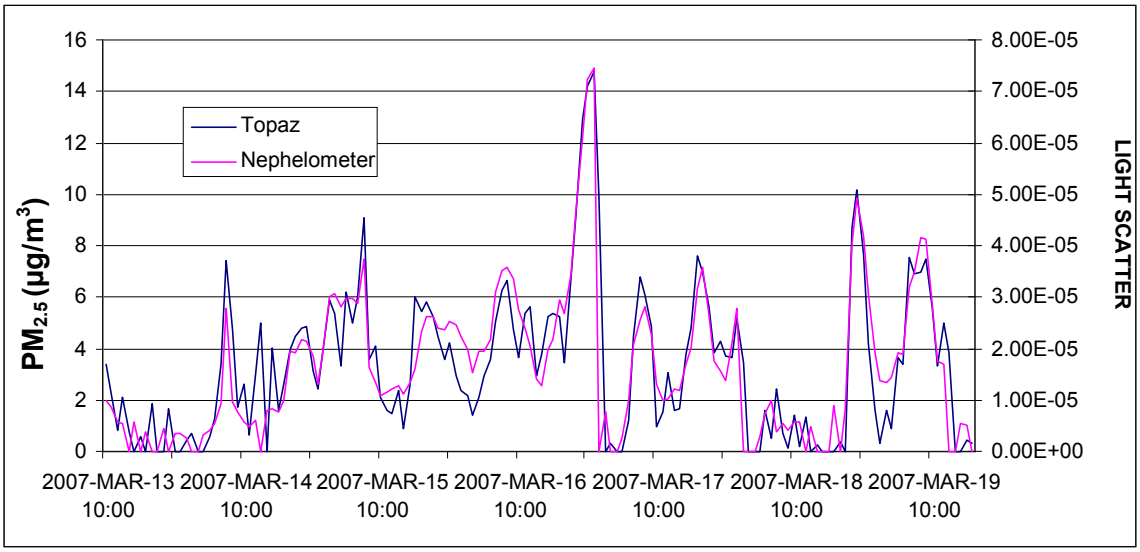


Figure 15. Hourly average $PM_{2.5}$ from the TEOM located at Topaz station and hourly average light scatter from a co-located nephelometer

[11] Fixed site measurements refer to those obtained from stations within the BC Ministry of Environment air quality monitoring network. $PM_{2.5}$ is measured continuously throughout the year using a Tapered Element Oscillating Microbalance (TEOM) monitor. This type of monitor draws in particles smaller than 2.5 microns from outdoor air which are then deposited on a filter. The change in filter mass affects the oscillating frequency of attached tubing which is used to estimate ambient concentrations.

Table 4. Regression model* for predicting nephelometer measurements (b_{sp}) from $PM_{2.5}$ observed at Topaz station.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.96E-06	9.93E-07	2.986233	0.003385
TOPAZ $PM_{2.5}$	4.17E-06	2.07E-07	20.15127	1.59E-41

*Alternatively, this model can be written b_{sp} = 2.96E-06 + 4.17E-06(TOPAZ $PM_{2.5}$)

Table 5. Nephelometer regression statistics for the model in Table 4.

Regression Statistics	
Multiple R	0.87
R^2	0.76
Adjusted R^2	0.76
Standard Error	0.00
F	406.07
Significance F	0.00
Observations	130

Table 4 shows that model coefficients and intercept are significant ($p < 0.00$), and the R^2 of the model is 0.76 (Table 5, $p<0.00$). The resulting equation for converting light scatter to $PM_{2.5}$ is:

$$PM_{2.5} = (b_{sp} - 2.96E\text{-}06)/4.17E\text{-}06 \quad (2)$$

Equation 2 provides $PM_{2.5}$ values 1.6 times lower than equation 1. Table 6 shows data for 7,424 nephelometer measurements. For each measurement, the 1 hour average $PM_{2.5}$ concentration of the 3 fixed monitors closest in time was recorded. Table 6 shows that for each nephelometer measurement, equation 1 produces a $PM_{2.5}$ value that is, on average, 3.14 $\mu g/m^3$ greater than the average of the fixed sites. Equation 2 yields concentrations 1.46 $\mu g/m^3$ lower. The Topaz monitor is located at an urban site impacted by vehicle traffic, whereas the Seattle co-located nephelometer, used to develop equation 1, was located in a woodsmoke impacted area. In addition, equation 1 was corroborated by a nephelometer co-location in Nanaimo by BC Ministry of Environment staff. Ideally, a nephelometer would be co-located with the other 2 fixed monitors in the CRD during the

winter heating seasons to support the use of either equation; however, time constraints preclude this option. Therefore, the ideal equation cannot be selected without uncertainty.

Table 6. Average $PM_{2.5}$ values for 7,424 nephelometer measurements using equation 1 and 2 as well as the fixed site 1 hour average closest in time to the corresponding nephelometer measurement.

	$PM_{2.5}$ ($\mu g/m^3$)
Fixed sites	11.35
Equation 1	15.25
Equation 2	9.53

The equation does not affect the subsequent modelling process because the ratio between $PM_{2.5}$ and b_{sp} are the same for both equations. When drawing conclusions about $PM_{2.5}$ levels, uncertainty must be communicated because equation 1 potentially overestimates values, and equation 2 potentially underestimates concentrations. Using light scatter measurements to build the model is a valid option until a $PM_{2.5}$ equation is confirmed; however, light scatter values are difficult to interpret; therefore equation 1 is chosen to err on the side of health caution.

Larson et al. (2007) adjust $PM_{2.5}$ measurements to account for changes in temperature across the heating season and because the model predicts a seasonal average. Nonetheless, no seasonal trends in $PM_{2.5}$ values were observed over 3 heating seasons. Figure 16 shows the linear trend in $PM_{2.5}$ for the 2004/2005 heating season based on the 3 monitoring stations. This pattern also holds for 2005/2006 and 2006/2007.

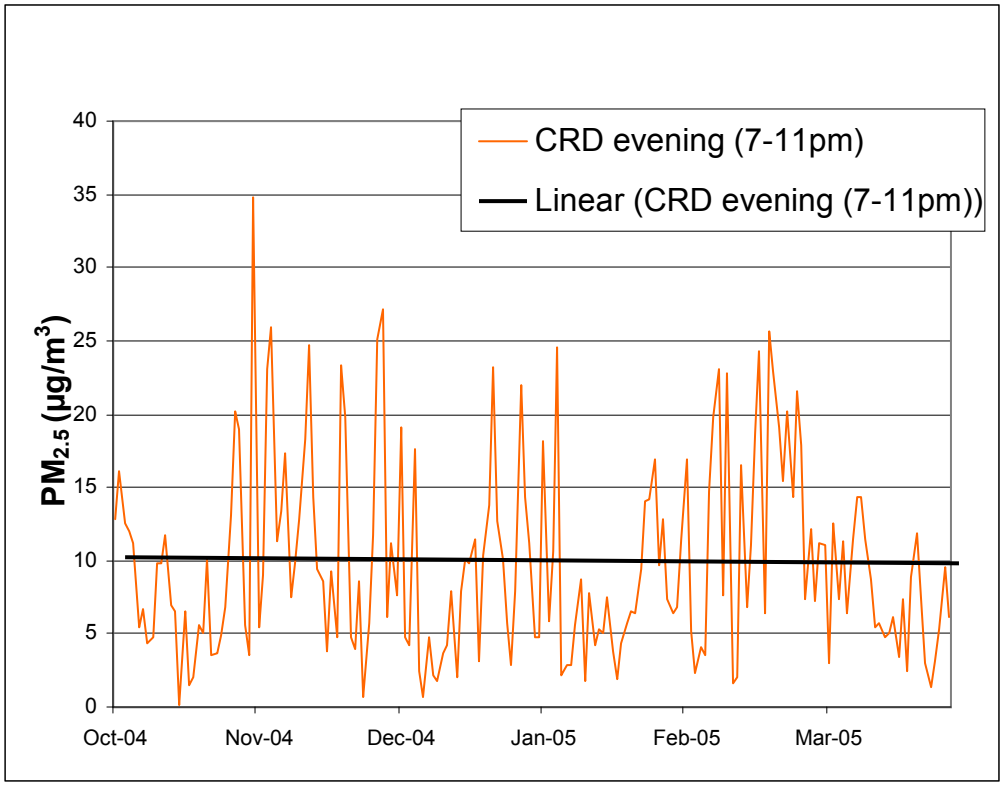


Figure 16. Evening (7-11pm) $PM_{2.5}$ and linear trend for the CRD winter heating season 2004/2005

3.2.2 Spatial and Temporal Dependence in Woodsmoke Particulate Data

In total, 14,196 individual measurements were taken over 3 heating seasons. During sampling, measurements are taken every 15 seconds resulting in both spatial and temporal dependence in the data set. Most traditional statistical analyses assume independence between data; therefore, the presence and structure of dependence was investigated. In addition, the presence of spatial autocorrelation in data affects the prediction accuracy of models and in determining which variables are significant (O'Neill et al. 2003). This section characterizes the spatial and temporal dependence in the dependent data, a precursor to modelling discussed in Chapter 4.

Semivariograms were applied to identify the distance where data are no longer autocorrelated. This distance defines the operational scale for modelling. The semivariogram is a function of distance and is based on the average sum of squared differences in attribute values for all pairs of points that are a defined distance apart. The semivariogram function is estimated by:

$$\hat{\gamma}(d) = \frac{1}{2n(d)} \sum_{\mathbf{s}_i - \mathbf{s}_i = d} \left(z_i - z_j\right)^2 \quad (3)$$

Where $\hat{\gamma}$ is the conventional symbol for a semivariogram, z_i and z_j are the attribute values of points $\mathbf{s}_i$ and $\mathbf{s}_j$. The summation is over all pairs of points that are separated by a distance d and $n(d)$ is the number of these pairs (O'Sullivan and Unwin 2003). To reduce the number of points on the semivariogram, pairs of points are binned based on their distance from each other. For each bin, the average distance (called a lag) and semivariance for all the pairs in that lag are plotted as a single point on the semivariogram (Figure 17) (Johnston et al. 2003). The number of pairs at each lag was always greater than 40 to ensure statistical reliability (Rossi et al. 1992; Burrough and MCDonnell 1998).

Typically, spatial dependence is summarized by fitting a model to the data points in the semivariogram (Figure 17). Atmospheric pollution is unevenly distributed at shorter distances making exponential and spherical models the most suitable (Moral et al. 2006). The range indicates the operational scale of the data set and has implications for subsequent analysis. For example, if the distance separating an unmeasured point and a measured point is greater than the range, the measured point cannot make a contribution to an estimate for the unmeasured location (Burrough and MCDonnell 1998). The range provides a defensible search radius for performing spatial analysis such as the use of spatial interpolators.

Semivariograms were developed and modelled using the spatial extension in Splus©. The software provided a goodness of fit measure for each model using the weighted least squares objective value. The weighted least squares objective value is the sum of the squared difference in variance estimates of the empirical and fitted semivariogram (Cressie, 1993, pg. 97). Semivariogram model results were corroborated using the 'Geostatistical Analyst' extension in ArcMap. Because the range is based on a fitted model, the range is an approximate value.

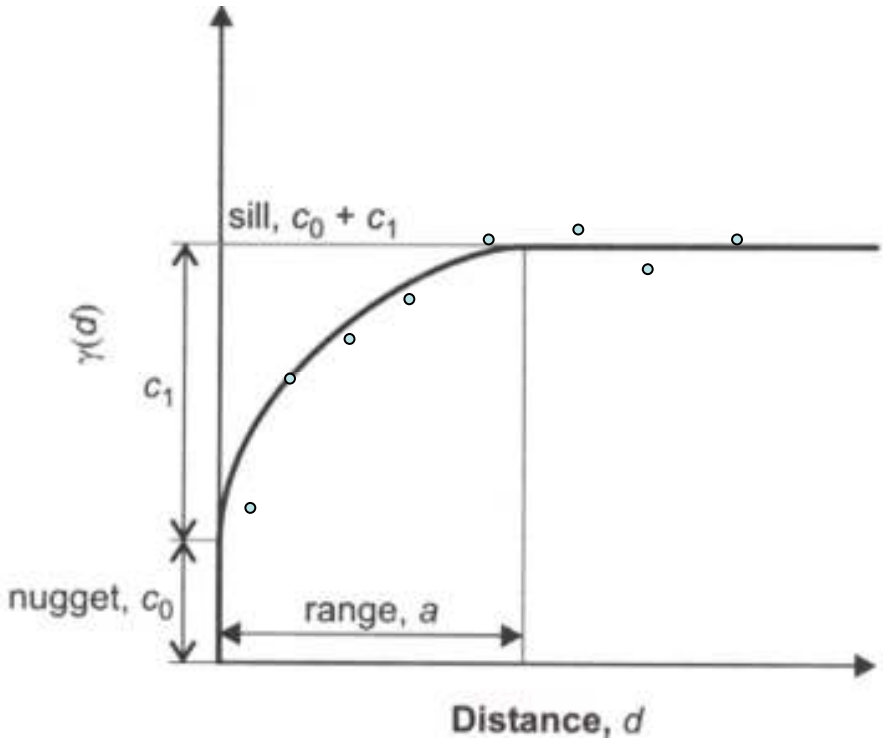


Figure 17. A spherical model fitted to a hypothetical semivariogram (adapted from O'Sullivan, 2003)

To investigate the effect of weather conditions on the range, fitted model parameters were compared for calm evenings (with wind speed less than 8 km/h), 'calm plus' evenings (wind speed less than eight 8 km/h plus clouds/fog) and non-calm conditions (wind speed greater than 8 km/h).[12]

Semivariograms require data to exhibit first order stationarity. If first order stationarity is present, smaller sub-regions have attribute values similar to the overall average (Jerrett et al. 2003a). Conversely, semivariance can change if the process is local and has a fine scale (O'Sullivan and Unwin 2003). Topography and socioeconomic characteristics vary throughout the study area which can influence the spatial correlation structure of woodsmoke. As shown in Figures 18 and 19, the semivariograms demonstrated the spherical shape characteristic of data exhibiting first order stationarity.

[12] 8km/hr was chosen based on the BC Ministry of Environment's threshold for windy conditions.

When data do not exhibit stationarity, the semivariogram has a concave-upward form (O'Sullivan and Unwin 2003).

Seasonally adjusted data collected during the first two heating seasons were appended to create one GIS layer to investigate spatial dependence for the entire data set. Data were seasonally adjusted based on the method and reasoning outlined in Larson et al (2007). Omnidirectional semivariograms using a variety of lag distances revealed no spatial dependence exists in this data set because the seasonal adjustment had the unintended effect of producing a high $PM_{2.5}$ value point from one sample evening located beside a lower value from another evening. The seasonal adjustment implemented by Larson et al. (2007) was meant to normalize the data; however, it did not. The lack of spatial dependence in this data set has implications for future spatial analysis of the woodsmoke data because the appended data contains too much noise for interpolation based on spatial dependence in the data set such as kriging. A semivariogram model for the combined data with no seasonal adjustments to the data produced a range of 2750m (CV=0.36), indicating the spatial scale of analysis for the heating season.

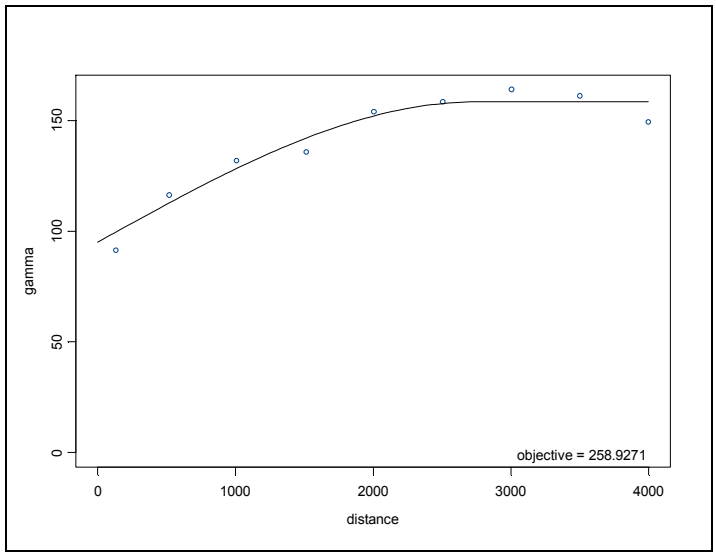


Figure 18. Global semivariogram for combined $PM_{2.5}$ data set from 3 winter heating seasons (calculated using 500m lags and a lag distance of 5000m)

To compensate for this sensitivity of semivariogram model parameters and goodness of fit to input parameters, semivariograms were calculated using various lag

and lag distances for each sample evening. According to Englund and Sparks (1991), there is confidence in consistent model parameters computed with different lag intervals. Changing the input parameters (lag size and distance) made little difference to the fitted model parameters, allowing the selection of common input parameters for comparison purposes (500m lags and a maximum lag distance of 5000m). Figure 19 is a typical semivariogram for the woodsmoke data. Three evenings were excluded from this analysis: January 12, 2006 and February 20th 2006 and March 12, 2006 because they did not produce consistent parameters. February 20th, 2006 and March 12, 2006 were not full sampling routes; therefore, the smaller area covered during those evening may have impacted the distance of spatial dependence. January 12, 2006 shows no distinguishing characteristics to suggest why it deviates from producing stable model parameters.

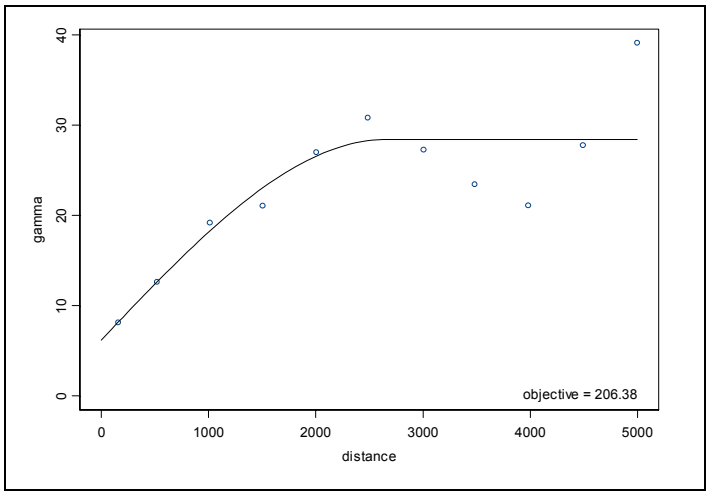


Figure 19. Semivariogram for a typical evening fitted with a spherical model (February 8th, 2005, 500m lags, 5000m lag distance).

Table 7. Summary statistics for semivariogram models fit for individual sample evenings (500m lags, 5000m distance)

Summary Statistics	Range (m)
Mean	2673
Standard Deviation	965
Median	2593
Min	932
Max	4222

The results for semivariograms computed with 500m lags to a maximum distance of 5000m show an average range of 2673m (Table 7). This finding is consistent

regardless of weather conditions (Table 8); however, there is a slight decrease in the range as wind increases. As discussed in Chapter 4, it was difficult to characterize an evening as windy/calm due to microclimates occurring throughout the CRD. This could explain the similarity in ranges.

Table 8. Mean semivariogram model parameter for different weather conditions (500m lag, 5000m lag distance).

Weather Conditions	Range (m)
Not calm	2486
Calm plus	2696
Calm	2811

For this research, the range (2,673 m) indicates the operational scale of the woodsmoke data. It also provides air quality managers with a tool to determine the spatial representation of a fixed monitoring network and provides insight into potential locations for new monitors (Figure 20).

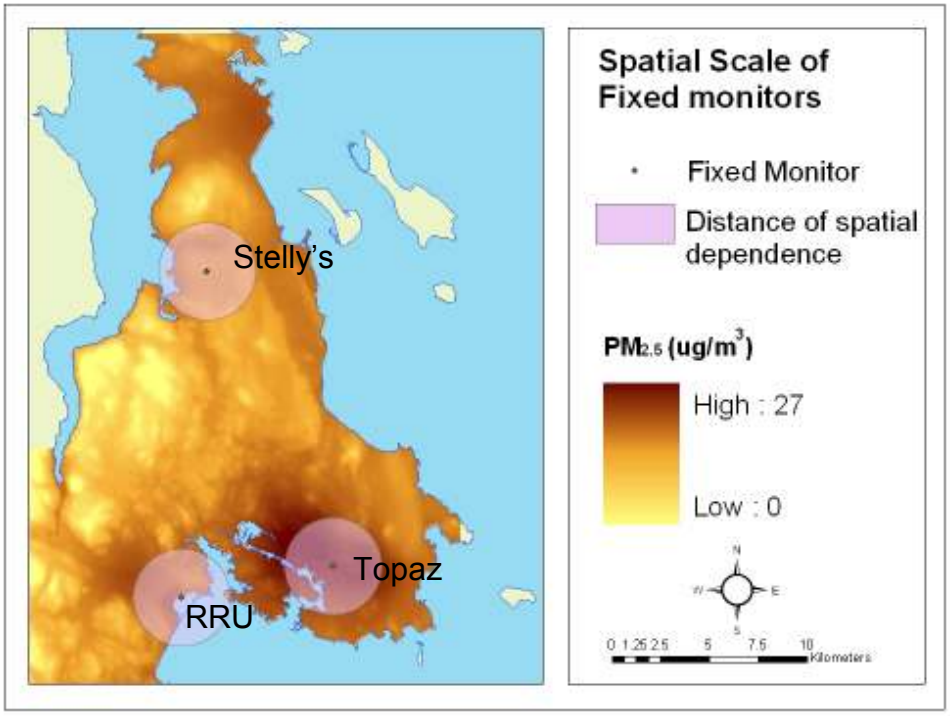


Figure 20. Fixed monitor sites in the Capital Regional District and the distance of spatial dependence in wintertime evening $PM_{2.5}$ concentrations

Serial autocorrelation occurs when measurements taken close in time are autocorrelated tending to inflate significance of tests. Figure 21 depicts the Autocorrelation Function (ACF) for a typical sample evening and demonstrates that temporal autocorrelation disappears after approximately 25 observations. This is consistant with the semivariogram range since every 25th observation is approximately 2500 (with a maximum of 4000m) apart.

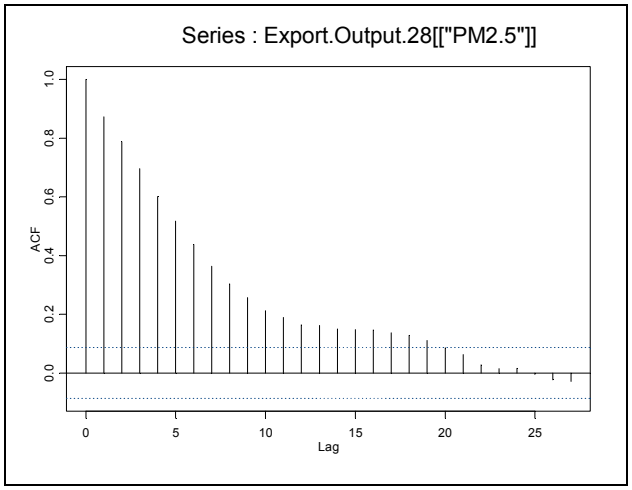


Figure 21. Temporal autocorrelation function (ACF) for $PM_{2.5}$ concentrations measured the evening of February 4th, 2007

Understanding the spatial and temporal autocorrelation in the woodsmoke data set provides useful information to ensure that modelling this data meets statistical assumptions.

3.3 Levoglucosan measurements

In the literature review (Chapter 2), neglecting to confirm woodsmoke as the source of $PM_{2.5}$ is cited as a major limitation in woodsmoke and health research. The evidence confirming the $PM_{2.5}$ being measured is derived from woodsmoke comes from a variety of sources. First, the difference between the diurnal pattern of $PM_{2.5}$ during the heating and non-heating seasons is examined. Figure 22 shows the diurnal patterns for the winter and summer seasons. The summer does not show the evening peak in concentrations after rush hour, suggesting a different source of $PM_{2.5}$ in the winter. The two peaks observed during the summer months in Figure 22 are driven by data from

Stelly's monitoring station (See Figure 23. For the location of Stelly's station refer to Figure 20.) This may be due to the proximity of this station to a fish smoking operation or it may be due to problems with the data from this station (The Ministry of Environment is investigating problems with data from Stellys for the last 2 years and has suggesting removing Stelly's data from any analysis).

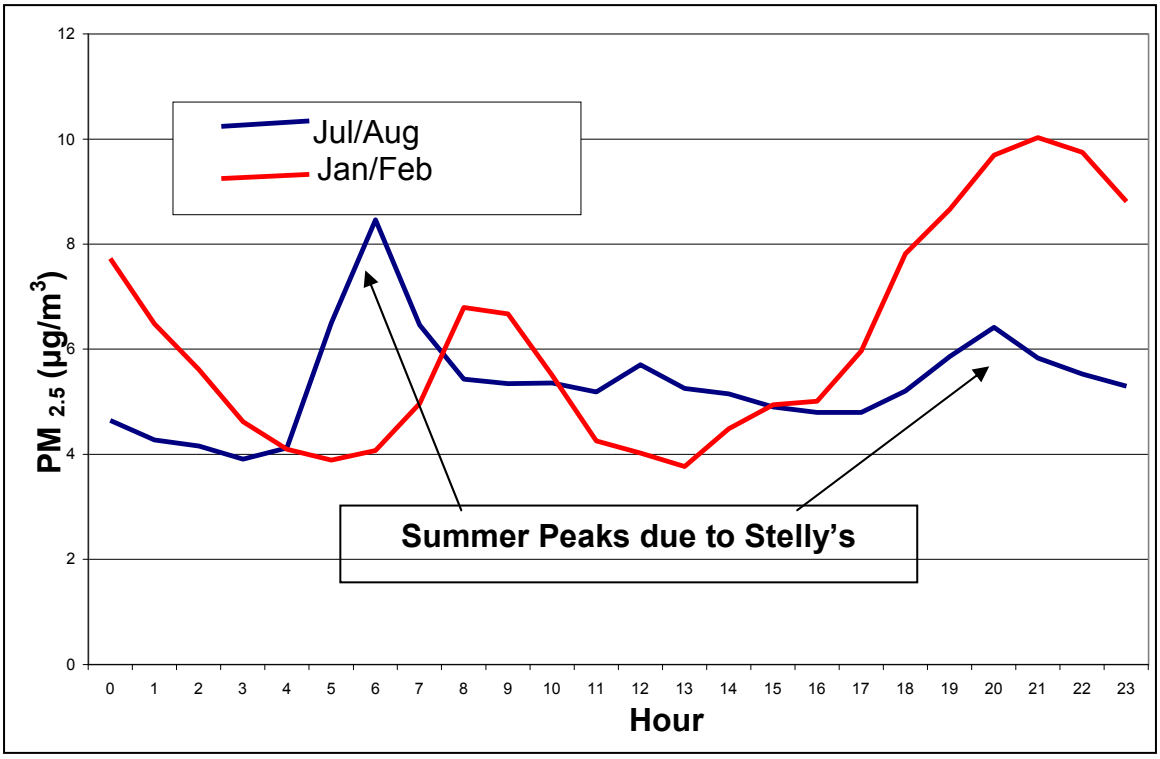


Figure 22. Average hourly $PM_{2.5}$ concentrations from 3 fixed monitors in the Capital Regional District during winter and summer months (average of 2004, 2005 and 2006/7 seasons)

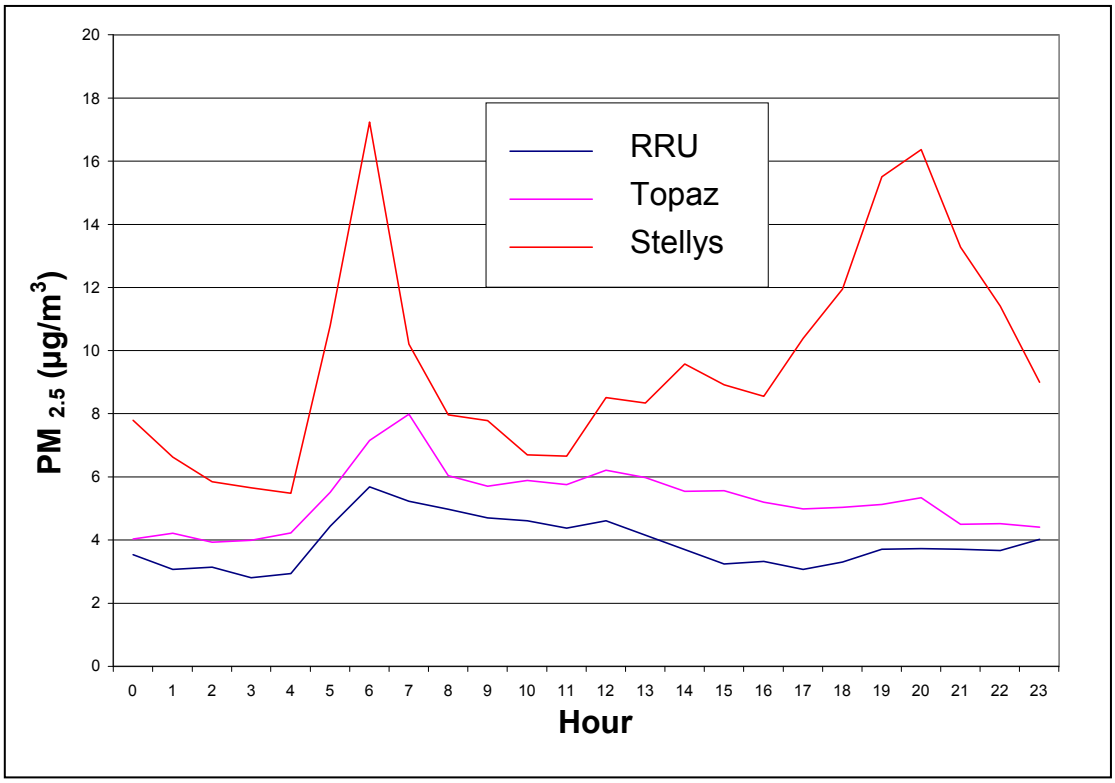


Figure 23. Average hourly $PM_{2.5}$ concentrations during the summer months in the Capital Regional District by station (average of 2004, 2005 and 2006 summers)

Another source of evidence that $PM_{2.5}$ is attributable to woodsmoke (and not vehicle exhaust, a common criticism) is from the time of measurement. Measurements were taken between 9 and 11pm, well past the commute period. During this time period, little to no traffic was on the roads.

The final piece of evidence comes from levoglucosan measurements, a unique tracer of woodsmoke (Larson and Koenig 1994). The Ministry of Environment loaned 4 Partisol Model 2000 Air Samplers with R&P $PM_{2.5}$ Cyclone inlets. The Partisol monitors ambient $PM_{2.5}$ deposited on 47mm Teflon filters with 2.0 µm pore size. The filters were sent to the University of British Columbia for levoglucosan analysis. Locations for Partisol placement were chosen to represent variation observed during the preceding heating seasons in $PM_{2.5}$ levels (Figure 24). One was placed near the coast, one in a known hot spot and 2 others placed in between those gradients. Partisols were placed greater than 2.5 km apart to avoid spatially autocorrelated results.

4 sampling periods of 12 hour durations occurred at each site (two 7am-7pm, and two 7pm-7am) to measure differences between day and night-time levels of levoglucosan. Twelve hour periods were chosen to ensure sufficient levoglucosan levels were deposited on the filter for analysis. The Partisols were then left to monitor for an additional week (24 hours a day) to examine a difference in spatial distribution of levoglucosan between the 4 sites. Two Partisols were run at a time due to a lack of resources available to operate all four at the same time. Results from the 12 hour samples are shown in Figure 24.

Figure 24 demonstrates the difference between night and day levels of levoglucosan, particularly during week 1 when weather conditions were calm and clear. The Marigold site (Site 1) was identified as a potentially high woodsmoke site and the City site (Site 2) was identified as moderately impacted by woodsmoke as observed from krigged surfaces of the measured data from the 2004/5 and 2006/7 heating seasons. The second week of sampling (the bottom two graphs) experienced unstable conditions so it is difficult to compare with the first week but evening concentrations of $PM_{2.5}$ were elevated compared with the days. The Partisol placed at the coast site (Site 3), was meant to provide background readings, but it also experienced elevated levoglucosan in the evenings compared with the day. The Langford/Colwood site (Site 4) was chosen as an additional heavily impacted site. This area had lower levoglucosan levels during the day, but night-time levels were not as high as Site 1 which is expected due to the difference in weather conditions.

Figure 25 shows the results of the week long monitoring of levoglucosan (24 hours per day). Unfortunately, the filter from the area heavily impacted by woodsmoke (Site 1) is not included because the Partisol was no longer able to pull air through the saturated filter past the third day of monitoring. The levoglucosan levels for the week shows that even during unstable conditions, woodsmoke is present in the CRD.

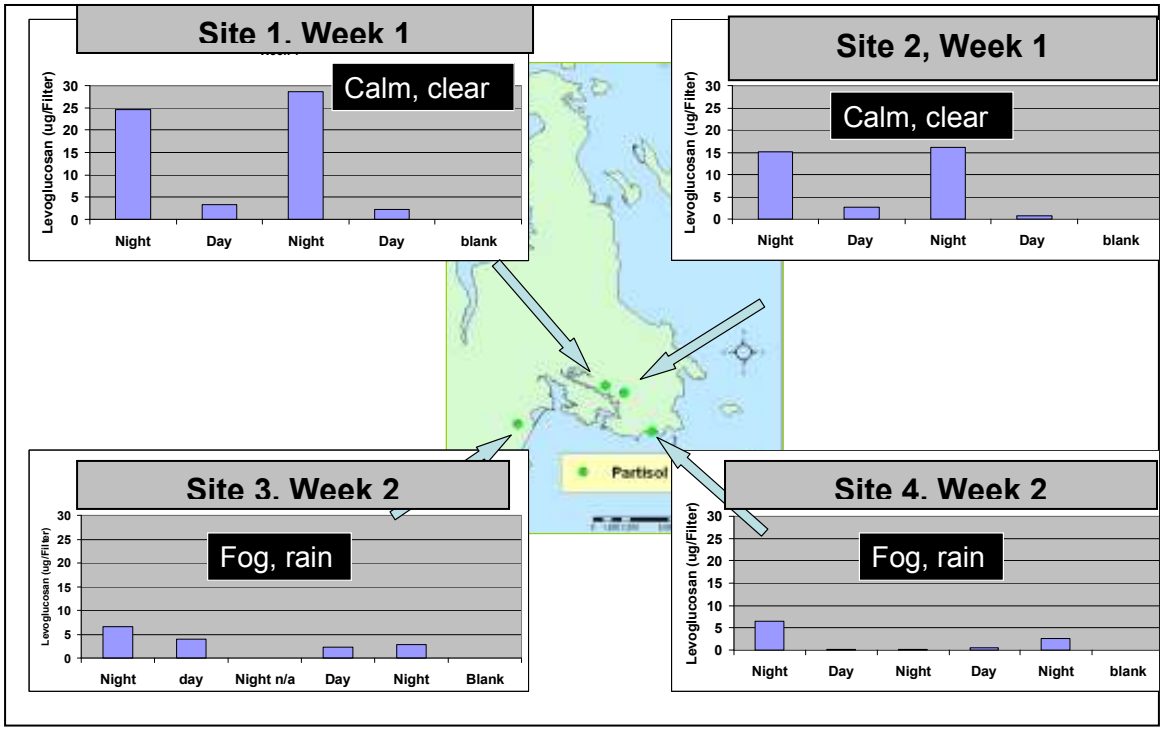


Figure 24. Levoglucosan levels for 12 hour sampling periods at Partisol locations throughout the Capital Regional District, March 2007

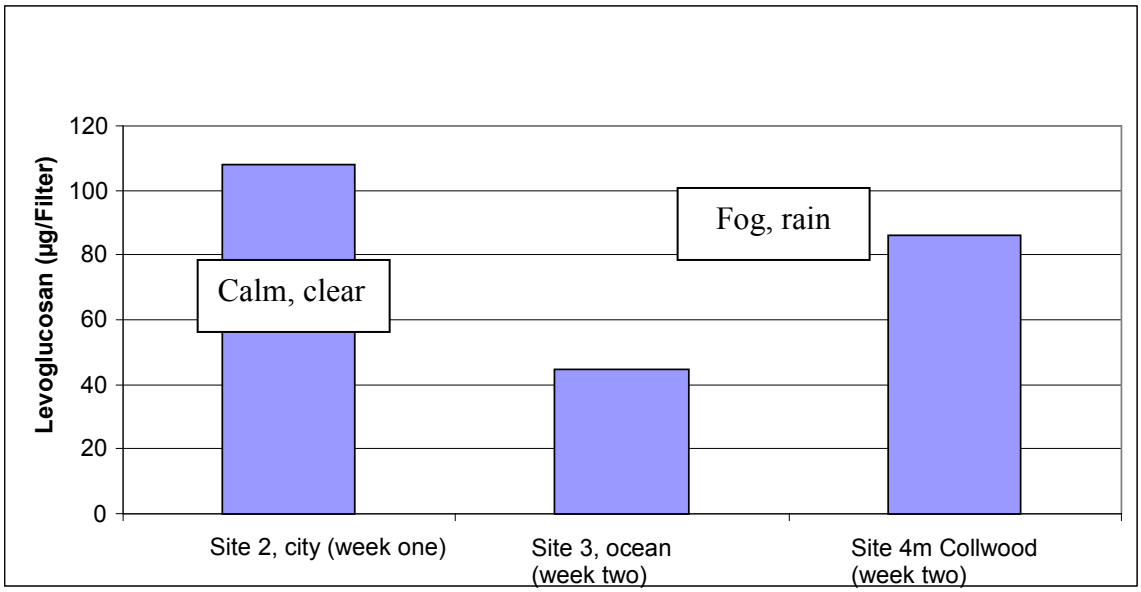


Figure 25. One week levoglucosan levels for 3 Partisol locations in the Capital Regional District, March 2007

Given the increase in $PM_{2.5}$ during the evenings in the winter heating season, the lack of mobile sources of $PM_{2.5}$, the elevated levels of levoglucosan during the evenings it is argued that the $PM_{2.5}$ measured during the monitoring campaign is attributable to woodsmoke.

3.4 Independent Variable Data Development

This section describes data development for independent variables (also called predictor variables) for woodsmoke modelling. Literature is sparse on the determinants of residential wood burning; therefore, the potential independent variables developed for modelling woodsmoke come from an understanding of air pollution dynamics, published research related to modelling woodsmoke and common knowledge regarding the drivers of residential wood burning (Table 9). Previous woodsmoke model variables included wind speed (Larson et al. 2007), topography (Larson et al. 2007), elevation (Tian et al. 2004), degree of urbanization (Tian et al. 2004), the presence of wood stoves (Larson et al. 2007), temperature (Tian et al. 2004), and population demographics (Larson et al. 2007).

Table 9. Potential independent variables for a woodsmoke model and the theory for their inclusion.

Theory	Potential Independent Variables
Need people to burn wood	Population density
Need wood to burn	Access to firewood, urban/rural
Need a place to burn wood	Fireplace, woodstove, housing type and age
Needs to be cold enough to require home heating	Temperature, heating degree days
Need to choose wood over other heating options	Income, primary heating source, housing type/age, degree of urbanization, cultural preference
Atmospheric conditions influence outdoor concentrations	Wind speed, wind direction, precipitation, distance to ocean, Environment Canada's stability index

3.4.1 Topographic Data

Tian et al. (2004) found elevation was the most significant predictor of potential RWB activity and Larson et al. (2007) cite the influence of drainage on the concentration of woodsmoke on cool, calm evenings. GPS elevation data were available for each measurement as well as a Digital Elevation Model (DEM) for the CRD. GPS elevation had a low but significant negative relationship (R= -0.12, α=0.01) with $PM_{2.5}$ and was a significant variable in all of the approaches to woodsmoke modelling. There was little correlation (R = -0.02, α=0.01) between $PM_{2.5}$ and the elevation extracted from the Digital Elevation Model (DEM) based on TRIM data (accuracy +/- 10m).

A watershed variable (*WSVAR*) was conceptualized to incorporate the catchment basin approach in Larson et al. (2007) using the DEM for the CRD. The relative height of each $PM_{2.5}$ measurement (*DEMelev*) to the maximum elevation in the basin (*WSmax*) was calculated to determine the effect of elevation within the watershed. The watershed variable was calculated as follows:

$$WSVAR = ([\text{WSmax}] - [\text{DEMelev}]) / ([\text{WSmax}] - [\text{WSmin}]). \quad (4)$$

Where a value of 0 is close to the top of the watershed and 1 is close to the bottom. The watershed maximum and minimum (*WSmin*) were calculated using 'Zonal Statistics' in 'Spatial Analyst'. The zone field was the 3 by 3km water basin layer created using the 'watershed tool' in ArcMap based on the DEM. No association was found between the watershed variable and $PM_{2.5}$ and this variable was not significant in the modelling process. Since elevation is significant and the catchment basin is not, perhaps the CRD serves as one large watershed when it comes to the scale of woodsmoke, or else the DEM was a poor interpolation of TRIM elevation data.

3.4.2 Geographic Data

Latitude and longitude coordinates were recorded for every $PM_{2.5}$ measurement. The X and Y coordinates were modelled using linear regression for a directional trend in $PM_{2.5}$. Table 10 shows the regression model results. Although not heavily influential in predicting $PM_{2.5}$ (R^2 = 0.07, p<0.00, n=596), these results indicate a slight increase in

$PM_{2.5}$ from west to east (positive X coefficient) and a stronger decrease in $PM_{2.5}$ from south to north (negative Y coefficient).

Table 10. Regression output for directional trend

Model Variable	Coefficient	Std.Error	t-Statistic	Probability
Intercept	-89.55	73.79	-1.21	0.23
X coordinate	0.00014	6.04E-05	2.36	0.019
Y coordinate	-0.00018	4.00E-05	-4.597	6.3E-06

3.4.3 Meteorological Data

Local meteorological conditions and topography affect the spatial distribution and concentration of air pollutants (Briggs et al. 1997). Ambient air pollutant concentrations are inversely related to volume of air into which emissions are mixed, which is a function of mixing height and wind speed (Cupitt et al. 1994). Larson et al. (2007) capture this effect of mixing height through the use of hydrological catchments trapping woodsmoke on cool, calm evenings. Nevertheless, this is limiting because the entire catchment has the same value, meaning the top of the basin is given the same value as the bottom of the basin where air pollution accumulates resulting in the ecological fallacy and exposure misclassification. Cupitt et al. (1994) and Tian et al. (2004) found woodsmoke was related to temperature as RWB varies with the need for heating, this relationship was also (Table 11, R=0.2, α=0.01).

Environment Canada has data for temperature (hourly average, °C), heating degree days (for the afternoon of a specific date) and wind speed (hourly average, m/s) for the Victoria weather station available for download.[13] Figure 26 shows the location of the monitor at the Victoria airport. Temperature, heating degree days and wind speed measurements were assigned to each data point based on the hourly average closest in time to the nephelometer measurement. For example, a measurement taken at 9:45pm was assigned the hourly average for 9pm and a measurement taken at 10:15 was assigned the hourly average for 10pm. The spatial resolution of the Environment Canada data are poor because they come from one meteorological station at the Victoria airport. This

[13] http://www.wunderground.com/history/airport/CYYJ/

scale is appropriate for temperature because data at a finer spatial resolution did not improve the association with $PM_{2.5}$ (Table 11). Nevertheless, this scale is not appropriate for wind speed. The Environment Canada monitoring data, which are common in most BC communities, were compared with data of a higher spatial resolution from the VWN which had anywhere from 24 to 82 stations operating throughout the CRD for the duration of the measurement campaign.

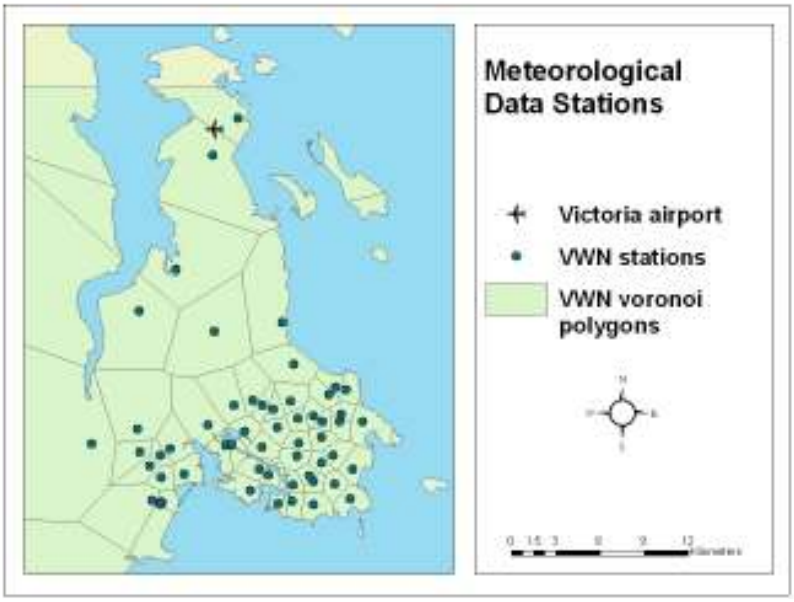


Figure 26. Spatial resolution for the Victoria Weather Network and airport monitor

Figure 26 shows the spatial resolution of the VWN at a given point in time. Temperature (°C) and wind data speed (m/s) were available for each clock minute. Temperature and wind speed closest in time and distance, as determined by Voronoi[14] polygons for each station, were assigned as attributes for each $PM_{2.5}$ measurement. The previous hour's wind speed average was calculated for each measurement but did not improve performance over wind speed at the closest minute. The higher resolution wind speed data from the VWN improved associations with wind speed (R = -0.22, α=0.01) over the airport data (R=-0.11, α=0.01), however, temperature from the airport (R=-0.17, α= 0.01) showed the best correlations with $PM_{2.5}$ over heating degree days or the higher

[14] Voronoi, or theissen, polygons define the area around a point in a pattern of point objects such that all locations within the polygon are closer to that point than to any other point in the pattern (O'Sullivan and Unwin 2003).

resolution temperature data from the VWN (Table 11). Therefore, temperature and wind speed operate at different scales with wind speed varying at a local scale impacting pollutant concentrations that is not captured by the monitor at the airport. Temperature, influential in determining when people will burn wood, varies at the regional scale in the CRD.

Environment Canada data related to Heating Degree Days, the Venting Index and Mixing Height were downloaded[15] for each sample evening (they are available on a daily basis) but had no significance in the modelling process.

3.4.4 $PM_{2.5}$ Data from Fixed Monitors

The BC Ministry of Environment provides $PM_{2.5}$ hourly averages ($\mu g/m^3$) for download from the Air Quality Index Database website. Figure 27 shows the Voronoi polygons for each fixed site monitor. The hourly average of the 3 fixed sites (*TEOM2*) as well as the measurement from the closest station (*TEOM*) at the closest hour was recorded for each $PM_{2.5}$ measurement. The average of the 3 stations showed a higher correlation with $PM_{2.5}$ (R=0.54, α=0.01) than the concentration from the closest station (R=0.50, α=0.01). The data from the fixed monitors were included as a background measure of $PM_{2.5}$, with the idea that including other variables in the model would improve the estimates from the closest monitor. Including the *TEOM* variable replaced wind speed and temperature as significant variables in the models.

[15] http://www.env.gov.bc.ca/epd/epdpa/venting/polled_data/

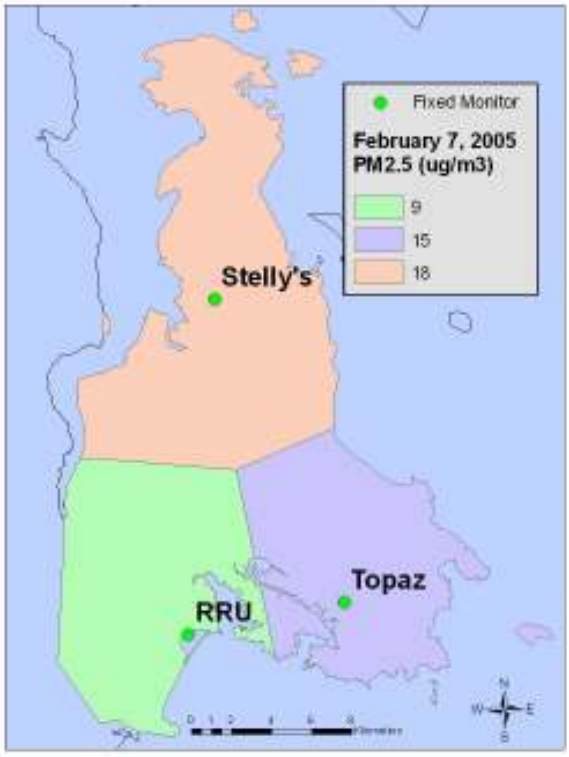


Figure 27. Voronoi polygons for the fixed site monitors in the Capital Regional District

3.4.5 Spatial Property Assessment Data (SPAD)

A major constraint to improving exposure assessment is the lack of available data at fine spatial resolutions (Setton et al. 2005). Property assessment data are available at the individual parcel level providing a resource for characterizing exposure at a fine spatial resolution (see the left hand panel in Figure 28). A limitation of this data arises from property size differences as rural parcels tending to be larger than urban parcels; nevertheless, the resolution is better than census data (Section 3.4.6). Spatial property assessment data for the CRD is maintained by the BC Assessment Authority and has been geocoded based on unique property identifiers (Setton et al. 2005).

Spatial property assessment data are available for 2004 and include several variables examined such as property size, property use code, land use code, land value, year built, total square footage, predominant heating type and number of fireplaces. Property use codes and land use codes were used to identify residential locations. Since this is a model for residential wood burning, only residential parcels were selected. To avoid the inclusion of multiple dwelling units skewing calculations, only single dwelling residences were used. Total value, a proxy for income, has a significant negative

relationship (*TVD25P*, *R*=-0.26, α=0.01) to $PM_{2.5}$ that is stronger than any correlations relating to income from the census data. Year built was hypothesized to indicate houses that are likely to have wood fireplaces. The density of houses built before 1951 and houses built after 1982 were chosen as thresholds because those years represent the first and third quartile of building ages. The density of houses built or improved before 1951 has a positive relationship (*R*=0.26, α=0.01) with $PM_{2.5}$.

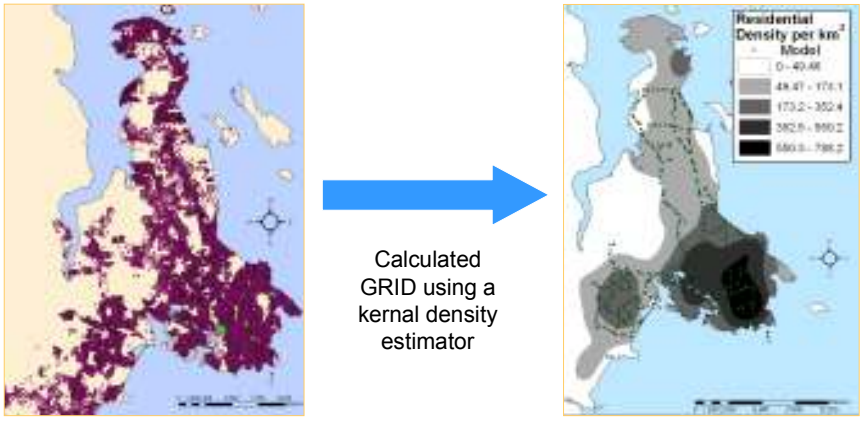


Figure 28. Example of converting independent variable SPAD data sets to grids

According to Tian et al. (2004), estimating the number of households with wood stoves or fireplaces is essential for estimating emissions of $PM_{2.5}$. Fireplace densities were calculated alone and in conjunction with primary heating type to rule out gas fireplaces. Although fireplace density has a positive significant relationship with $PM_{2.5}$ (*R*=0.19, α=0.01), it is collinear with residential density which shows a stronger relationship to $PM_{2.5}$ (*R*=0.21, α=0.01), therefore, fireplace density was not a significant variable when residential density was included. Residential density became insignificant in models that included the density of houses in the 25th percentile for value (*TVD25P*).

Independent variable data sets derived from SPAD vector data sets in a GIS were converted to raster (also referred to as grids) for ease of model development.[16] Vector

[16] Models were also developed using the vector data sets which performed marginally better than using the rasterized data sets; however, the ease of model development using raster outweighed the marginal improvement in model performance.

data sets were converted to raster using kernel density estimation, a point density technique that converts point patterns to a continuous density surface based on the number of points in a given area (O'Sullivan and Unwin 2003). A cell size of 50m was chosen for the rasters based on the resolution of points in the SPAD. Each point represents a property, which could be closer than 50m; however, computing power precluded a smaller resolution. The optimal operational and analysis scale for the woodsmoke particulates is approximately 2700m and informs how to combine particulate data with other sources of spatial data (Section 3.3). Examining the spatial dependence structure was typically observed at 1100m. To determine the ideal search radius which was hypothesized as being between 1100 and 2700m, a selection of independent variables were converted to rasters using a kernel density estimator with search radii ranging from 250m to 5000m (Figure 28), a method described in Henderson and Brauer (2005).

3.4.6 Socioeconomic Data from the Census

The census provides data that are commonly used in modelling environmental and health phenomena. Census data are maintained by Statistics Canada and were downloaded through the University of Victoria's Library Gateway. The most recent census was taken in 2001 and it is updated every five years. The 2006 census data was not released in time to incorporate into the thesis.

Census data are spatial referenced and are available at the Dissemination Area (DA) level (Figure 29). DAs include one or more neighbouring blocks with a population of 400 to 700 people (compare with 1 point per residence for SPAD). As shown in the figure, DAs vary in size throughout the study area representing a potential limitation due to the MAUP and the loss of variation within each DA. Variables from the census included: population, population in manufacturing employment, immigration, income, education and mortality. Demographic data were used to generate density measures. Demographic variables identified in Larson et al. (2007) include population in manufacturing employment and the number of immigrants. Other variables investigated include those related to income which may have an effect due to the lower cost of using wood as fuel for heating and education as a proxy for income levels or reflecting awareness of health impacts. Although some census variables (percent low income for

example) did show associations with $PM_{2.5}$ they were rendered insignificant by the inclusion of similar variables from the SPAD and, as a result, no census data were used in modelling woodsmoke (with the exception of the Larson Model).

Figure 29. Census dissemination areas in the Capital Regional District, 2001

Census variables were developed for use in regression modelling by converting the DA polygons to raster using Spatial Analyst (Figure 30). Focal statistics were used to determine the average value of census variables (i.e., mortality, low income, education) within a 3km radius of each $PM_{2.5}$ measurement (Figure 31). Associations with census variables were also investigated using the DA intersecting the $PM_{2.5}$ measurement; however, associations declined.

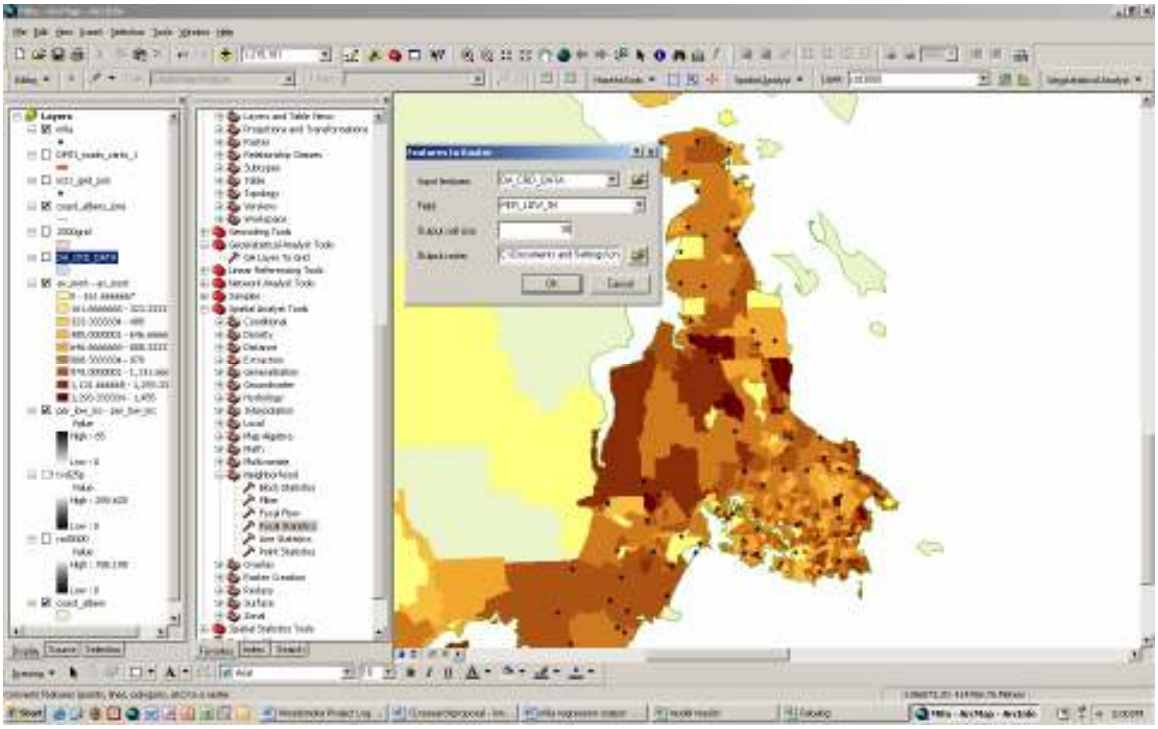


Figure 30. Conversion of census DA to raster for the Capital Regional District

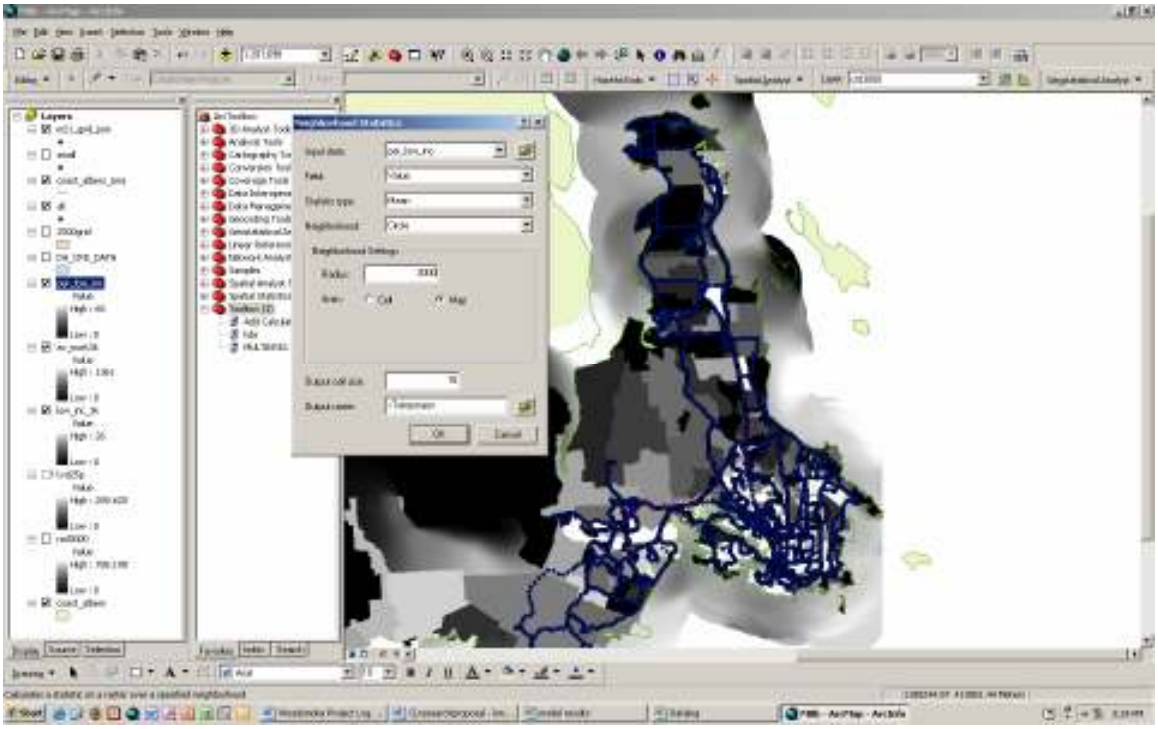


Figure 31. Focal statistics to obtain average census variables within a 3km radius of each data point in the Capital Regional District

3.4.7 Road data

To determine if road density is a predictor of $PM_{2.5}$ during the evenings, a road density variable was generated using a street network file distributed by DMTI Spatial Inc. The road categories, including highways, major and residential roads, were selected from this file. The sum of the road length (m) within 3 km of $PM_{2.5}$ measurements were calculated using 'Spatial Analyst' 'Neighbourhood Statistics'. There was a positive relationship between road density and $PM_{2.5}$ (R=0.32, α=0.01); however, this variable is collinear with residential density (R=0.98, α=0.01) and is dropped in the modelling process as insignificant when residential density or density of low value housing (*TVD25P*) is included.

DMTI data were also used to calculate the distance to the ocean using 'Near' in 'Analysis Tools.' Distance of the $PM_{2.5}$ measurement to the ocean had a weak relationship (R=0.08, α=0.01) but the variable was consistently significant in models. Amount of open areas within 3km of each measurement also showed a similar weak relationship (R=0.08, α=0.01); however, it was an insignificant model variable and the positive direction of the relationship is counter to the hypothesis that open areas result in less woodsmoke.

3.4.8 Selection of model variables

Data values from the independent variable grids were extracted to the $PM_{2.5}$ data points, the dependent variable (see Figure 28, right hand map), and the correlation between $PM_{2.5}$ and the independent variable value was examined for independent variable layers created using different search radii. Figure 32 shows the correlation between $PM_{2.5}$ and residential density calculated using a variety of search radii. Although Figure 32 is one example, the same relationship was observed for the other independent variables developed. As a result, a search radius of 3000m was selected for independent variable development since it represents the maximum correlation, beyond which no new information is gained and it is supported by the spatial dependence in the particulate data.

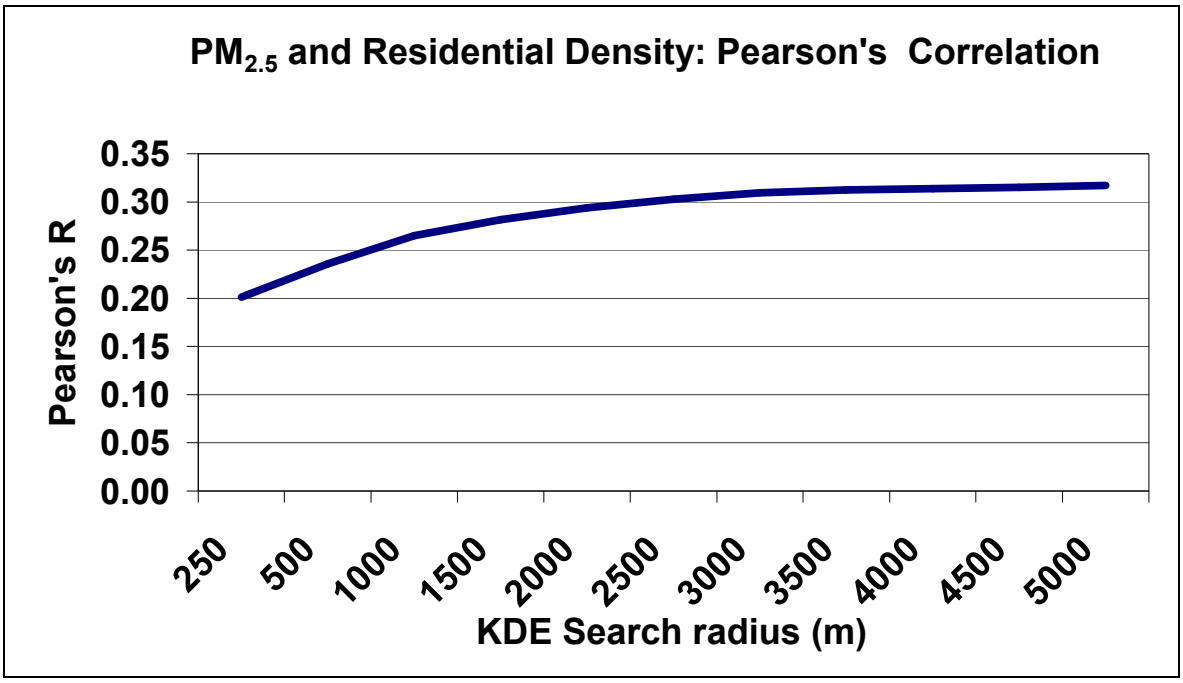


Figure 32. Correlation between $PM_{2.5}$ and residential density calculated using a variety of search radii

GPS elevation, location coordinates, meteorology and background $PM_{2.5}$ concentrations were the only independent variable data sets that were not converted to raster. The GPS logged latitude, longitude and elevation of each measurement during the mobile monitoring campaign; therefore, this data was already associated with each $PM_{2.5}$ measurement and did not need to be extracted from rasters.

Meteorological data associated with each measurement were also obtained. There were two sources of meteorological data, Environment Canada and the Victoria Weather Network (VWN), and their development is discussed in Section 3.4.3. The hourly average $PM_{2.5}$ concentrations from each fixed monitoring site monitoring were downloaded from the BC Ministry of Environment website[17] for every hour of the heating season. The average $PM_{2.5}$ from the fixed monitoring sites and from the closest monitor for the closest hour in time were also added as attributes to the $PM_{2.5}$ data.

[17] http://www.elp.gov.bc.ca:8000/pls/aqiis/archive.report?cookie=PBLHBJNJCP&ID=01&report=extract

Table 11 summarizes the independent variables investigated, the data source, the method used for development, as well as the correlation with measured $PM_{2.5}$ calculated in SPlus. The correlation values are approximate because the correlations varied depending on the approach for selecting dependent variable data for model development (discussed in Chapter 4). Nonetheless, the direction and significance listed in Table 11 is indicative of the correlations observed. The list in Table 11 is not exhaustive, additional variables were developed for some variable groups; however, they are not included in the interest of brevity. Only the independent variable showing the highest correlation from a particular group is included where several similar variables were developed. For example, several census variables related to education and income are not included in Table 11 because percent low income was collinear with many of these variables and showed the highest correlation with $PM_{2.5}$.

Table 11. Independent variables and correlations with $PM_{2.5}$, the dependent variable.

Variable groups (units)	Independent variables (nomenclature)	Data Source	Development method	Correlation (Pearson's R)*
Topographic (m)	GPS Elevation (ELEV)	GPS	n/a	-0.12
	DEM (DEMELEV)	DMTI	n/a	-0.02
	Catchment basin (WSVAR)	DMTI	Spatial analyst – Zonal statistics	0.05
Geographic	Latitude (X)	GPS	n/a	0.05
	Longitude (Y)			-0.19
Meteorology (°C temperature, m/s wind speed, degrees wind direction, m venting index and mixing height)	VWN temperature (TEMP)	Victoria Weather Network (VWN)	Closest in time from closest station (as determined by voronoi polygons) to the nephelometer measurement	-0.16
	VWN wind speed (WS)			-0.22
	VWN hourly wind speed (WSA)			-0.10
	VWN wind direction (WD)			-0.14
	Hourly temperature (TEMP2)	Environment Canada	Closest hour to the nephelometer measurement	-0.17
	Heating Degree Days (HDD)		Daily value	-0.20

Variable groups (units)	Independent variables (nomenclature)	Data Source	Development method	Correlation (Pearson's R)*
	Hourly wind speed (WS2)		Closest hour	-0.11
	Hourly wind direction (WD2)		Closest hour	-0.06
	Venting Index (VENTI)		Daily value	-0.07
	Mixing Height (MXGH)		Daily value	-0.10
Background $PM_{2.5}$ concentration ($\mu g/m^3$)	Hourly average of 3 fixed monitors (TEOM2.)	Ministry of Environment	Closest hour	0.54
	Hourly average from closest station (TEOM)		Closest station (voronoi polygons), closest hour to nephelometer measurement	0.50
Residential Density (residences/km^2)	Residences per km^2 (RZD3000)	SPAD	Kernal Density Estimator	0.21
	Number of residences in 3km of PM measurement (COUNT1)		Spatial join of 3km buffer and SPAD data for single residential addresses	0.28
	Number of residences per square area of 3km buffer clipped with coast (COUNT)		Spatial join of 3km buffer and SPAD data clipped with coastline	0.28
Income (%) and property value ($)	Percent low income (PER.LOW.INC)	Statistics Canada DA	Spatial analyst – Focal Statistics	0.14
	Density of houses in 25^{th} percentile of Total Value (TVD25P)	BC Assessment	Kernal Density Estimator	0.26
	House value per km^2 (VAL.AREA)		Total Value/Area.	0.07
Mortality	Average mortality in 3 km (AV.MORT)	Statistics Canada DA	Spatial analyst – Focal Statistics	-0.09
Distance to Ocean (m)		DMTI	Analysis tools – NEAR	0.08
Land use (ha)	Open (OPEN)	DMTI	Spatial analyst – Focal Statistics	0.08

Variable groups (units)	Independent variables (nomenclature)	Data Source	Development method	Correlation (Pearson's R)*
Fireplace Density	Fireplaces per km^2 (FP3000)	BC Assessment	Kernal Density Estimator	0.19
Building Age (houses/km^2)	Houses built/improved after 1982 - 75^{th} percentile (BAD82.3K)	BC Assessment	Kernal Density Estimator	0.10
	Houses built/improved before 1951 – 25^{th} percentile (BAD.51.3k)			0.28
Road Density (length/km^2)	Length of roads classified as RD1, RD2 and RD3 within 3km (RD_3KM)	DMTI	Spatial Analyst – Neighbourhood Statistics	0.32

*All correlations are significant at α=0.01

Since environmental variables tend to be lognormallly distributed and Pearson's *R* assumes a normal linear relationship, Pearson's *R* may not be an appropriate statistic to use to measure correlation. Variables were transformed to become more normally distributed by taking the square root in addition to performing a non-parametric test (Spearman's rho) to determine the effect of non-normality. The distributions of the variables did not affect the general strength and direction of associations, or model performance as discussed in Chapter 4, although Spearman's rho did tend to provide slightly smaller associations, by about 0.05. Since modelling results were robust to changes in variable distributions, using statistical tests that assume normality was deemed appropriate.[18] Appendix A includes variable distributions as well as scatter plots with $PM_{2.5}$ for selected variables.

Variables for modelling woodsmoke were selected based on the following criteria:

- A significant association with woodsmoke (Pearson's *R* values);
- No collinearity between independent variables;

[18] This relates to data sets where there was an attempt to remove spatial and temporal dependence prior to statistical analysis. When they were not removed, correlations and model performance were lower, counter to the published literature stating autocorrelation inflates performance.

- Independent variables are significant at the 1% confidence interval;
- Inclusion of independent variables improves model performance significantly; and,
- Independent variables fit established theory.

ELEV, *TEOM2* and *TVD25P* were selected for modelling based on these criteria. *TEMP2* and *NEAR.DIST* can be included to marginally improve models; however, they are excluded for ease of comparison.

The next chapter discusses the methods and results of different approaches to modelling exposure to woodsmoke.

Chapter 4:
Spatial Modelling of Woodsmoke

This chapter outlines the development and evaluation of exposure models. The first section of the chapter (Section 4.1) examines the typical epidemiological scenario for characterizing exposure. The Larson Model is evaluated in Section 4.2, which moves beyond the typical epidemiological scenario. The Larson Model is evaluated with data excluded from the model development process. Drawing from the strengths and weaknesses of the Larson Model, a different approach to modelling woodsmoke at a finer spatial resolution is explored (Section 4.3). Finally, Bayesian theory is applied to the woodsmoke model to determine if exposure modelling benefits from this approach. The chapter concludes with a discussion of the models (Section 4.4) with recommendations for a model to use in the spatial characterization of health risk.

4.1 Baseline Scenario: Concentrations from Fixed Monitors

As discussed in the literature review, most studies relating air pollution and health use concentrations from fixed site monitors, and often only one, to characterize exposure for individuals or a population. This section examines the woodsmoke model produced under this scenario. Figure 33 shows the 3 fixed monitors in the CRD and their associated Voronoi polygons. The population falling within each monitor's polygon are assigned the concentration from that monitor. Figure 33 shows a typical evening where the population within the Stelly's monitoring site polygon is assigned a concentration of 18 $\mu g/m^3$, the Topaz population is assigned 15 $\mu g/m^3$, and the Royal Roads University (RRU) population is assigned 9 $\mu g/m^3$.

A spatially stratified random sample of points from 15 sample evenings (i.e., randomly selecting one point from each 2.5 km grid cell for each sample evening to remove spatial and temporal dependence, n=595), has a correlation of $R = 0.51$, $\alpha = 0.01$, between the $PM_{2.5}$ value from the closest monitor and the nephelometer measurement. Figure 33 shows concentrations from a typical evening, and the potential for exposure misclassification. For example, the RRU population is assigned the lowest concentration

which fails to capture a measured woodsmoke hot spot a few kilometers northeast of the RRU station (see figure 34 in Section 4.2).

Table 12 shows the regression results using the values from the nearest monitor to predict measured $PM_{2.5}$ ($R^2 = 0.25$, $p<0.00$). Better results are obtained if the average of the 3 fixed sites is used to characterize exposure (Table 13, $R^2 = 0.33$, $p<0.00$).

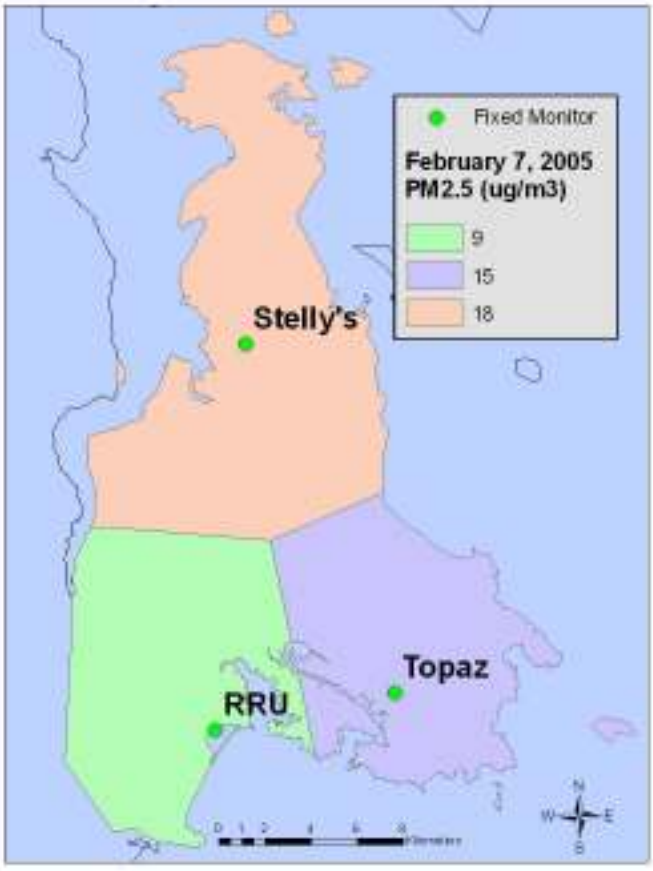


Figure 33. Evening (9-11pm) average $PM_{2.5}$ concentration for fixed monitor's Voronoi polygons in the Capital Regional District

Table 12. Regression model results using nearest monitor to predict measured $PM_{2.5}$.

Model*	Value	Std. Error	t value	Pr(>|t|)
Intercept	8.54	0.63	13.47	0.00
TEOM	0.43	0.04	10.63	0.00

*F-statistic: 112.9 on 1 and 337 degrees of freedom, the p-value is 0

Table 13. Regression model results using average of 3 monitors to predict measured $PM_{2.5}$.

Model*	Value	Std. Error	t value	Pr(>|t|)
Intercept	5.51	0.75	7.37	0.00
TEOM2.	0.75	0.06	12.87	0.00

*F-statistic: 165.5 on 1 and 337 degrees of freedom, the p-value is 0

4.2 Kriging

Kriging is a statistical interpolation technique that uses the underlying spatial structure in data to develop a continuous surface of attribute values based on point data (O'Sullivan and Unwin 2003). Kriging is a method designed to retain the measured data values, in contrast to linear regression, where the line of best fit does not pass through all the measured points (O'Sullivan and Unwin 2003). Another difference between kriging and regression is that kriging exploits spatial and temporal dependence in the data, whereas, regression requires independence.

Kriging assumes first order stationarity in the data set. There is a minor increase in $PM_{2.5}$ southerly trend (Section 3.5, R^2=0.07, p<0.01); however, it is assumed that the trend is weak enough to merit the use of ordinary kriging. In addition, woodsmoke data semivariograms have the characteristic spherical shape indicative of first order stationarity (O'Sullivan and Unwin 2003).

The krigged surfaces based on measured $PM_{2.5}$ data were developed in ArcMap's 'Geostatistical Analyst' extension. The 32 sample evenings were krigged using the 'Geostatistical Wizard'. Each evening was krigged using the semivariogram parameters based on the analysis in Section 3.3 (500m lags, 2700m range). The 'Mosaic to New Raster' tool in 'Spatial Analyst' was used to mosaic the 32 evenings together and find the average $PM_{2.5}$ concentration where layers overlapped. Since the 32 sample routes did not cover identical areas, some areas were sampled more than others (Figure 34). The 15 similar routes were mosaiced, showing the average cell value in Figure 35.

Kriging performs well (R=0.84, α=0.01, n=517) when predicting $PM_{2.5}$ values for the sample evening used to create the surface. For example, the krigged surface for February 7, 2005 predicts the measured values for that evening well because it is a method that retains the original measurements. Nevertheless, it does not perform well when predicting measurements from the seasonal average surface (the mosaiced surface). The relationship between a selection of $PM_{2.5}$ measurements (a spatially stratified random sample, as outlined in Section 4.1), and the values extracted for those points from the seasonal average krigged surface has a low correlation (R=0.25, p<0.01, n=333).

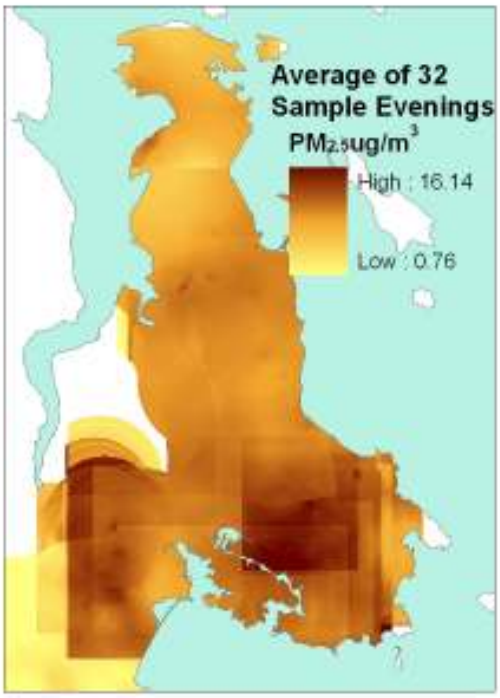


Figure 34. Average PM2.5 concentrations of mosaiced krigged surfaces, the average of the 32 routes

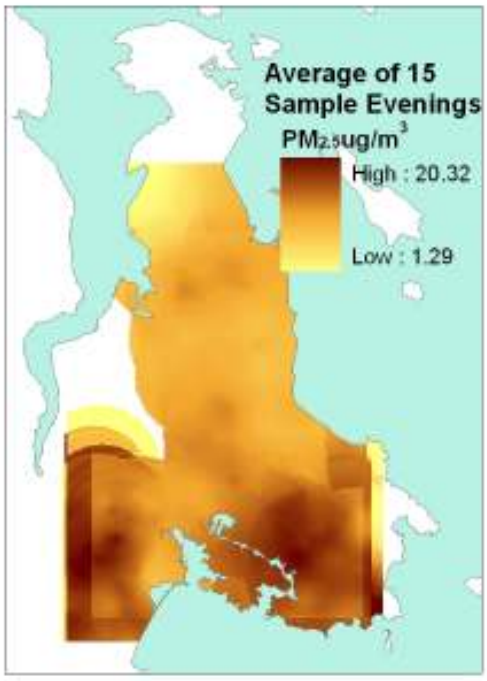


Figure 35. Average $PM_{2.5}$ concentrations of mosaiced krigged surfaces, the average of 15 evenings

Another weakness of this approach is apparent when examining Figure 34: the average surface is calculated based on the average of overlapping cells; therefore, areas that were sampled infrequently are biased towards the evening's values they were measured on. So, for example, the Metchosin area (the southwest corner of Figure 34), shows low $PM_{2.5}$ values; however, this area was only sampled twice, compared to sampling more than 15 times in other areas, therefore, the low $PM_{2.5}$ values may not be representative of that area.

4.3 Land Use Regression Modelling

This section moves from measurement to modelling to determine if:

- The spatial pattern of $PM_{2.5}$ attributable to woodsmoke from residential heating can be explained through the inclusion of independent (predictor) variables;
- $PM_{2.5}$ measurements can be enhanced to improve exposure estimates;
- A model can be developed to predict and construct past exposure; and
- A model can be developed that is transferable to times and places where little or no measurements of $PM_{2.5}$ exist?

There is interest in employing the Larson Model both from an air quality management perspective and an epidemiological and health risk assessment perspective. Before it is employed, the model requires validation and possibly calibration. Prior to examining the Larson Model, a summary of land use regression analysis is provided as a background for the remainder of the models discussed in this chapter.

LUR applies OLS regression to make predictions based on independent (or predictive) variables such as land use. The model for OLS regression is:

$$X = \beta_0 + \beta_1 Y_1 + \varepsilon \quad (5)$$

Where X is the dependent variable, β_0 is the intercept, β_1 is the regression coefficient representing the amount of change in the dependent variable for a one unit change in the independent variable Y_1, and ε is the residual or error term (Hair et al. 1984).

The first step in developing a land use regression model is to establish a monitoring network to measure the dependent variable (described in Section 3.2). Independent variables are then derived from GIS layers (Section 3.5). The model variables are chosen by examining individual correlations (Table 11) and dropping collinear or insignificant variables. Regression models were developed in Splus and the resulting model coefficients were applied to independent variable GIS layers using Spatial Analyst's 'Raster Calculator' to develop a map of predicted air pollution levels. In Figure 36, the rasters on the left represent independent variable GIS layers. These are multiplied by the model coefficients derived from regression modelling in Splus. The resulting rasters are combined using arithmetic grid operations to produce a surface of predicted woodsmoke concentrations (the far right hand grid).

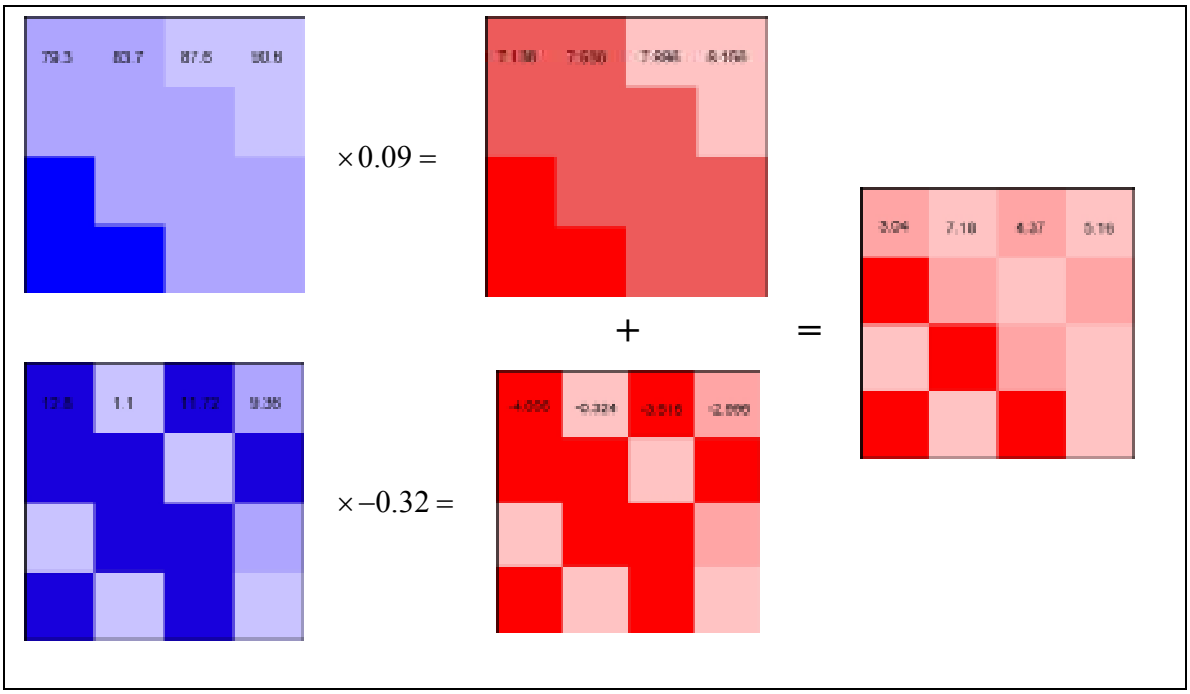


Figure 36. Creating a predicted surface of $PM_{2.5}$ attributable to woodsmoke using land use regression

Mann (1987) lists the assumptions inherent in linear regression modelling:

1. The dependent (X) and independent (Y) variables are measured without error
2. X and Y are normally distributed and independent

3. A linear relationship exists between X and Y;
4. The error term (ε) is normally distributed with a mean of zero; and
5. ε are not autocorrelated.

The first assumption is not possible to meet; however, it is not crucial (Mann 1987). The second assumption was tested by examining variable distributions. These were often lognormally distributed thereby violating the second assumption. The effect of this violation was investigated by transforming the variables to be more normally distributed and it was found to have little effect in the final model (Section 4.4.2.4 and 4.4.2.8). As for independence, any collinear variables were dropped to meet this assumption (with the exception of the Larson Model). A linear relationship was investigated by examining scatter plots of X and Y suggesting this assumption is not met (Appendix A). Assumption 4 was tested by examining residual plots to determine if there is a narrow band of randomly placed standardized residual values around 0 for measured and predicted values. The sixth assumption was tested using the serial autocorrelation function and spatial statistical software for tests of spatial autocorrelation, the residuals were then mapped in ArcMap. With this as the backdrop, the Larson Model is now examined.

4.3.1 Larson et al. (2007) Catchment Basin Model

Validating a model produces an estimate of how well the model predicts values at unmeasured locations (Johnston et al. 2003). To evaluate the Larson Model, data were divided into two subsets: one subset to create the model (the training subset) and the other subset to assess model performance (the test subset).

The CRD model was constructed using data from the 2004/2005 heating season (the training subset) and was validated using data from the subsequent heating season (2005/2006, the test subset). The evaluation consisted of the following steps:

1. Running model diagnostics to test linear regression assumptions;
2. Seasonally adjusting the test subset using the method in Larson et al. (2007) to render the data comparable to the training subset;
3. Calculating predicted light scatter using the Larson model;
4. Averaging test subset data to the catchment level to obtain observed values;
5. Performing Pearson's *R* between the training and test subset to determine performance of the model.

The Larson model is as follows:

$$ADJb_{sp} = 2.00E\text{-}005 + 9.76E\text{-}005[perlpop4] - 0.038[totimm4] + 0.024[av\_fp4] \quad (6)$$

Where $ADJb_{sp}$ is a seasonally adjusted light scatter value, *perlpop4* is the percentage of low income population, *totimm4* is the total number of immigrants and *av_fp4* is the average number of fireplaces. The number 4 at the end of each variable refers to the catchment buffer area. The catchment buffer area includes those upstream catchments where the centroid is within 4 km of the catchment centroid of interest (see Figures 7 and 8 Chapter 2).

The coefficient of determination ($R^2 = 0.73$, $p<0.00$) for the Larson model shows the model predicts the variation in light scatter well. The model is significant ($F = 13.705$ and $p < 0.00$, Table 14) and all independent variables are significant at $\alpha= 0.05$. Model diagnostics indicate collinearity between variables. The first indicator of collinearity is the *Totimm4* variable changing signs from a positive correlation with light scatter (R=0.53) to a negative β coefficient (-0.038) when it is entered into the model. According to Flaherty (2007) this is an indication of collinearity, artificially inflating the R^2 associated with model performance. In addition, collinearity diagnostics such as the tolerance and the Variance Inflation Factor (VIF) indicate collinearity. A tolerance of 1 signifies no collinearity and values approaching zero indicate collinearity. An accepted guideline is the tolerance must be greater than 0.1 (Flaherty 2007). *Av_fp4* and *Totimm4* have tolerance values that do not meet this guideline (Table 14).

Table 14. Larson regression model coefficients and collinearity statistics.

Model Variables	Coefficients		t	Sig.	Collinearity Statistics	
	B	Std. Error			Tolerance	VIF
Intercept	2.00E-005	.000	11.285	.000		
perlpop4	9.76E-005	.000	4.603	.000	.203	4.919
totimm4	-.038	.011	-3.277	.005	.064	15.599
av_fp4	.024	.010	2.550	.022	.096	10.385

Table 15 shows Pearson's correlation's (*R*) for the model variables, indicating strong associations between independent variables. For example, *perlpop4* shows associations of *R*=0.89 and *R*=0.83 (α=0.01) with *totimm4* and *av_fp4* respectively.

Table 15. Pearson's Correlation for Larson model variables

	ADJbsp	perlpop4	totimm4	av_fp4
ADJbsp	1.000	.735	.529	.588
perlpop4	.735	1.000	.891	.831
totimm4	.529	.891	1.000	.950
av_fp4	.588	.831	.950	1.000

The current model includes *totimm4*, the number of immigrants, an absolute value, when most land use regression models are based on density measures (Briggs et al. 1997; Brauer et al. 2003). The number of immigrants is misleading and potentially controversial to include because it is likely a surrogate for total population: as the total population increases, so can the number of immigrants. While immigrants can settle in particular areas, Figure 37 and Table 15 show that as population increases, the number of immigrants increase in the CRD (the association is ecological). The scatter plot (Figure 37) and Pearson's Correlation (Table 16), reveal a strong association between the population and immigrant data used to build the model (*R*=0.99, α = 0.01) and population is actually better correlated with light scatter.

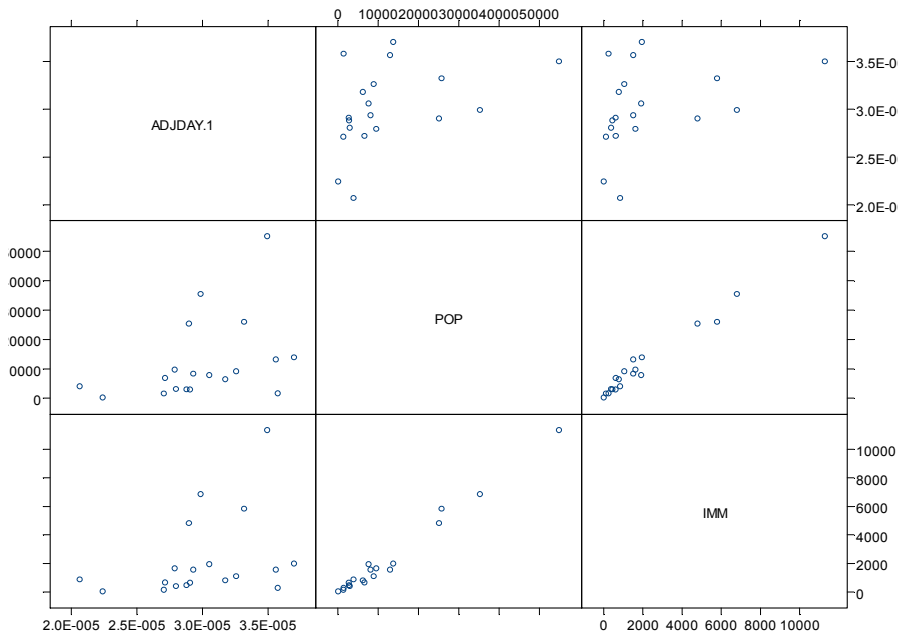


Figure 37. Matrix scatter plot for light scatter (ADJDAY.1), Population (POP) and Total Immigrants (IMM)

Table 16. Pearson's correlation matrix for light scatter, population and total immigrants.

	Light scatter	Population	Total Immigrants
Light scatter	1.00	0.40	0.35
Population	0.40	1.00	0.99
Total Immigrants	0.35	0.99	1.00

To evaluate the Larson Model using the test subset, test data required the seasonal adjustment applied to the training subset. The adjustment for seasonal variation involves adjusting each measurement to the sample evening average and then again to the $PM_{2.5}$ average of all the season's sample evenings. There is no seasonal variation in $PM_{2.5}$ rendering the adjustment unnecessary (Section 3.3, Figure 16).

Figure 38 shows the catchment boundaries. The points outside the catchment boundaries are nephelometer measurements from the test subset outside of the predicted area and were not included in the test subset.

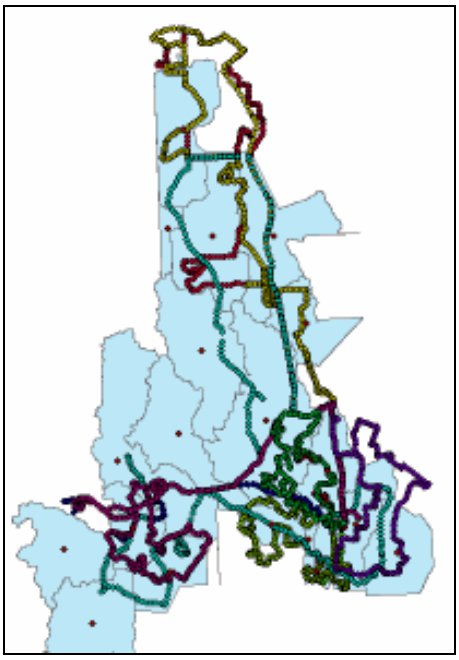


Figure 38. Larson model catchment basins (and centroids) and nephelometer mesaurements from the 2005/2006 heating season

The predicted values were calculated using the model (equation 6). The seasonally adjusted measurements from the test subset were averaged for each catchment basin as per Larson et al. (2007) using the spatial join in ArcMap and averaging the $PM_{2.5}$ values from the test points within each catchment polygon.

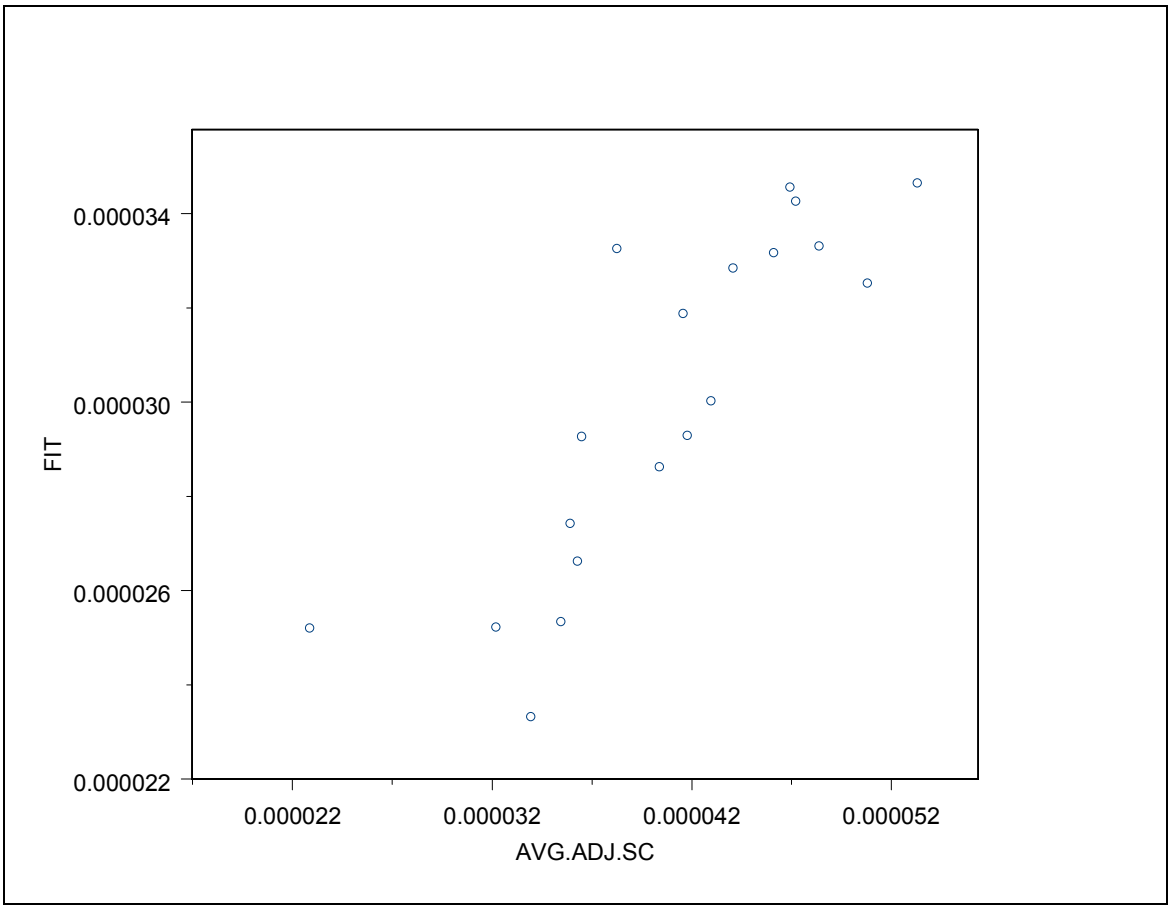


Figure 39. Scatter plot of predicted seasonal average light scatter (fit) and observed seasonal average light scatter (AV.ADJ.SC) for each catchment basin

The correlation between the predicted values (fit) and the test values is high (Figure 39, $R = 0.84$, $\alpha=0.01$, $n=19$). Figure 40 shows the Larson Model predicted $PM_{2.5}$ values. Figure 41 is a map of observed seasonally adjusted values. The range of values for each map differs: both maps are in quartiles, however the Larson Model predicts a range of 4 $\mu g/m^3$ and the observed values span 10 $\mu g/m^3$. While the Larson Model looks to predict the spatial distribution reasonably well, it does not predict a wide range of values.

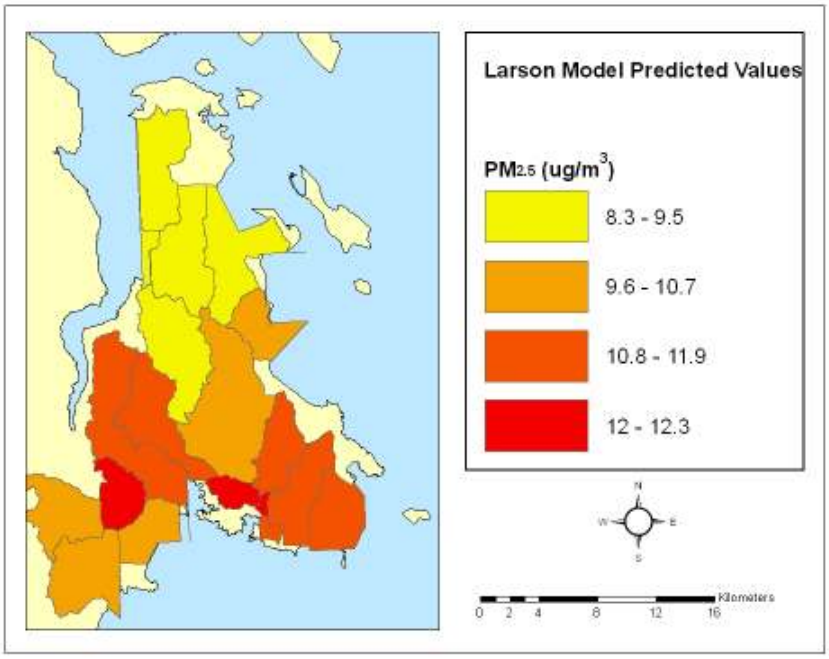


Figure 40. Larson model predicted seasonal $PM_{2.5}$ values

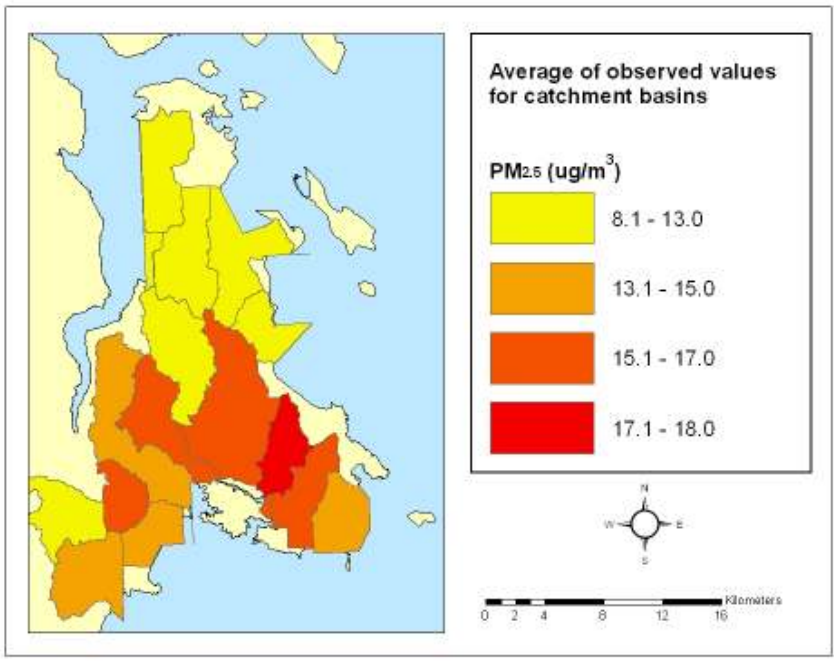


Figure 41. Observed seasonally adjusted $PM_{2.5}$ seasonal value for each catchment from 2005/2006 heating season

Due to the collinearity in the Larson Model, an alternate model that meets regression assumptions as previously described is presented as part of this thesis:

$$ADJb_{sp} = 0.00 + 0.0001\ [perlpop4] \quad (7)$$

Model performance drops (R^2=0.54, p<0.00, Table 17); however the correlation between the observed and predicted values is equal to the Larson Model (R = 0.83, α=0.01). The Low Income Model has the same limitation of predicting a small range of values. The low income model, together with a population mask covering areas of low population density, is depicted in Figure 42.

Table 17. Regression model results for Low Income Model.

Model Parameter	Value	Std. Error	t value	Pr(>|t|)
Intercept	0.0000	0.00	15.71	0.00
Perlpop4	0.0001	0.00	4.47	0.00
R-squared	0.54			

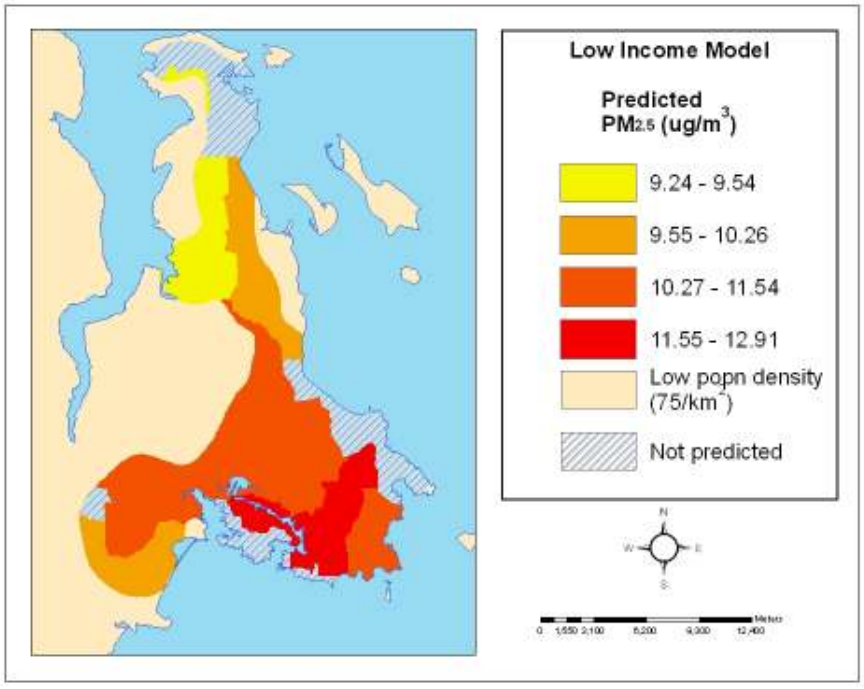


Figure 42. Low income regression model predicted seasonal average $PM_{2.5}$ attributable to woodsmoke (shown with a population density mask)

The Low Income Model is shown with a population density mask to highlight a limitation with the catchment basin approach: every person within a catchment basin is assigned the same exposure, regardless of their elevation within the catchment basin. This is the ecological fallacy and subjects exposure assessment based on this model to exposure misclassification. It is also counter to the theory behind the model that air pollution flows downward during cool, clear evenings.

The assumption that exposure to woodsmoke is only a concern during stable conditions is negated for two reasons. The first is because of the difficulty in characterizing an entire evening as such. During the monitoring campaign, the CRD experienced several different micro climates: while one area was experiencing stable conditions another was experiencing unstable conditions. Temperature remained relatively constant throughout the CRD during an evening; however, wind speed and precipitation varied. This is demonstrated in Figure 43 where both evenings experienced the same average wind speeds and the same average $PM_{2.5}$ concentrations; however, the spatial distribution differed. On December 28th (the map on the left), wind speeds were high and PM concentrations were low in the South of the CRD. Under the Larson Model assumptions, the model does not apply to this evening because of the unstable conditions in the south.[19] Yet the northwest shows low wind speeds and high PM concentrations. The opposite conditions are seen on December 26th (the map on the right) where wind speeds were low and PM concentrations were high in some areas of the south; however the north shows high winds and low concentrations. For this reason, it is difficult to characterize an evening as cool, calm and clear since these conditions vary throughout the study region.

The second reason to negate the assumption that woodsmoke is only an issue during stable conditions comes from levoglucosan measurements during unstable conditions. Figures 24 and 25 in Section 3.2 show levoglucosan is present during unstable conditions.

[19] The decision to sample was made based on observed conditions in the downtown area (southeast CRD)

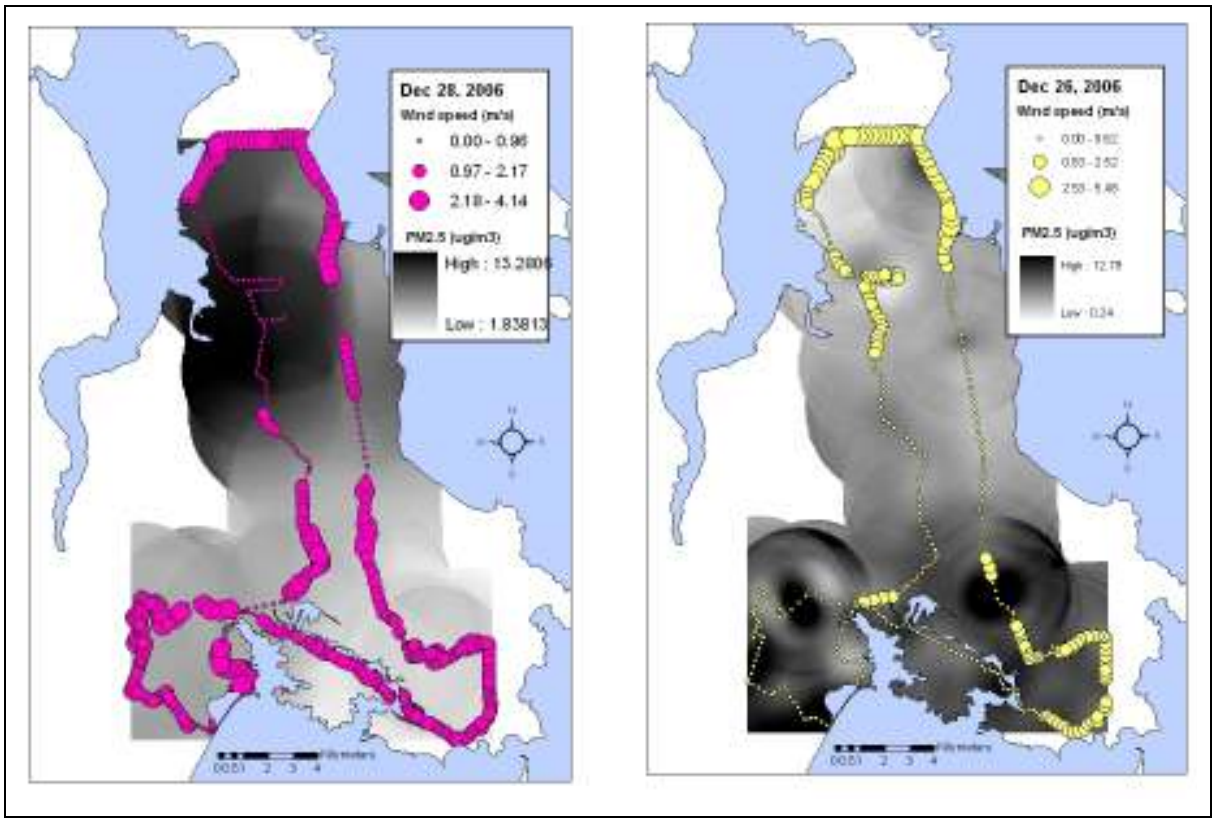


Figure 43. Wind speed and $PM_{2.5}$ concentrations for two different sample evenings.

In summary, the model predicts a small range of values and consistently under predicts observed values. The model usually predicts the spatial pattern of woodsmoke reasonably well (R = 0.84, α=0.01, n=19), but is limited for predicting the range of levels. The Larson Model does not predict unadjusted values averaged to the basin level (R = 0.48, α=0.01), therefore, this adjustment is crucial to model performance. The lack of seasonal trend observed in the data during any of the winter heating seasons brings measured values closer to the mean consequently reducing existing variation. This explains the small range of values predicted by the model. There is potential for exposure misclassification to occur when assigning exposure to an entire catchment, as the population living higher in the basin may not experience the same exposure as those lower in the basin. In addition, this model applies to calm, cool clear evenings (an invalid assumption) and therefore predicts a seasonal average consisting of those conditions which may not be an appropriate assumption.

4.3.2 A New Modelling Approach

The Larson Model is an ecological approach to modelling exposure subject to exposure misclassification and ecological fallacy. The potential to improve the spatial resolution of the model exists with the high resolution woodsmoke, SPAD and meteorological data from the VWN as sources of independent variables, thereby reducing the problems associated with an ecological model.

The Larson Model assumes exposure is negligible on meteorologically unstable evenings. To address this assumption, measurements were taken on unstable evenings during the 2005/2006 and 2006/2007 winter heating seasons to support a model applying to a variety of weather conditions. Temperature, wind speed and stability conditions were recorded as attributes for each measurement. Including meteorological data removes the need for seasonal adjustments because stability conditions associated with each measurement are available for modelling. This introduces flexibility to the model because it is no longer limited to predicting a seasonal average.

To retain the spatial resolution of the data, different approaches for dealing with spatial and temporal dependence were investigated. There are two approaches in the literature. The first is Luc Anselin's (2005) approach of modelling spatial dependence post regression (Section 4.3.2.6). The second approach suggested in Jerrett and Finkelstein (2005) is to remove spatial and temporal dependence prior to modelling (Section 4.4.2.7). Table 18 summarizes methods investigated, the results of which are discussed below. Models 1 through 5 use the first approach, Models 6 and 7 use the latter approach.

For comparison purposes, the models in Table 18 were developed using the following formula:

$$PM_{2.5} = Intercept - \beta_1[ELEV] + \beta_2\,[TEOM2] - \beta_3[TVD25P] \qquad (8)$$

Where *ELEV* is the GPS elevation in metres, *TEOM2* is the average $PM_{2.5}$ concentration in $\mu g/m^3$ of the 3 fixed monitors and *TVD25P* is the density of houses with a total value in the 25^{th} percentile. A discussion of how these were chosen is provided in Section 3.5.8.

Table 18. Different approaches to regression modelling of $PM_{2.5}$ attributable to woodsmoke from residential wood burning.

Model sampling option	Number of Observations	Model Performance	Number of resamples	Description
M1	449	R^2=0.32	1001	Stratified random sample from 15 sample evenings (one point from each 500X500m grid cell)
M2	6,625	R^2=0.32	0	All data points from 15 sample evenings (minus 10% for model validation)
M3	12,906	N/A	N/A	All data points from 32 sample evenings (minus 10% for model validation).
M4	687	R^2=0.31	0	Random selection of 10% of points from subset (15 sample evenings)
M4T	687	R^2=0.34	0	Transformed variables using square root to be more normally distributed.
M5_100	2017	R^2=0.34	0	Take average of all data points within 100m grid cell.
M5_500	449	R^2=0.40	0	Take average of all data points within 500m grid cell.
M5_1000	179	R^2=0.42	0	Take average of all data points within 1000m grid cell.
M1s	339	Log likelihood improves	0	Spatial error model: Maximum Likelihood Estimation in GeoDa.
M2s, M4s, M5_100s	3315, 6625, 2017	N/A	0	Modelling spatial error term. M6_100s shows spatial lag and spatial errors that cannot be rectified. M2 and M5 have too many points to build a sufficient weight matrix to model spatial error term correctly.
M6	596	R^2=0.33	1001	Removing spatial and temporal dependence prior to fitting model (random selection of points from each 2.5 km grid from each evening). PM2.5 shows most normal distribution of all models and residuals show no spatial dependence and are normally distributed.
M7	596	R^2=0.33		Bayesian using M6 data points (spatial and temporal dependence removed prior to fitting model).
M8**		R^2 range from 0.02 to 0.57	32	OLS model for each evening.

* These results are for the linear model: PM2.5 ~ ELEV + TEOM2. + TVD25P (except M8).
** Cannot use TEOM as a variable since it is constant for each evening.

4.3.2.1 Model 1 (M1)

This approach attempted to remove temporal autocorrelation in the woodsmoke data prior to regression modelling (spatial dependence is dealt with post modelling and results are summarized in Section 4.3.2.6). Data from 15 sample evenings were appended in ArcMap. A 500x500m grid, generated using Hawth's Tools, overlaid the points (Figure 42). Using 'Hawth's Tools's' 'Random Sample Within Subsets,' one point was randomly selected from each 500m cell to reduce autocorrelation because adjacent points are likely from different evenings. The randomly selected data set is shown in the right hand panel of Figure 44. This method had the unintended effect of increasing serial autocorrelation (Figure 45), likely because there is an abundance of noise when appending the 15 sample evenings together, and as a result, the serial autocorrelation is lost (left hand graph in Figure 45).

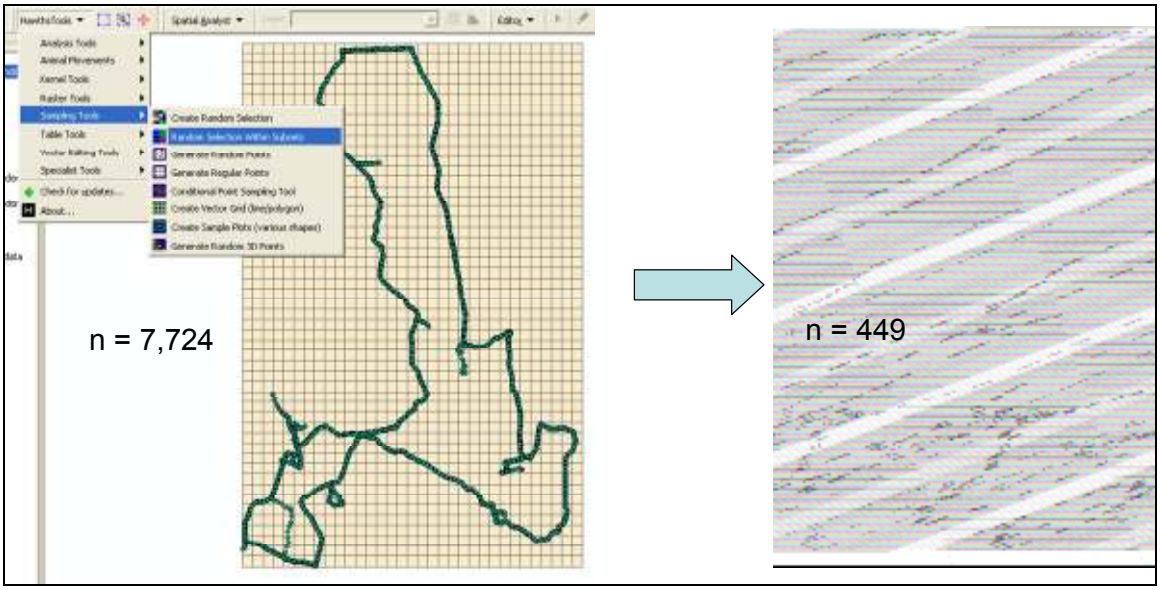


Figure 44. Random selection of one point from each 500 x 500m cell.

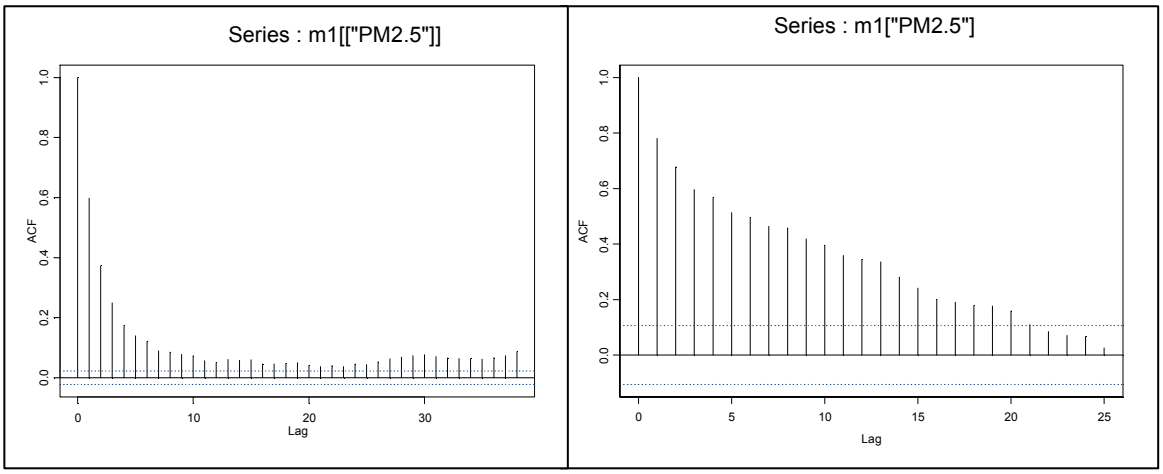


Figure 45. Autocorrelation Function (ACF) for M1 data set prior (left hand graph) and post (right hand graph) stratified random sampling

The independent variable values were extracted for each point (n=449) based on the methodology outlined in Section 3.4. The database file was imported to Splus and linear regression modelling was performed. The model is summarized in Table 19. The original model appeared promising with an R^2 of 0.47, p<0.00. Regression diagnostics show no collinear variables (multicollinearity condition index is 5.06) [20]. Figure 46 shows the normal QQ plot for the model residuals with residuals deviating from normal, mostly at the higher end.

Diagnostics for spatial dependence in residuals show significant but low global spatial correlation (Moran's I = 0.09 α<0.01, using 2500m search radius) which could explain the deviation from normal for the residuals (Figure 46). Results from the Local Moran's I cluster analysis show that the cluster of spatially autocorrelated residuals in the south-western area of the CRD deviate significantly from the overall trend of low spatial autocorrelation. Mapping the residuals can indicate other independent variables to include. A distance to the ocean variable was included based on this Moran's I analysis to little affect on model performance.

[20] A multicollinearity index value greater than 15 is problematic and a value greater than 30 is unacceptable (Anselin 2005).

Table 19. Model results for M1

Model Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	2.55	0.51	5.02	0.00
ELEV	-0.04	0.01	-4.86	0.00
TEOM2	0.49	0.03	13.94	0.00
TVD25P	0.02	0.00	7.36	0.00

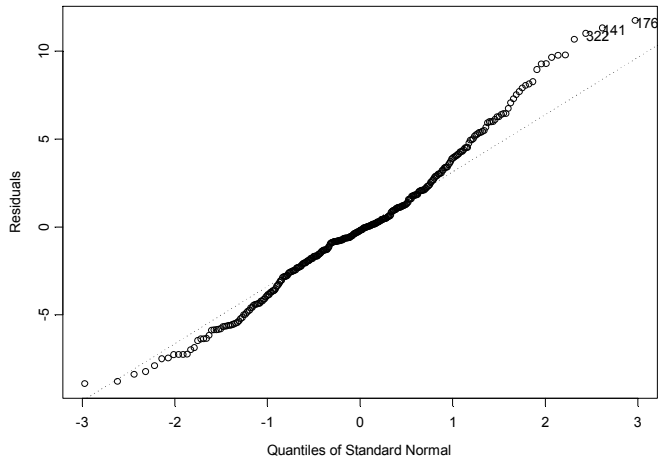


Figure 46. Normal QQ plot for M1 residuals

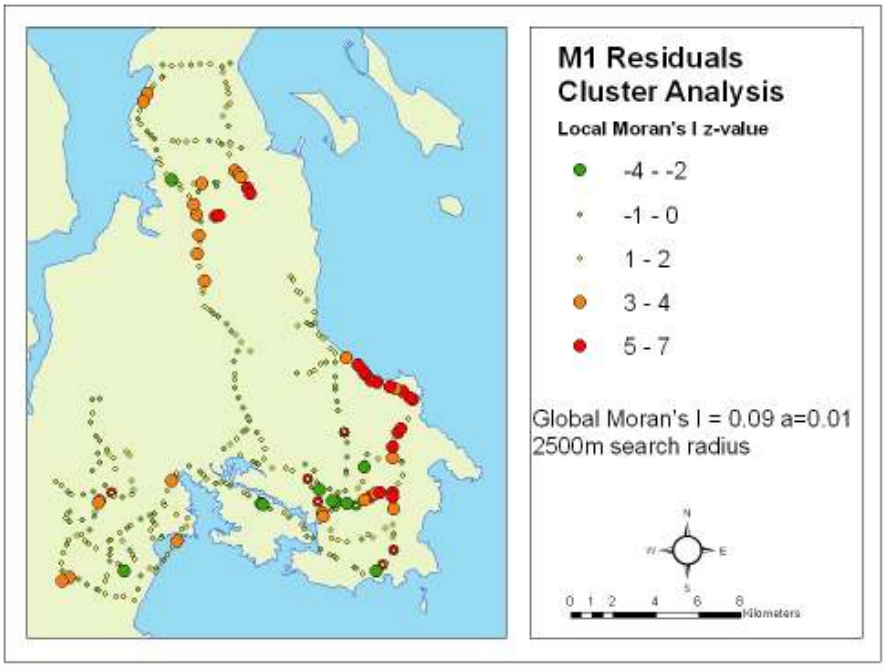


Figure 47. M1 residuals cluster analysis using Local Moran's *I*
(Global Moran's *I* = 0.09, p<0.01)

A modified bootstrap analysis was conducted to investigate the validity of M1 results. In Splus, using the script in Appendix B, the original data set (n=7,724) was randomly sampled 1000 times. One point was randomly selected from each 500 x 500m grid cell to create a data set where n=449. The model '*PM2.5~ELEV+TEOM2+TVD25P*' was calculated for each random sample i. Model coefficients and R^2 were recorded for each random sample. The bootstrap summary statistics are shown in Table 20. The intercept and *TEOM2* variables are slightly different from those in Table 19, but the most significant difference is the model performance: an average R^2 of 0.33 indicating that the model in Table 19 is overfitted to the specific points used to build the model.

Table 20. Summary of bootstrap results for M1

Model Parameter	Observed	Mean	Bias	SE
Intercept	7.02	6.86	-0.16	0.61
ELEV	-0.05	-0.05	0.00	0.01
TEOM2	0.66	0.66	0.00	0.05
TVD25P	0.02	0.02	0.00	0.00
R^2	0.32	0.33	0.01	0.04

This is not a conventional bootstrap procedure. Most bootstrap results resample from the data used to build the model. For example, for n=449, a traditional bootstrap drops the ith observation and replaces it with another observation from n=448. Using the traditional bootstrapping procedure, the R^2 of 0.47 for M1 is significant with a small standard error. Figure 48 shows the distribution of R^2 using the traditional bootstrap approach. The woodsmoke data set is unique since there are an abundance of observations. The bootstrap program was rewritten to resample from the entire data set (n=7,724) reducing the average R^2 from 0.47 to 0.33. The abundance of data allows for more robust model validation than typical data sets.

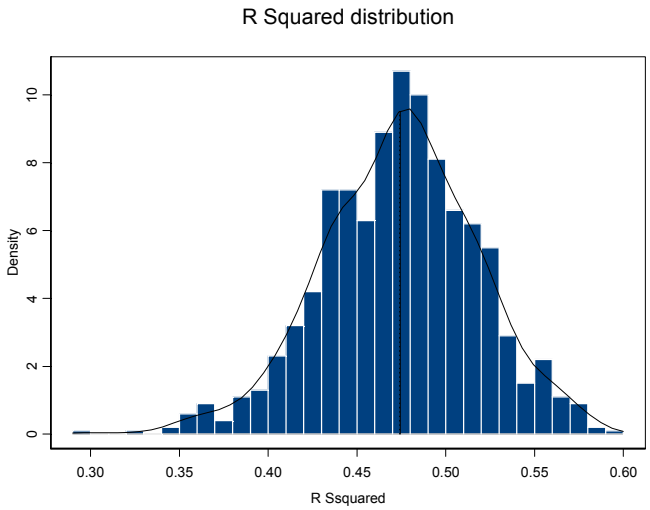


Figure 48. R^2 distribution for M1 using traditional bootstrapping procedure (resampling 1000 times from M1 data set)

4.3.2.2 Model 2 (M2)

This method retains all data points from the 15 similar data collection routes (minus 10% for model validation). This data set is shown in the left hand panel of Figure 44, with the exception that n = 6,625 due to the removal of 10% of points for model validation. M2 results are shown in Table 21:

Table 21. Model results for M2.

Model	Value	Std. Error	t value	Pr(>|t|)
Intercept	7.01	0.21	32.67	0.00
ELEV	-0.06	0.00	-13.19	0.00
TEOM2.	0.66	0.01	50.94	0.00
TVD25P	0.02	0.00	14.20	0.00
R.squared	0.32			

Retaining all of the points violates the regression assumption of normally distributed residuals. Figure 49 demonstrates the tail ends of the QQ normal plot seriously deviating from normal. Nevertheless, the results are almost identical to the bootstrap results for M1 (Table 20), indicating that the M1 method did not improve the data set for regression analysis. Data dependence for this model is dealt with in section 4.3.2.6.

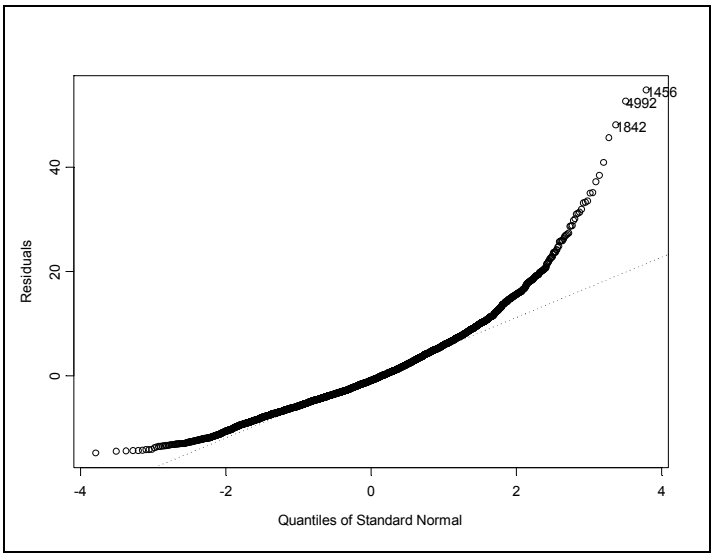


Figure 49. QQ normal plot of residuals for M2

4.3.2.3 Model 3 (M3)

Model 3 uses all data points from the 32 sample evenings (*n*=12,906); however, model development exceeded computing capacity making this approach unfeasible.

4.3.2.4 Model 4 (M4)

Model 4 is a random selection of 10% of points from each sample evening to reduce serial autocorrelation within each evening prior to appending the data sets together. Figure 50 shows the reduction in serial autocorrelation for one evening using this approach. The reduced evenings were appended in ArcMap (*n*=742). Results are similar to M1 and M2 (Table 22). The residual distribution also violates the assumption of normality, although to a lesser degree than M2 (Figure 51). Spatial dependence in model residuals is similar to M1 (Moran's $I = 0.06$, $p<0.01$, 2500m search radius).

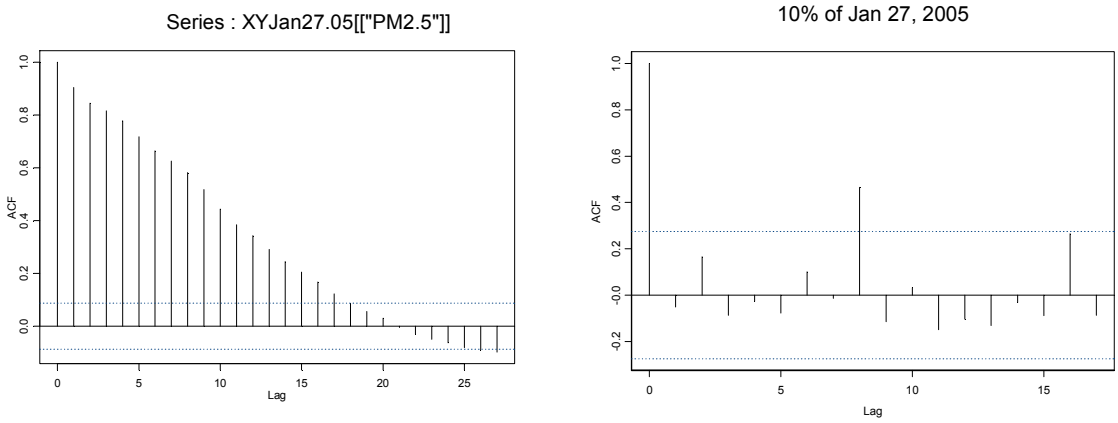


Figure 50. Serial autocorrelation function (ACF) for January 27, 2005 (graph on the left), and ACF for a random selection of 10% of points (graph on the right)

Table 22. Model results for M4.

Model	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	7.25	0.54	13.32	0.00
ELEV	-0.05	0.01	-5.09	0.00
TEOM2.	0.61	0.04	16.69	0.00
TVD25P	0.02	0.00	7.26	0.00
R squared	0.31			

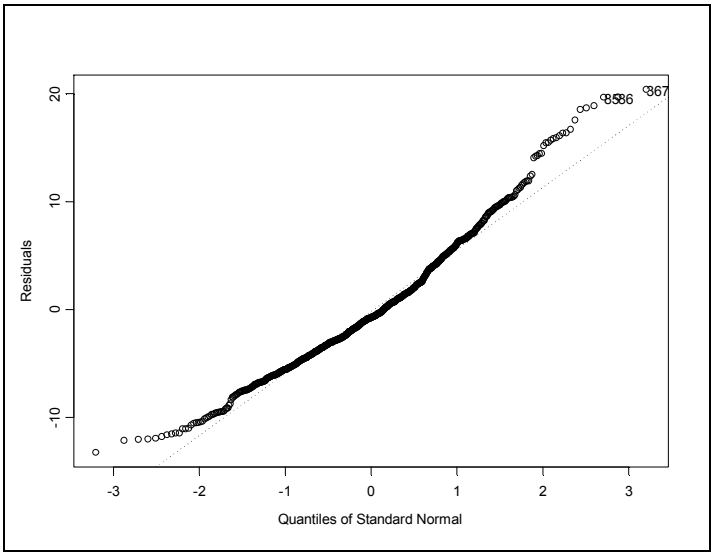


Figure 51. M4 residual QQ normal plot

Since environmental variables tend to be lognormally distributed and regression analysis assumes a normal distribution, the M4 variables were transformed to determine the effect on model diagnostics. Figure 52 shows untransformed variables on the left and transformed variables on the right. Variables were transformed using the square root function because it provided the distributions closest to normal. Table 23 shows model results using transformed variables. Only those variables that were improved by this transformation were altered, therefore ELEV remained the same. Transforming variables changes model performance (R^2 = 0.34, p<0.00), the intercept, the effect of *ELEV* and, to a lesser extent, *TEOM2* and *TVD25P*. In addition, the residual normal QQ plot is improved (Figure 53).

Since the residuals, independent and dependent variables are more normally distributed and model performance is improved; transforming variables may be required for modelling woodsmoke.

Table 23. Model results for M4 using transformed variables.

Model	Value	Std. Error	t value	Pr(>|t|)
Intercept	1.70	0.12	14.52	0.00
ELEV	-0.01	0.00	-5.32	0.00
sqrtTEOM2	0.56	0.03	17.81	0.00
sqrtTVD25P	0.05	0.01	7.66	0.00
R squared	0.34			

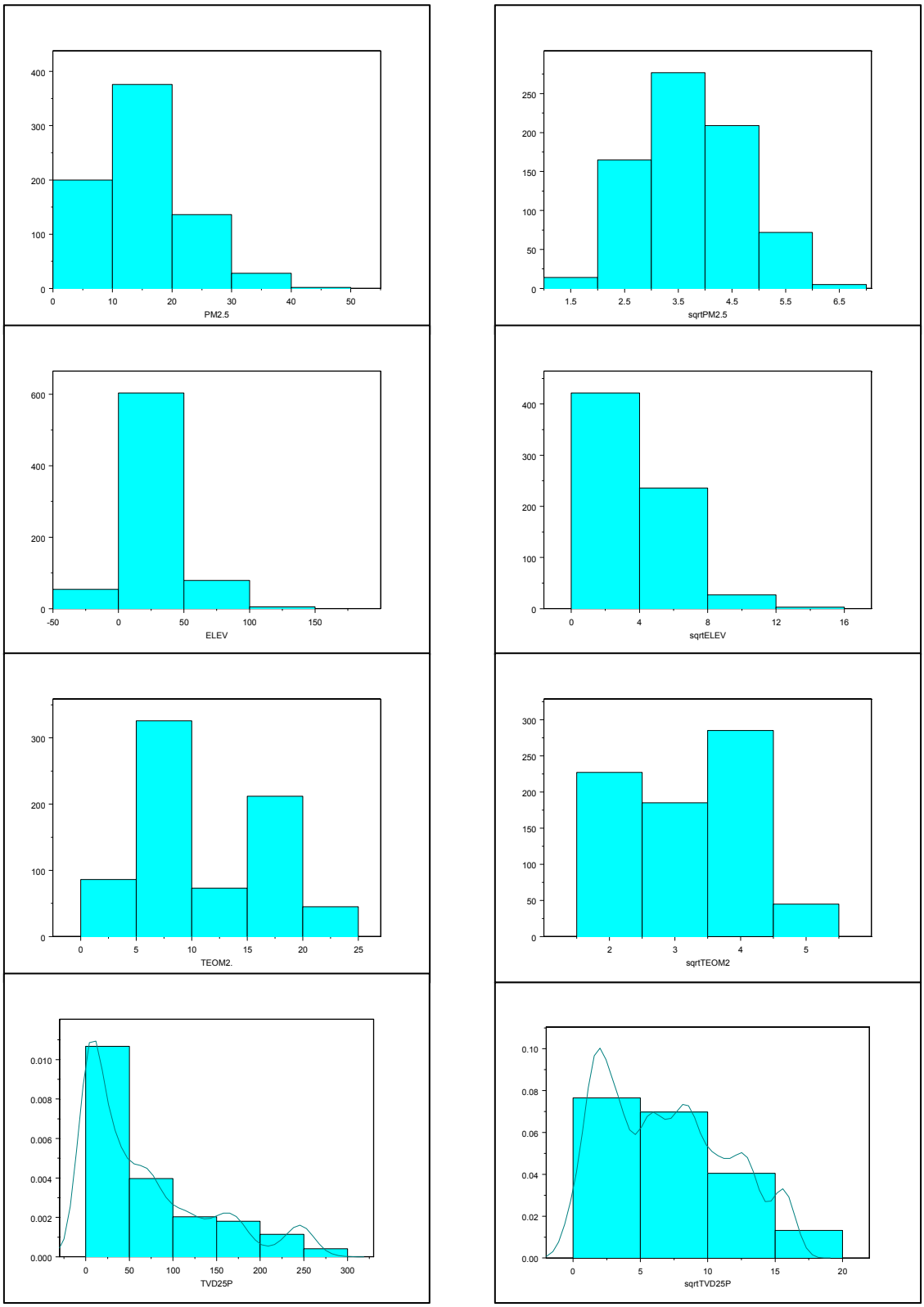


Figure 52. M4 variable distribution on the left, and transformed M4 variables using square root function variables on the right

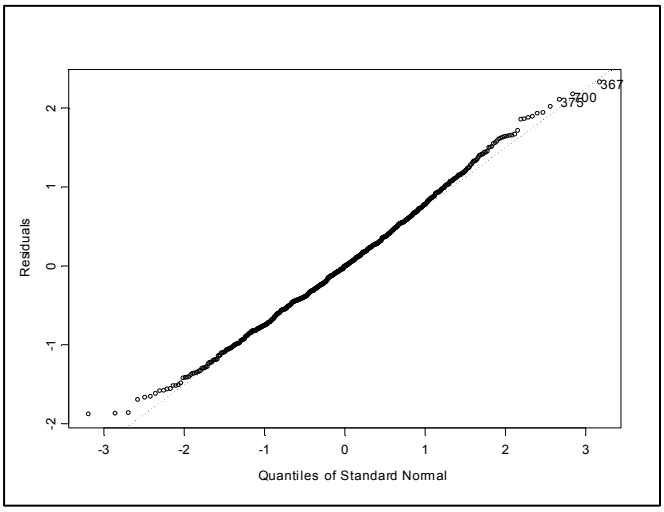


Figure 53. Residual normal QQ plot for M4 using transformed variables.

4.3.2.5 Model 5

Model 5 is a different approach to dealing with temporal dependence. The data from the 15 sample evenings were averaged to 3 grid sizes to evaluate the distribution of $PM_{2.5}$ values within each grid cell. A spatial join of the $PM_{2.5}$ point layer with the grid cell layers provided summary statistics of the $PM_{2.5}$ and independent variables intersecting each grid cell. Figure 54 shows the number of points that are summarized for each 100m, 500m and 1000m cells. If $PM_{2.5}$ values are normally distributed within each cell, then a summary statistic such as the mean or median can be used to model $PM_{2.5}$. However, as Figure 55 demonstrates, $PM_{2.5}$ values are not normally distributed within each cell making the mean or median an inappropriate measure to use for modelling.

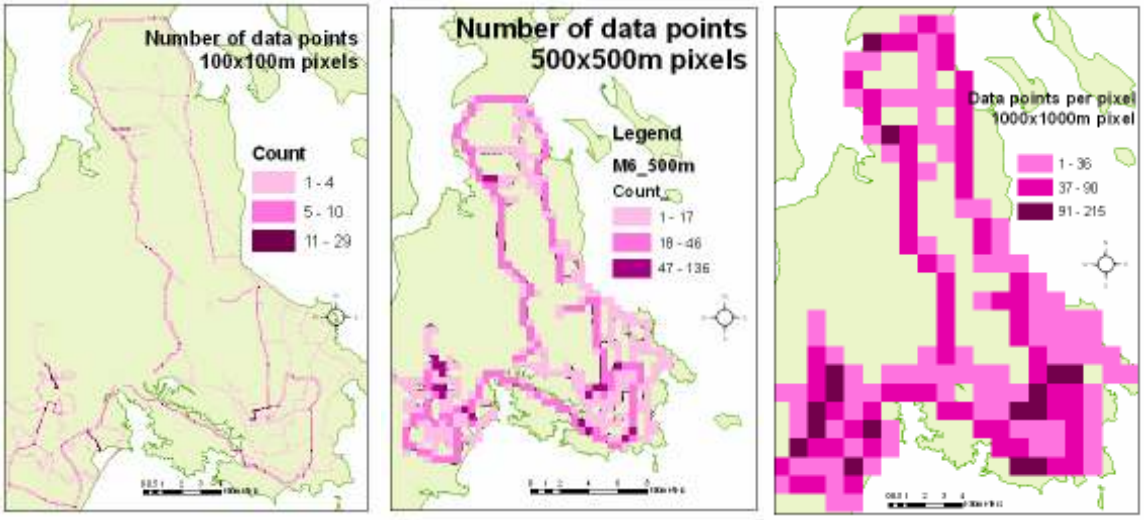


Figure 54. 100m, 500m and 1000m cells and the number of points summarized for each cell.

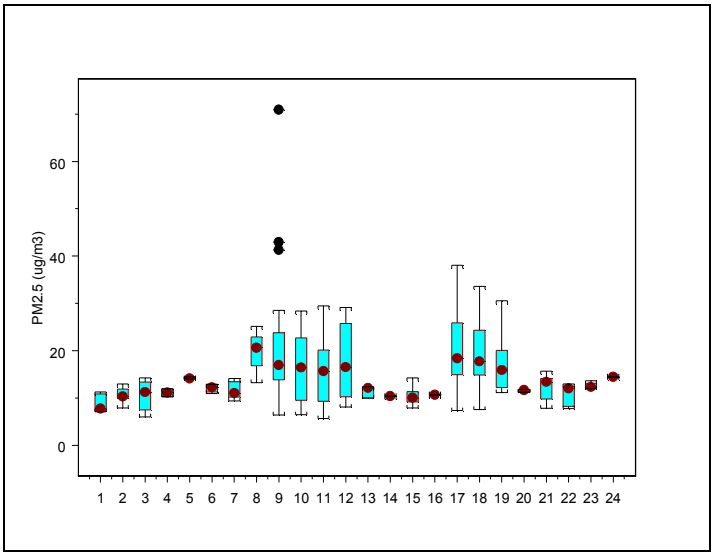


Figure 55. Box plots of PM $_{2.5}$ values for a selected number of 500 x 500m cells

The average coefficient of variation was calculated for each grid layer to determine the cell size with the lowest variation within each cell indicating the optimal cell size for modelling. The coefficient of variation statistics are summarized in Table 24. According to these statistics, the 100m cells have the least variation so averaging model parameters to this cell size was investigated. Table 25 has the model results indicating there is little difference from the models investigated thus far (transformed variables

aside). In addition, the residuals deviate from normal. Since $PM_{2.5}$ is not normally distributed within cells, model parameters remain unchanged and residuals still deviate from normal, there is little to gain from this approach.

Table 24. Summary statistics of coefficient of variation for three grid sizes.

Coefficient of Variation Summary Statistics						
Grid size (m)	Mean	Median	Standard Deviation	Minimum	Maximum	Number of grid cells
100	0.23	0.23	0.22	0	0.87	2017
500	0.3	0.37	0.21	0	0.93	449
1000	0.36	0.41	0.17	0	0.79	179

Table 25. Model results for M5 using averages for 100m cells.

Model	Value	Std. Error	t value	Pr(>|t|)
Intercept	6.21	0.35	17.73	0.00
Avg.ELEV	-0.05	0.00	-12.28	0.00
Avg.TEOM2.	0.71	0.03	26.69	0.00
Avg.TVD25P	0.02	0.00	14.71	0.00
R squared	0.34			

Models were also calculated using the 500m and 1000m averages. As table 18 indicates model performance improves with each level of aggregation (R^2=0.34 for 100m, R^2=0.40 for 500m and R^2=0.42 for 1000m). Since this method is deemed inappropriate for modelling woodsmoke, this indicates that the MAUP is an issue with woodsmoke modelling.

4.3.2.6 Spatial regression modelling

Aside from the model using transformed variables, problems exist with autocorrelation in the residuals. When data are not spatially independent, Keitt et al. (2002) have shown that model variables can change significantly when spatial autocorrelation is taken into account, making this an important aspect to consider in the spatial modelling of woodsmoke. Anselin (2005) provides free spatial statistical software called GeoDa to diagnose and model spatial autocorrelation in regression. The spatial regression process is shown in Figure 56 and was applied to models 1, 2, 4 and 5 to determine if the violations of regression assumptions could be rectified and to determine if model performance could be improved.

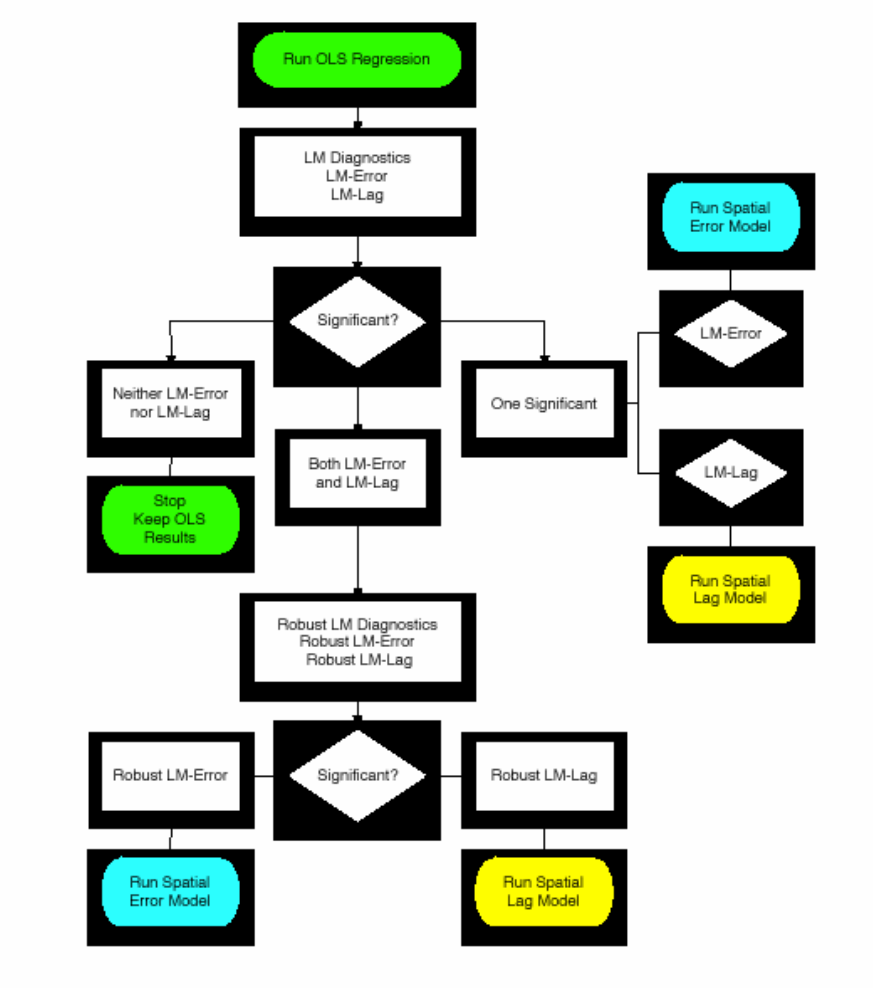


Figure 56. Spatial regression modelling process in GeoDa (from (Anselin 2005)

GeoDa provides spatial dependence diagnostics for OLS regression models: it provides diagnostics for spatial dependence within the model, using the Lagrange Multiplier (LM) lag test, and for model residuals, using the LM error test. As in Splus, the OLS model is run in GeoDa. The user then specifies distance weight matrices for the spatial dependence diagnostics. The choice of weight matrices influences the spatial dependence diagnostics; therefore, they must be chosen based on prior knowledge of the spatial process or by examining the effect of a variety of weight matrices (Boots 2002). In addition, the number of neighbours included in the calculation must be at least 8 (Nelson et al. 2005). Both methods were explored and the results discussed below pertain to M1 using a weight matrix of 2500m, a distance based on the semivariogram range (Section 3.2.2).

After the OLS regression is run, the next step is to examine the LM diagnostics (Table 26). Moran's *I* is 0.09, $p<0.00$ indicating a slight but significant spatial autocorrelation. If Moran's *I* is significant, the LM statistics must be considered (Table 26). If both lag and error statistics are insignificant, than the OLS model is the best model. LM diagnostics for M1 show significant LM error problems ($p<0.00$), therefore, the spatial error model is the model run (see Figure 56).

Table 26. Diagnostics for spatial dependence in M1 using weight matrix of 2500 m (row-standardized weights).

TEST	Moran's *I*/ Degrees of Freedom	VALUE*	PROB
Moran's *I* (error)	0.09	6.70	0.00
Lagrange Multiplier (lag)	1	1.85	0.17
Robust LM (lag)	1	9.18	0.00
Lagrange Multiplier (error)	1	33.08	0.00
Robust LM (error)	1	40.41	0.00

*All LM statistics are one-directional test statistics distributed as χ^2.

A spatial effect in the error terms suggests deviations from the global model take place at a very local scale (Fotheringham et al. 2000). Figure 47 shows the spatial effect in residuals for M1. Since the error terms are not independent, then OLS predictions of β_0 and β_1 are not justified (Fotheringham et al. 2000). The spatial error model estimates the spatial trend in error terms by maximum likelihood estimation (Anselin 2005). This model is:

$$X = \beta_0 + \beta Y + \varepsilon, \text{ with } \varepsilon = \lambda W\varepsilon + u \quad (9)$$

where X is a vector of observations on the dependent variable, W is the spatial weights matrix, Y is a matrix of observations on the independent variables, ε is a vector of spatially autocorrelated error terms, u is a vector of independent errors, and λ and β are model parameters (Anselin 2005). The resulting spatial error model for M1 is shown in Table 27.

Table 27. Spatial error model results for M1 using 2500m distance weight.

Variable	Coefficient	Standard Error	z-value	Probability
CONSTANT	2.44	0.69	3.54	0.00
ELEV	-0.05	0.01	-5.31	0.00
TEOM2_	0.53	0.03	15.24	0.00
TVD25P	0.02	0.00	4.08	0.00
LAMBDA	0.58	0.10	5.78	0.00

The R^2 is not an appropriate measure of performance for spatial regression models because it is not comparable to the OLS R^2 (Anselin 2005). The measure of fit for comparison is the Log-Liklihood (LL). The LL of the OLS model is -942.97 and the LL for the spatial error model is -932.74 indicating the spatial error model is an improvement over the OLS model because the LL closer to zero is a better model.

There are 3 classic specification tests comparing the null model (i.e., the OLS model) to the spatial regression model: the Wald test (W), the Liklihood ratio (LR) and the LM error test. The results need to occur in the following order, W>LR>LM, to reject the null model. The Wald test is the square of the λ z-value ($5.78^2 = 33$), the LR test is 20.1 and the LM error test is 33. This is not compatible with the expected order; therefore, the null model cannot be rejected. Anselin suggests two ways to improve the spatial model: changing model weights and changing model variables. A variety of weights were investigated ranging from 664m, the threshold weight where each point had at least one neighbour, to 5000m. No models met the specification order although the model using the threshold weight matrix came the closest. This leaves specifying new model variables as an option. All of the variables listed in Table 11 were explored with no improvements to model diagnostics; therefore, model misspecification may be failing to improve the model. Nonetheless, exploring other variables goes beyond the scope of this research.

Spatial regression could not be completed for M2 and M4 because modelling the spatial error term exceeded computing capacity due to the large number of points included in each model. M5_100 showed spatial lag and spatial errors that could not be rectified using a variety of weight matrices.

Although the spatial regression approach proposed by Anselin (2005) showed marginal improvements over the OLS models as indicated by the improvement in the LL,

since the null cannot be rejected suggests that misspecification exists in the model, since a variety of distance weight matrices were explored, none of which produced results that were compatible with the order of specification tests.

The next section explores the option of removing spatial and temporal dependence prior to fitting an OLS model.

4.4.2.7 Model 6 (M6)

This approach removes spatial and temporal dependence prior to running OLS regression to meet assumptions. According to Jerrett et al. (2005b), spatial modelling of exposure requires the anticipation and elimination of spatial autocorrelation prior to fitting models because no methods exist to diagnose spatial autocorrelation in residuals (cf. Anselin 2005). M6 is built using data from the 15 similar routes. A 2.5km grid was laid over measurements from a sample evening (Left hand panel in Figure 57). Using Hawth's tools, one point was randomly selected from each cell (Middle panel in Figure 57). This process was completed for the 15 sample evenings. The 15 layers were then appended in ArcMap (Right hand panel in Figure 57).

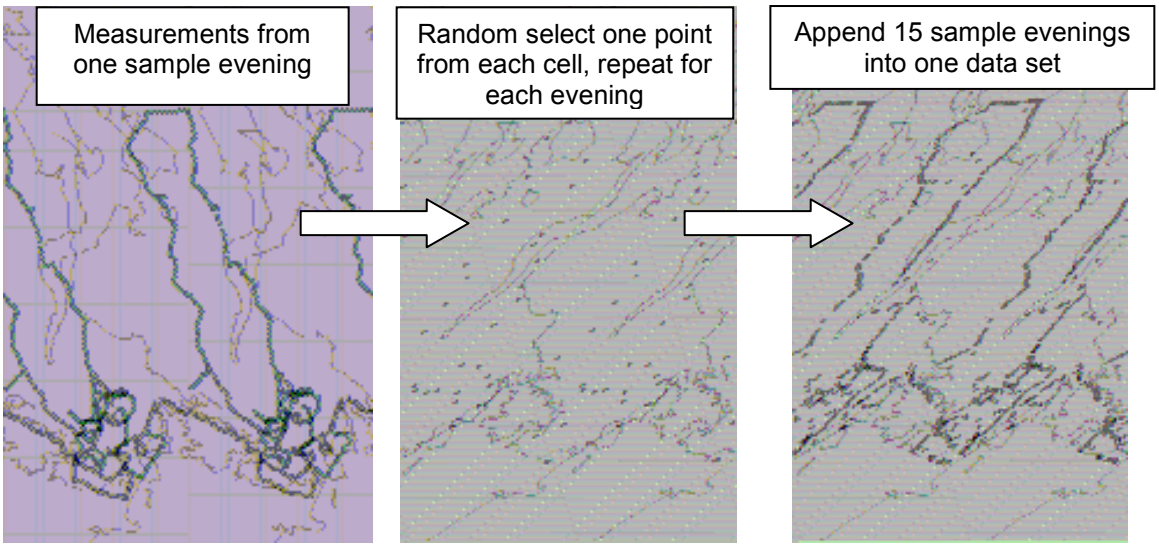


Figure 57. Creation of the M6 dataset

2.5 km was chosen because beyond this distance, spatial dependence in the woodsmoke data disappears. This also served to remove serial autocorrelation as shown by the autocorrelation function graphs in Figure 56.

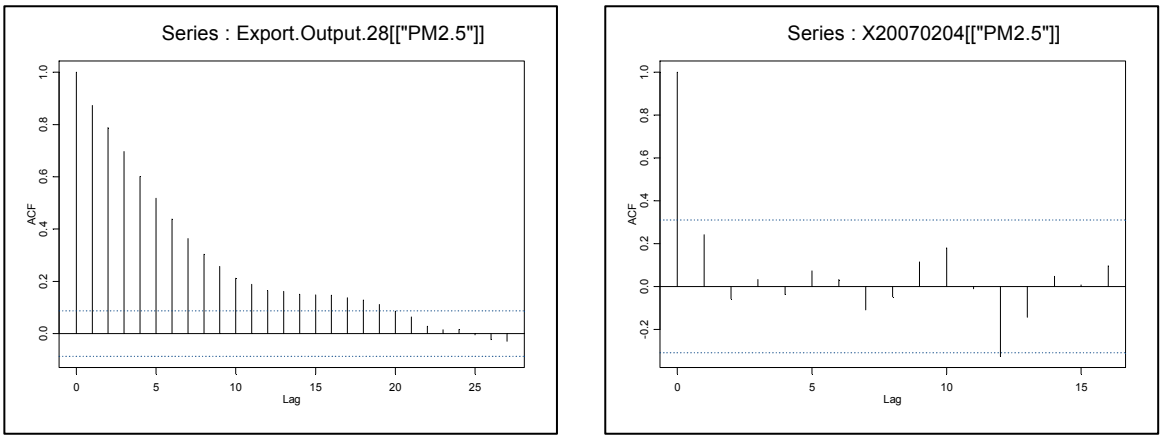


Figure 58. Serial Autocorrelation (ACF) for a sample evening on left graph, ACF after random selection of a point from each 2.5km grid square

Model results for M6 are presented in Table 28 (R^2=0.37, p<0.00). Diagnostics for spatial dependence indicate that there is no spatial autocorrelation in the model (Moran's I = 0.01, p>0.05) and diagnostics show it will not benefit from a spatial regression model.

Table 28. Regression model results for M6.

Model	Value	Std. Error	t value	Pr(>|t|)
Intercept	7.01	0.55	12.74	0.00
ELEV	-0.05	0.01	-3.48	0.00
TEOM2.	0.62	0.04	16.55	0.00
TVD25P	0.03	0.00	7.18	0.00
R squared	0.37			

A bootstrapping procedure similar to the one conducted for M1 (randomly sampling one point from each 2.5km grid cell for each sample evening, 1000 times) showed that this result is on the higher end of the performance distribution (Table 29), with this model producing similar average results to the other models. Nevertheless, the residuals are more normally distributed (Figure 59). Figure 60 shows the clusters of residual values deviating from normal. While there is no global spatial autocorrelation in the data set (Moran's I = 0.03, p<0.01, 2500m search radius), Figure 58 shows where local spatial autocorrelation in residuals is occurring. The spatial distribution of residuals show the model is over predicting in the north and under predicting $PM_{2.5}$ in the Langford/Collwood area.

Table 30 demonstrates that the model can be improved by the addition of temperature (*TEMP2*) and distance to the ocean (*NEAR.DIST*) to an average R^2 of 0.36. On the other hand, the inclusion of variables to marginally improve the model detracts from its simplicity and transferability. In addition, it does not change the spatial distribution of residuals in the model.

Table 29. Bootstrap results for M6 (1000 iterations).

Model	Observed	Mean	Bias	Standard Error
Intercept	6.93	7.10	0.16	0.33
ELEV	-0.05	-0.04	0.01	0.01
TEOM2.	0.67	0.63	-0.04	0.02
TVD25P	0.02	0.03	0.00	0.00
R squared	0.33	0.33	0.00	0.02

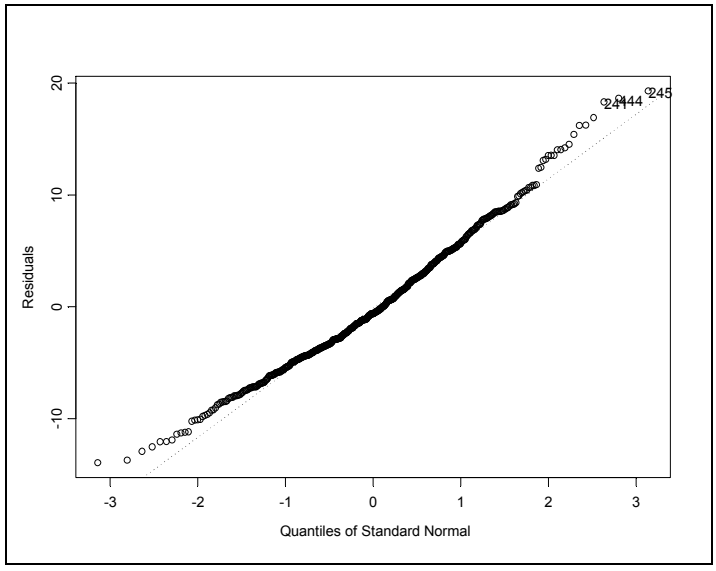


Figure 59. Normal QQ plot for M6 residuals.

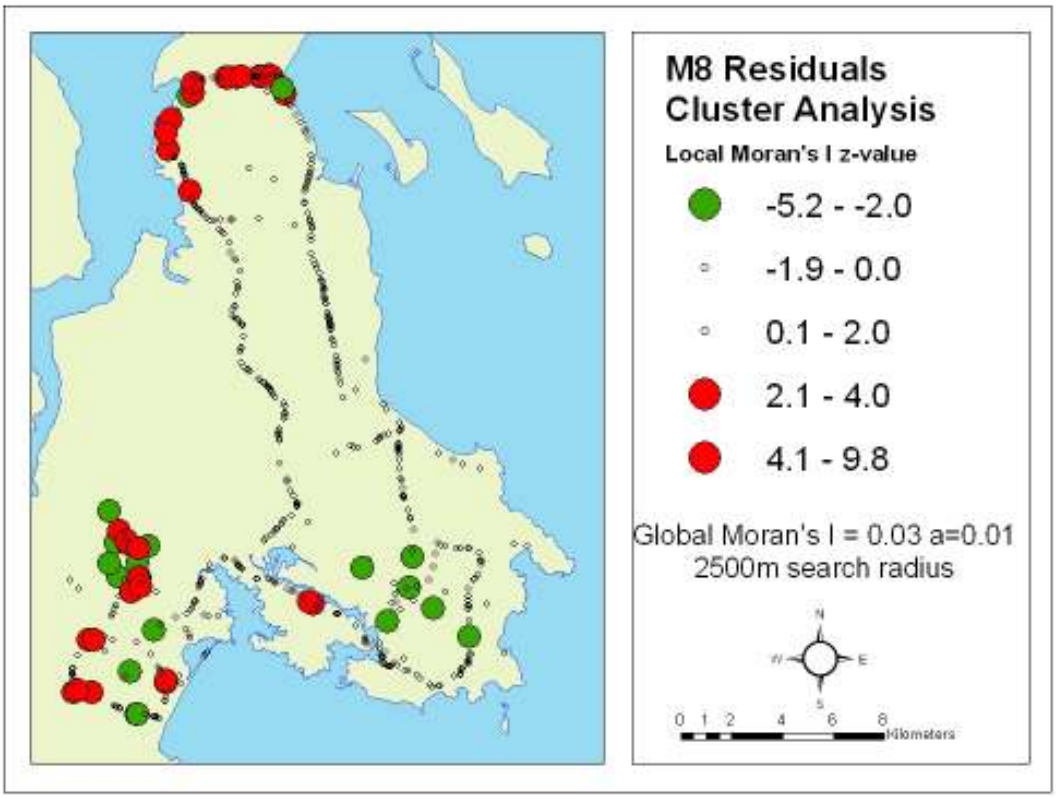


Figure 60. Cluster analysis of M6 residuals using Local Moran's *I* (Global *I*=0.03)

Table 30. Bootstrap results (1000 iterations) for M6 plus 2 more variables.

Model	Observed	Mean	Bias	Standard Error
Intercept	4.65	4.36	-0.29	0.41
TEOM2.	0.72	0.70	-0.02	0.02
ELEV	-0.07	-0.07	0.01	0.01
TVD25P	0.02	0.02	0.00	0.00
NEAR.DIST	0.0001	0.00	0.00	0.00
TEMP2	0.18	0.30	0.12	0.04
R^2	0.35	0.36	0.01	0.02

4.4.2.8 Model 7 (M7): The Bayesian Approach

Spatial regression (Section 4.4.2.5) uses maximum likelihood estimation, a method for *estimating* models, whereas the Bayesian approach simulates *exact* solutions where the accuracy depends on the computation (Gilks et al. 1996).

In Bayesian inference joint probability distributions are created for all model entities where D is the observed data (i.e. $PM_{2.5}$, elevation etc.) and θ represent model

parameters. Prior distributions $P(\theta)$ and likelihood distributions $P=(D|\ \theta)$ are set up to make up a full probability model:

$$P(D,\ \theta) = P(D|\ \theta)P(\theta) \tag{10}$$

Since D is observed, Bayes theorem is applied to determine the distribution of θ conditional on D (called the posterior distribution of θ, the object of Bayesian inference):

$$P(\theta|D) = \frac{P(D|\ \theta)P(\theta)}{\int P(D|\ \theta)P(\theta)d\ \theta} \tag{11}$$

Building the model in WinBUGS includes specifying the model variables (See Appendix C for the model specifications). First θ is defined which is a vector of regression coefficients and the model intercept. Prior distributions for θ were specified. For the woodsmoke model, there is no prior information incorporated; therefore, a non-informative prior was specified. The likelihood function was specified for the observed data describing the conditional probability of observing the data given a specific value of θ. The likelihood function reflects the relationship between D, covariables and θ. The posterior distribution is determined using Bayes' theorem and summarizes all the information contained in the observed data and prior information. It provides the probability of observing each value of θ conditional on D. The posterior distribution of θ can be expressed as the posterior expectation of functions of θ ($E[f(\theta|D)]$), or $E[f(X)]$. Evaluating $E[f(X)]$ is almost impossible and Markov Chain Monte Carlo (MCMC) simulation is a method that is typically used for evaluating $E[f(X)]$. MCMC draws samples from input distributions to approximate:

$$E[f(X)] \sim 1/n\sum f(X_t), \tag{12}$$

where n is set by the user (i.e., n=1000) and $X_{t,}$ is a randomly selected variable from the input distribution. Markov Chains sample from the input distributions where X_{t+1} does not depend on the history of the chain, X_t. Usually the first 1 to 2% of results from the first m iterations (called the burn-in iterations) are discarded giving the ergodic average (Gilks et al. 1996).

Model output was monitored in WinBUGS to determine the length of *m*. Summary statistics for θ (called nodes in Table 31) were calculated and output for inference about true values of unobserved nodes. The table specifies credibility intervals, a statement of the probability distribution of the variable of interest, as opposed to confidence intervals of the OLS approach. The model intercept differs from the OLS approach (6.93 for M6, 14.65 for Bayesian) due to the centering process applied to the Bayesian model.[21] The remaining model coefficients, are the same as the bootstrapped values of coefficients for M6 listed in Table 29 The start column in Table 31 refers to *m*, which were discarded. The results for Table 31 are based on the 14,001 samples run after *m*.

Table 31. Results for M7 using Bayesian approach.

Node	mean	sd	MC error	2.50%	median	97.50%	start	sample
Intercept	14.65	0.24	0.0078	14.18	14.66	15.1	1000	15001
beta1 (TEOM2)	0.62	0.038	9.66E-04	0.55	0.62	0.70	1000	15001
beta2 (ELEV)	-0.046	0.013	3.75E-04	-0.07	-0.046	-0.020	1000	15001
beta3 (TVD25P)	0.027	0.0036	1.08E-04	0.019	0.026	0.033	1000	15001
sigma.model	5.73	0.16	0.0043	5.44	5.73	6.06	1000	15001

To evaluate the OLS and Bayesian models, the predicted values from the OLS model (OLSfit) and the Bayesian model (B.pred) were calculated. The predicted values were compared with observed values of $PM_{2.5}$ using Pearson's Correlation (Table 32). According to Pearson's *R*, the values predicted by the Bayesian and the OLS model are exactly the same (R=1.0, α=0.01, Figure 61). Essentially, the Bayesian and OLS models produce the same model, even though their intercepts are different due to the centering of variables used to calculate the Bayesian model. The difference is in the interpretation of results. The Bayesian interprets the findings in terms of probability: there is 95% probability that the model intercept is between 14.18 and 15.1. Whereas the frequentist OLS interpretation states that if the experiment were repeated, 95% of the time, the confidence intervals will contain the true value of the intercept. Also, the frequentist

[21] The centering process refers to subtracting a constant (in this case, the mean) from the independent variables. Centering serves two purposes: the first is to improve computational efficiency, and the second is for ease of interpreting model coefficients.

approach provides a point estimate of the model parameters. This has implications for because the variance associated with the estimate is neglected with the frequentist interpretation.

Table 32. Pearson's Correlation for predicted values using the Bayesian and OLS models.

	B.pred	PM2.5	OLSfit
B.pred	1.00	0.61	1.00
PM2.5	0.61	1.00	0.61
OLSfit	1.00	0.61	1.00

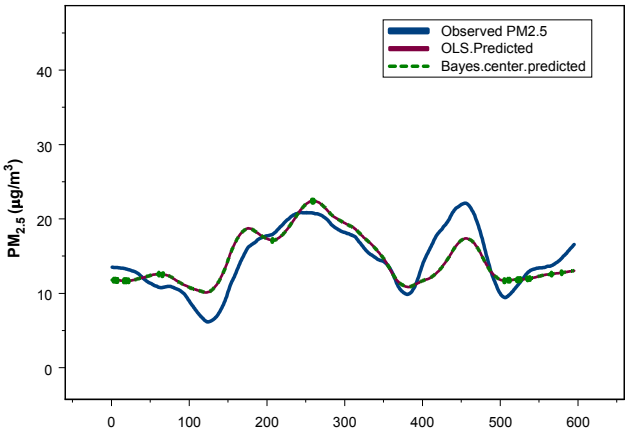


Figure 61. Comparison of Bayesian versus OLS model (n=595)[22]

4.4.2.9 Model 7

Individual models for each sample evening (no attempt was made to remove spatial or temporal dependence) were developed to determine if there is a difference in model parameters depending on the type of data collection route (a full or partial route) and the meteorological conditions. The 32 models are included in Appendix D.

There was no difference in model performance due to type of route (Table 33) or due to wind speed (Table 34). Elevation was rarely significant for smaller routes since the variation in elevation was minimal. One observed difference was that as temperature increased, model performance decreased, possibly because there is less need to burn wood for heat as temperature rises (Table 35).

[22] Graphed using Friedman's super smoothing function in Splus because it maintains the trend while removing noise.

Table 33. T-Test for difference in mean R^2 for two samples (full and partial routes)

	Partial route	*Full route*
Mean	0.26	0.30
Variance	0.02	0.03
Observations	17	15
Pooled Variance	0.025	
Hypothesized Mean Difference	0	
df	30	
t Stat	-0.75	
P(T<=t) one-tail	0.23	
t Critical one-tail	1.7	
P(T<=t) two-tail	0.46	
t Critical two-tail	2.04	

Table 34. T-test for difference in mean R^2 square for 2 samples (windy and non-windy evenings)

	Below average windspeed	*Above average windspeed*
Mean	0.28	0.27
Variance	0.02	0.03
Observations	19	12
Pooled Variance	0.026	
Hypothesized Mean Difference	0	
df	29	
t Stat	0.14	
P(T<=t) one-tail	0.45	
t Critical one-tail	1.7	
P(T<=t) two-tail	0.89	
t Critical two-tail	2.045	

Table 35. Pearson's Correlation for model performance and temperature.

	R^2	Average temperature (°C)
R^2	1	
Average temperature (°C)	-0.46	1

4.5 Comparison and Discussion of Models

This section summarizes the model findings to determine the best approach for characterizing exposure. In the literature, the best land use regression models are chosen based on the R^2 (i.e., Hoek et al. 2002; Brauer et al. 2003; Henderson and Brauer 2005;

Larson et al. 2007). This is not the only means for assessing model performance because the coefficient of determination does not consistently reflect the accuracy of prediction (Comrie 1997; Diem 2003). For example the R^2 is not an adequate assessment of the Larson Model where the R^2 was the sole driver in selecting variables yet the model exhibits limitations such as collinearity between variables. Therefore, the models are evaluated both quantitatively and qualitatively. The models are evaluated quantitatively by comparing the R^2, as well as by examining the association between the values predicted by the model and the measured values (not used in model construction) as indicated by Pearson's *R*. The models are assessed qualitatively by addressing the following questions:

- Does the model meet regression modelling assumptions?[23]
- Is the model transferable to other areas?
- What is the spatial resolution of the model?
- Is there potential for exposure misclassification?
- Is the model predictive or explanatory?

The model evaluations are summarized in Table 36.

The baseline scenario performs well in comparison with the other models and is a key variable in the LUR models developed in Section 4.4.2. Figure 33 displays concentrations from a typical evening, and demonstrates the potential for exposure misclassification: the RRU population is assigned the lowest concentration, failing to capture one of the woodsmoke hot spots occurring within the polygon. This model is transferable to areas where fixed monitoring exists but performance depends on the density of monitors. Because the model is based on PM measurements only, it does not attempt to explain, only predict, woodsmoke concentrations.

[23] Since no model can meet Mann's first assumption outlined in Section 4.3, this assumption is disregarded.

Table 36. Evaluation and comparison of exposure models.

Model	R^2	Prediction of measurements* (Pearson's *R*)	Meets Regression Model Assumptions	Transferability	Spatial Resolution	Exposure Misclassification Potential	Explanatory or Predictive
Baseline	0.25 (closest monitor) 0.33 (average of monitors)	0.51	No: Observations/ residuals not normally distributed	To areas with fixed monitoring	Coarsest. Dependent upon density of fixed monitors	High	Predictive
Kriging	n/a	0.84 (for sample evening) 0.25 (seasonal surface)	n/a	Not transferable	Fine (50m cells)	High	Predictive
Larson et al. (2007) Catchment Basin Model	0.73	0.84 (adjusted values) 0.48 (unadjusted)	No: violates assumption of independence between variables	Yes	9km^2	Moderate	Predictive
	0.54	0.83 (adjusted)	Yes				
M6 (2.5km grid) and M7 (Bayes)	0.33	0.62	Violates linear relationship assumption	Yes	Fine (50 m cells)	Low	Explanatory

* Prediction of the test subset of nephelometer measurements.

Kriging, like the baseline scenario, is based on measurements, but these measurements come from nephelometer measurements during sample evenings. Strengths of krigging include:

- It indicates pollution hot spots not captured by the baseline scenario
- The mobile monitoring method provides the ability to interpolate at a high spatial resolution
- It provides an estimate of the error at each location; and
- It is simple to implement.

For predicting measured values not used to construct the surface, kriging performs poorly (Table 36, Pearson's $R = 0.25$, $\alpha=0.01$). Kriging does not produce a transferable model because the krigged surface is dependent upon the time and place of measurements and cannot extend beyond the spatial dependence in the data set. This also creates the potential for developing an unrepresentative prediction since the krigged surface reflects a snap shot in time of a phenomenon that is continuously changing (O'Sullivan and Unwin 2003), increasing the possibility of exposure misclassification. In addition, krigging requires numerous measurements to create a stable semivariogram which is not always possible when modelling air pollution. Finally, although kriging is easy to implement, it is not easy to understand the process resulting in the potential for statistically weak applications.

Kriging provided a surface at a fine spatial resolution capable of highlighting areas impacted by woodsmoke undetected by the fixed monitors. In addition, modeling the semivariogram to produce the krigged surface provided insight into the spatial scale of woodsmoke data. Therefore, kriging's strength is not as a model for characterizing exposure but as an exploratory tool and validation method used in conjunction with other modelling approaches.

LUR modelling is an approach that incorporates both measurement and predictive variables providing the potential to model a more realistic exposure estimate. A specific LUR model, the Larson Model, is the first attempt at a measurement-based spatially resolved estimate of woodsmoke exposure. It has the best performance (R^2=0.73, p<0.00, n=19) that is inflated due to the inclusion of collinear variables; although it is simple to

drop collinear variables to meet regression assumptions without jeopardizing model performance (R^2=0.54, p<0.00, n=19). The performance of the Larson Model is dependent upon a seasonal adjustment of the data that may not be valid since there is no seasonal trend in $PM_{2.5}$. The Larson Model is an improvement over the baseline scenario because it captures woodsmoke hot spots undetected by the fixed monitors.

The Larson Model is an ecological approach demonstrating the MAUP since the aggregation and adjustment of the data improves model performance. Because it is ecological, it is subject to the ecological fallacy and exposure misclassification where the entire population within a catchment basin is assigned the same exposure level regardless of their elevation in the basin.

The Larson Model predicts the spatial pattern of woodsmoke (high, medium and low areas). In addition, it is a novel approach to developing informed spatial units for modelling. Although the model violates the assumption of independence between predictor variables and the inclusion of immigrants is questionable, if the sole purpose of a model is to predict (not explain), then meeting the assumptions of regression modelling are not important (Mann 1987). Nevertheless, Larson et al. (2007) drew erroneous conclusions regarding the effect of immigrants on woodsmoke. The correlation between the number of immigrants and $PM_{2.5}$ attributable to woodsmoke is positive (R = 0.54), however, the model coefficient is negative. Model coefficients changing signs when entered into a regression model, is indicative of collinearity. Despite the positive correlation between immigrants and woodsmoke data, the authors conclude that immigrants have a negating affect on woodsmoke concentrations.

There are further limitations warranting consideration prior to the adoption of this model in the field of epidemiology and air quality management. If it is employed there needs to be an understanding of the trade off's between high model performance and limitations with an ecological model, the requirement for data adjustments, and the assumption of cool, calm and clear evenings (an invalid assumption). The model cannot be used to predict concentration levels, but is useful for highlighting areas exposed to higher concentrations.

The final model in Table 36 attempts to address the limitations of the Larson Model by:

- Improving the spatial resolution;
- Applying to a variety of meteorological conditions;
- Requiring no data adjustments;
- Providing flexibility for the exposure period to be predicted (i.e., a seasonal average, constructing cumulative exposure); and
- Building an explanatory model as opposed to a predictive model.

Nevertheless, the performance of the Larson Model cannot be matched.

The Model series (M1 through M7) was developed using SPAD improving the spatial resolution over the Larson Model. No data were aggregated during model development reducing the potential for exposure misclassification and the ecological fallacy. These models are more flexible because they apply to all weather conditions during the winter heating season and do not require any seasonal adjustments. The models can be used to predict and construct past exposure as defined by the user, as opposed to being limited to a seasonal average.

These models also attempt to explain the pattern of woodsmoke: As elevation increases, woodsmoke concentrations decrease, a phenomenon supported by theories of air pollution dynamics (Larson et al. 2007). As the density of low income homes increase, so does the concentration of woodsmoke. The theory is that the availability of wood as a low cost fuel renders it attractive to low income earners.[24] This theory is also supported by a survey conducted by the Ministry of Environment that found the low cost of wood burning to be the determinant for choosing wood as a residential heating source. This is also supported by the Larson Model because the income variable is the only defensible variable in the model. Through the *TOEM2* variable (a measure of the background $PM_{2.5}$), the model series incorporates stability conditions because concentrations at each monitor reflect the stability of conditions at that time and place. The further locations are from these monitors (i.e., Sydney and Metchosin) the more poorly the model performs. Temperature and wind speed may be better predictors at these locations as opposed to the *TEOM2* variable.

[24] The assumption here is that low income earners live in lower value housing.

Two different approaches were investigated to deal with spatial and temporal dependence in the data set: post modeling via spatial regression in GeoDa and prior to modeling by removing dependence in the data set based on the spatial and temporal dependence present. For the woodsmoke model, removing spatial and temporal dependence prior to modeling produced the most robust approach, meeting the assumptions of OLS regression, although model performance and parameters were consistent across modeling approaches. Spatial regression showed improvements over OLS models; however, the model specification order was not met indicating model specification problems. Since available census, meteorological, SPAD and geographical data were all exploited; there were no further variables available to examine, presenting an area for future improvements to the model.

Given the problems associated with meeting OLS regression assumptions, the inability to meet spatial regression model specifications and the lack of improvement from the Bayesian approach, M6 was chosen as the new modeling approach because it meets modeling assumptions and predicted values show a strong, positive association with measured values ($R = 0.63$, $\alpha=0.01$).

Is M6 an improvement over the Larson Model? Quantitatively, the Larson Model performs better, qualitatively, M6 is superior. This is where the trade off between spatial resolution and model performance arises. As data are aggregated, model performance improves as an artifact of the MAUP. As the spatial resolution increases, the potential for exposure misclassification declines; however, it is difficult to predict accurately at such a fine resolution due to the presence of so many variables at this scale (i.e., it is difficult to predict individual behaviour).

M6 is recommended even though its performance is inferior to the Larson Model because:

- It meets modeling assumptions;
- It applies to all meteorological conditions;
- It allows the user to specify the period of exposure to be predicted;
- It has less potential for exposure misclassification;
- It is an explanatory model providing a policy tool to support air pollution reduction strategies.

The final model is presented in Figure 62 and was constructed using Spatial Analyst's 'Raster Calculator':

$$PM_{2.5} = 7.01 - (0.05*[ELEV]) + (0.62*[TEOM2]) + (0.03*[TVD25P]) \quad (14)$$

Because *ELEV* and *TVD25P* remain relatively constant (*TVD25P* can be updated with new releases of SPAD data), they predict the spatial distribution of woodsmoke and *TEOM2* determines the concentration levels predicted. This provides flexibility because the *TEOM2* variable can be defined by the user as the average value for an evening, a winter heating season, or several heating seasons. Past exposure can be reconstructed by adding surfaces together using the raster calculator to provide a more realistic construction as opposed to a seasonal average. Figure 60 shows the woodsmoke model for an evening where the average of the 3 fixed sites for the evening was 19 µg/m^3.

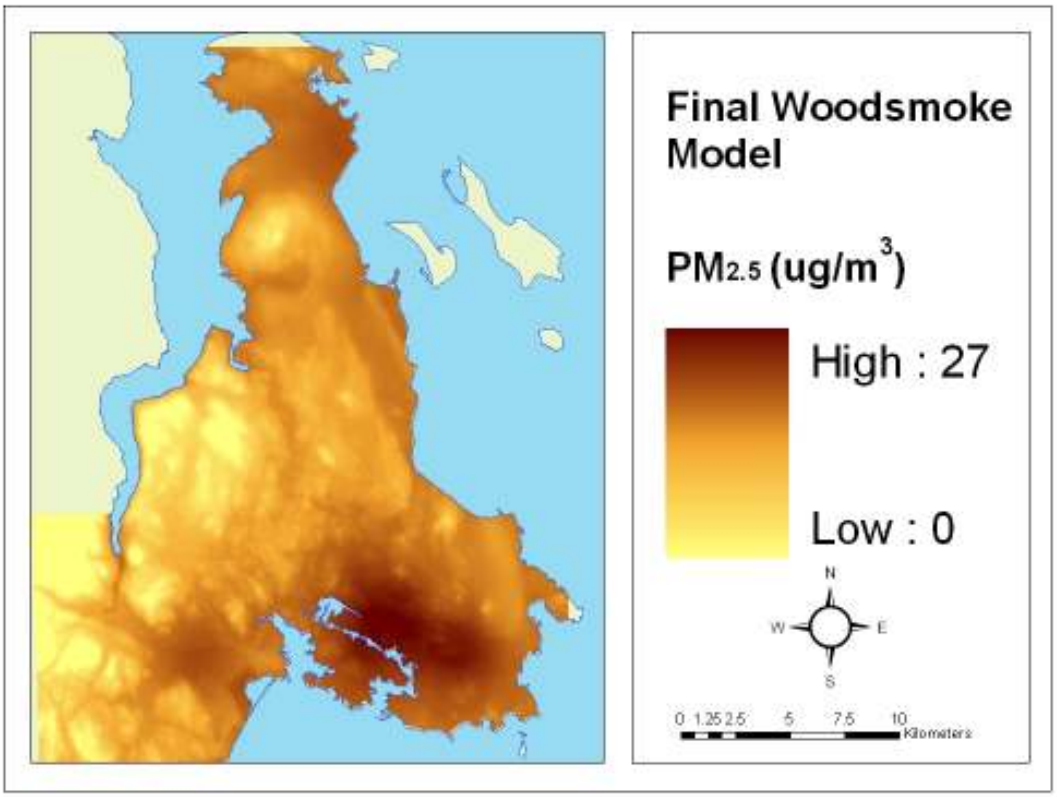


Figure 62. Spatial distribution of woodsmoke concentrations from residential wood burning for a hypothetical evening in the Capital Regional District

Figure 63 is a map of the residual error between observed and predicted values (see also Figure 60, a cluster analysis of residuals). The green circles show where the model under predicts and the red shows where the model over predicts. Generally, in the north, the model is over predicting $PM_{2.5}$ whereas in the southwest (Langford/Colwood), the model is under predicting $PM_{2.5.}$ This suggests that different determinants of residential woodsmoke concentrations are at work in these areas and that one model does not apply equally to all areas unless other variables capturing this difference can be incorporated.

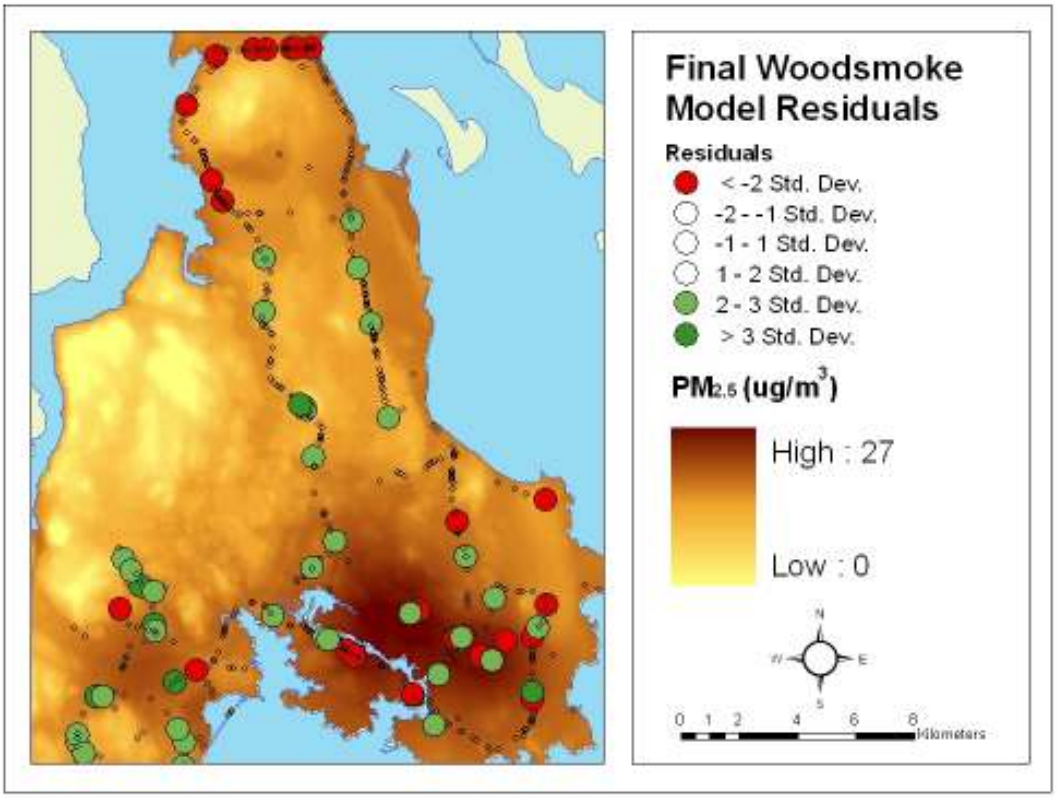


Figure 63. M6 regression model residual error

The next Chapter discusses the results of applying M6 to a real world scenario.

Chapter 5:
Applying the Woodsmoke Model to a Practical Example

5.1 Introduction to the Risk Characterization

Results of an exposure assessment are often used in the risk assessment process (Figure 6). This chapter is a preliminary characterization of the health risk associated with woodsmoke in the CRD and demonstrates how the woodsmoke model can be applied in a real-world context. As discussed in the literature, there is no evidence for quantifying carcinogenic effects, nor is there enough evidence to develop a dose-response relationship for exposure to woodsmoke making it impossible to quantify the health risk associated with woodsmoke. This risk characterization then is based on the evidence that woodsmoke poses a health risk and that health risk varies depending on susceptibility, mobility and inhalation rates within a population.

Columns H, I, and J in Table 3 represent the ideal information to include in a risk characterization. The exposure assessment process identifies the pollution sources, pollutant concentrations and duration and intensity of exposures based on time spent in different environments. This is combined with available dose-response information (columns B-D) to arrive at estimates of health effects (columns H-J). Since most of these data are unavailable this risk characterization is an exploratory step.

5.2 Sources of Woodsmoke

According to Williams and Paustenbach (2002), the first items to identify in a risk characterization are the sources of the air pollution. Fortunately, this is simple: woodstoves and fireplaces used for residential heating are the sources of air pollution. Figure 62 shows the density of residential fireplaces throughout the CRD, displaying the hot spots for potential air pollution due to residential heating. This map was developed using SPAD by selecting residential homes with greater than or equal to one fireplace. A kernel density estimator was run based on the selected points with a search radius of 1km and a cell size of 50m. The search radius of 1km was chosen based on spatial dependence observed in the fireplace data using semivariogram analysis.[25] There is a field in the

[25] The search radius differs from the one used to develop indicator variables because this density measure is not related to the scale of woodsmoke, only the locations of fireplaces.

SPAD indicating the primary heating source of the home. Homes using gas as the primary heating source with a fireplace were removed from this analysis based on the assumption that the fireplace is gas; however, this made no discernable difference to the output in Figure 64.

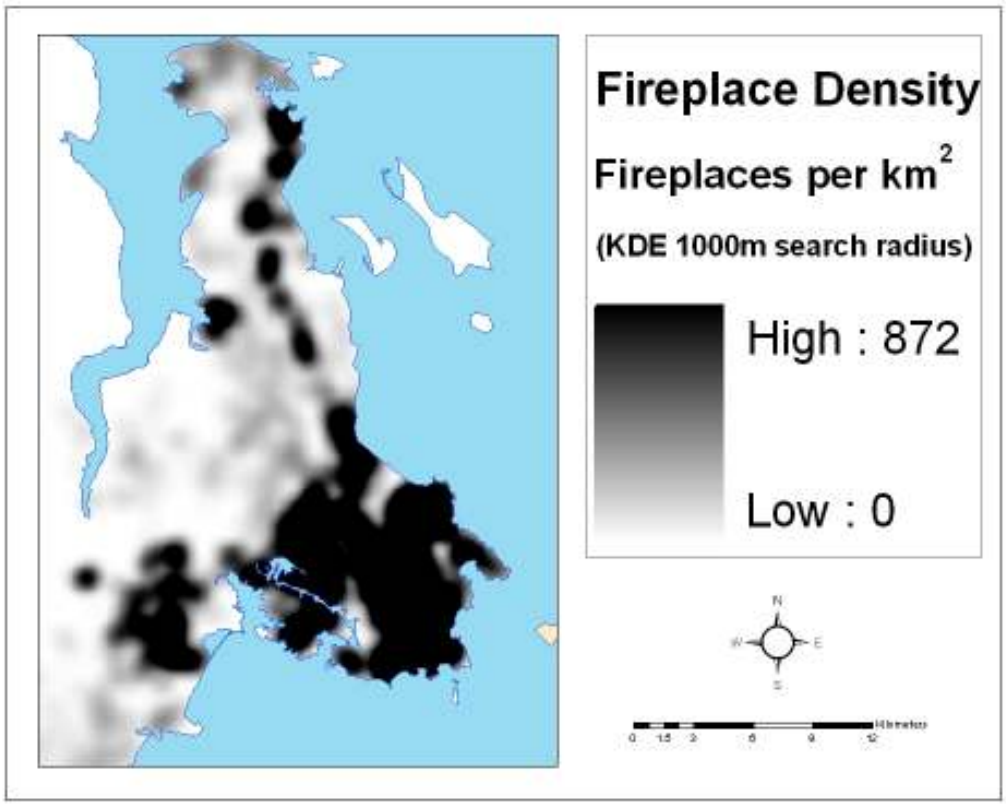


Figure 64. Residential fireplace density in the CRD

5.3 People Likely Exposed

The next step in a risk characterization is to identify the number of people likely exposed (Williams and Paustenbach 2002). Figure 65 shows residential density, looking no different from fireplace density. Since wood burning for heat occurs where people live, at a time when they are at home (Zelikoff et al. 2002), most residents in the CRD are likely exposed.

Figure 66 is a crude estimate, similar to the one reviewed in the literature by Scoggins et al. (2004), showing the number of people exposed to low, medium and high woodsmoke concentrations attributable to residential heating. The low, medium and high

rankings are relative and were divided using natural breaks.[26] Although this map provides crude estimates of the numbers of people exposed by DA, it demonstrates how the woodsmoke model can be combined with other data sets to estimate health risk. The map in Figure 66 was created by averaging the $PM_{2.5}$ values from M6 to the DA level using Spatial Analyst's 'Zonal Statistics.' The DA's were then divided into the bottom third, middle third and top third values for $PM_{2.5}$ concentrations based on natural breaks. The total population in each category was then summed.

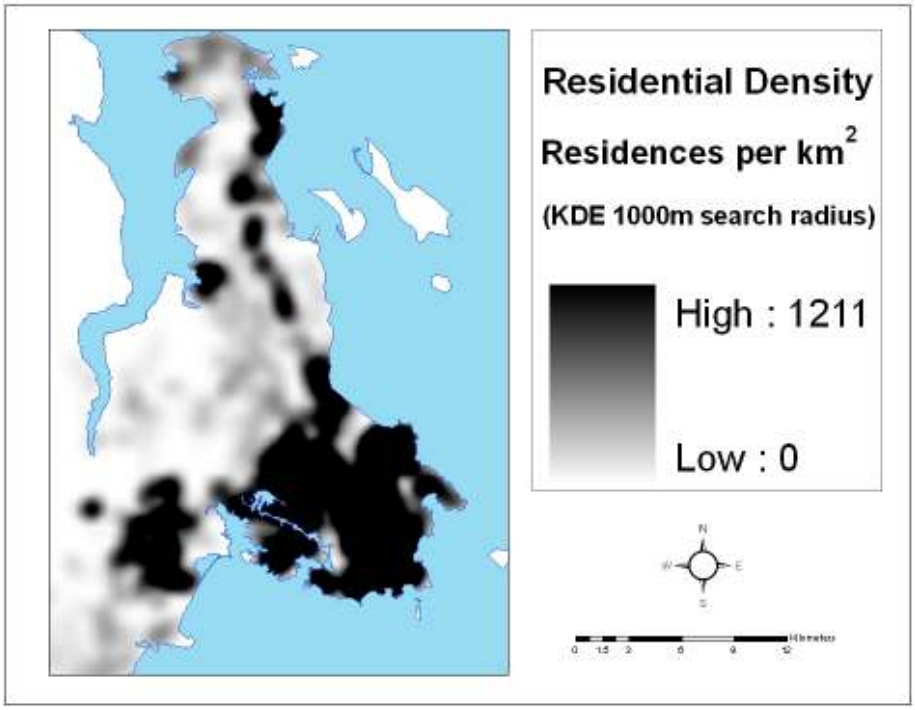


Figure 65. Residential density in the CRD

[26] The natural break method was chosen because it defines groups based on the variation in the data. It minimizes variation within groups while maximizing variation between groups thereby defining natural clusters. Natural breaks are used throughout the risk characterization to define groups.

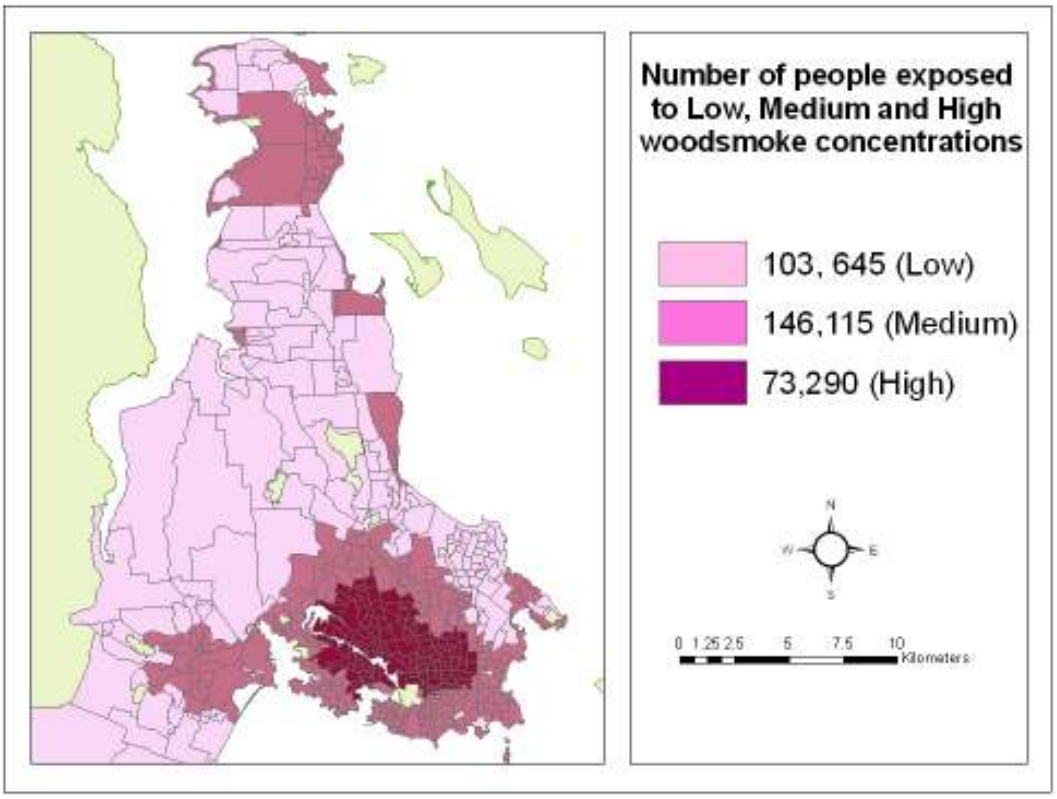


Figure 66. The number of people exposed to low, medium and high woodsmoke concentrations by DA (displayed using natural breaks)

Understanding the air pollution and health relationship requires identifying areas of maximum overlap between exposure and susceptibility. This overlap indicates the areas of highest risk (Jerrett and Finkelstein 2005). The exposure surface was created in Chapter 4 leaving the spatial distribution of susceptibility to be defined.

Health research typically uses census data, such as age and income, to characterize inequalities in susceptibility. Since susceptible population data are not available through SPAD or at the individual level, the data used to define susceptible populations for this risk characterization come from the census at the DA level. Percent low income, children under 5 and people over 70 were selected to describe susceptible populations. This is not a thorough definition of susceptible populations; however, these populations were chosen because they are considered more susceptible to health effects from air pollution in the literature (Larson & Keonig 1994; Naeher et al. 2007). Figure 67 shows the percentage of each DA population defined as low income.

To define the Geography of Risk for low income populations, areas with the highest percentage of low income residents (greater than 28% low income), overlapping with areas experiencing high woodsmoke concentrations (greater than 21µg/m^3) were selected in ArcMap and exported (Figure 68). 10 DA's were selected in the downtown, Marigold and Gorge areas.

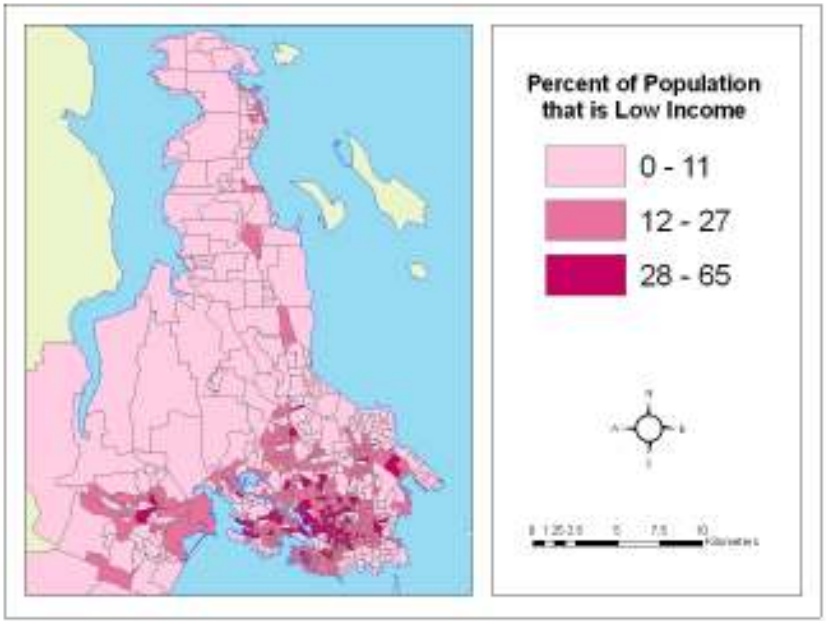


Figure 67. Percentage of population that is low income by dissemination area (presented using natural breaks)

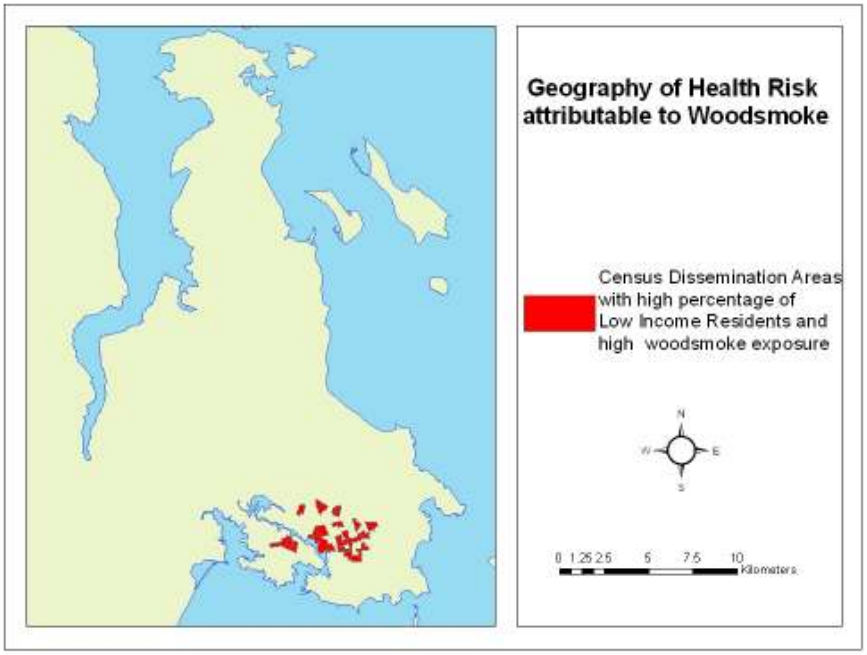


Figure 68. Geography of health risk for low income populations by DA

Figure 69 shows the percentage of the DA population under age 5. Figure 70 shows the area of overlap for DAs with a high percentage of children under 5 (over 7.1%) living in an area categorized as having high woodsmoke exposure. The Langford/ Colwood area demonstrates a significant limitation in the risk characterization: M6 under predicts $PM_{2.5}$ in this area where there are a number of DAs with a high percentage of children under 5. As discussed in the literature review, negative health outcomes associated with woodsmoke were particularly pronounced among children (Larson and Koenig 1994; Boman et al. 2003; Naeher et al. 2007); therefore, it is important to be able to identify these areas accurately in a risk characterization.

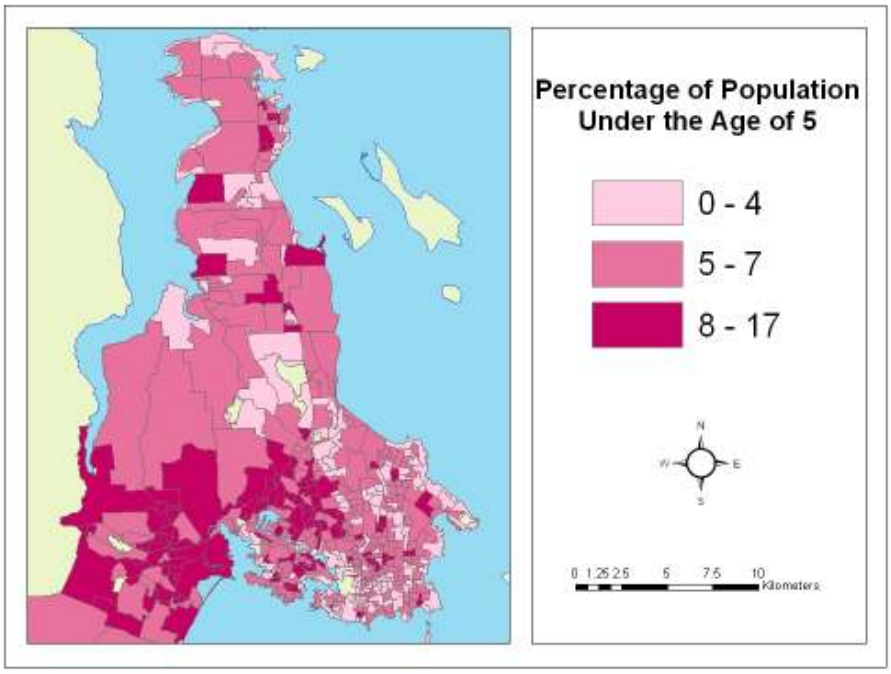


Figure 69. Percentage of population that is under 5 years old by dissemination area (presented using natural breaks)

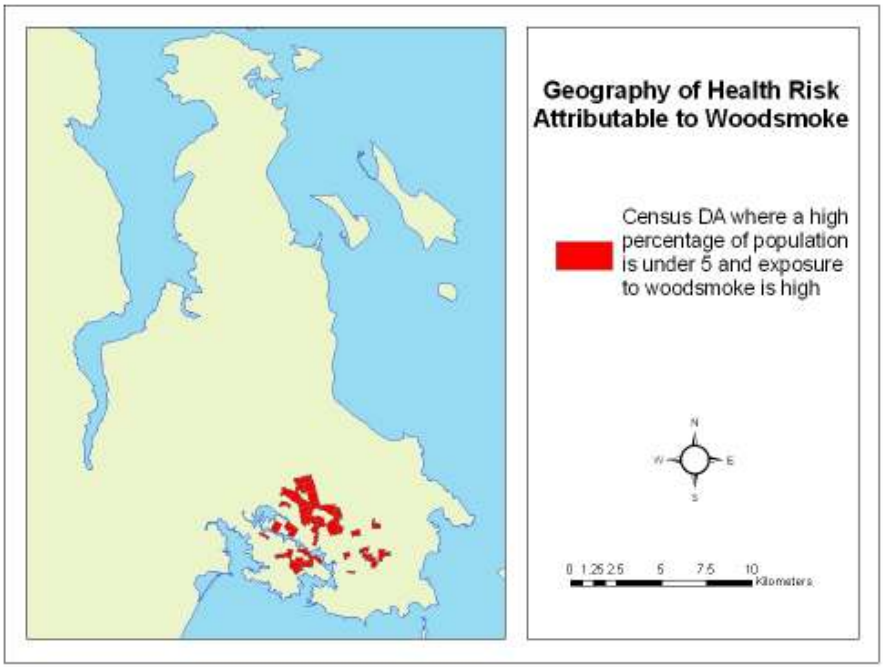


Figure 70. Geography of health risk for children under the age of 5 by dissemination area

The percentage of the population over the age of 70 by DA is shown in Figure 71. Population over the age of 70 does not show much overlap with high exposure to woodsmoke; however, it does occur in similar areas to low income populations and children under the age of 5 (Figure 72). Figure 73 shows the Geography of Health Risk attributable to woodsmoke as defined by the percentage of the population that is low income, the percentage of children under the age of 5 and the percentage of population over the age of 70. The high risk areas occur in the Gorge/Marigold area; however, Langford and Colwood have a high percentage of children that may be exposed to higher levels than those indicated by the woodsmoke model.

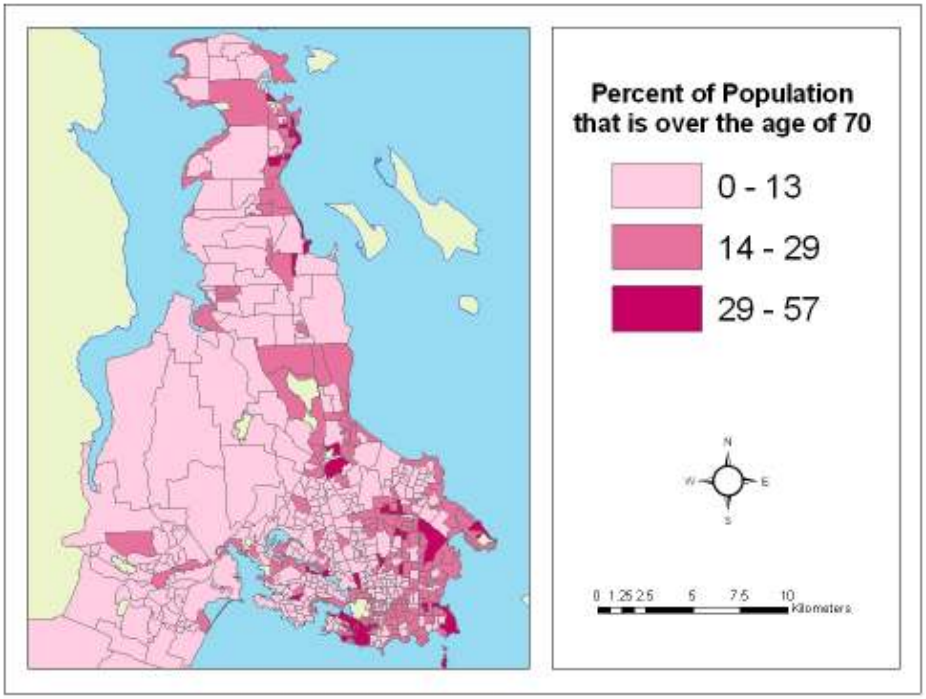


Figure 71. Percentage of population over age 70 by dissemination area

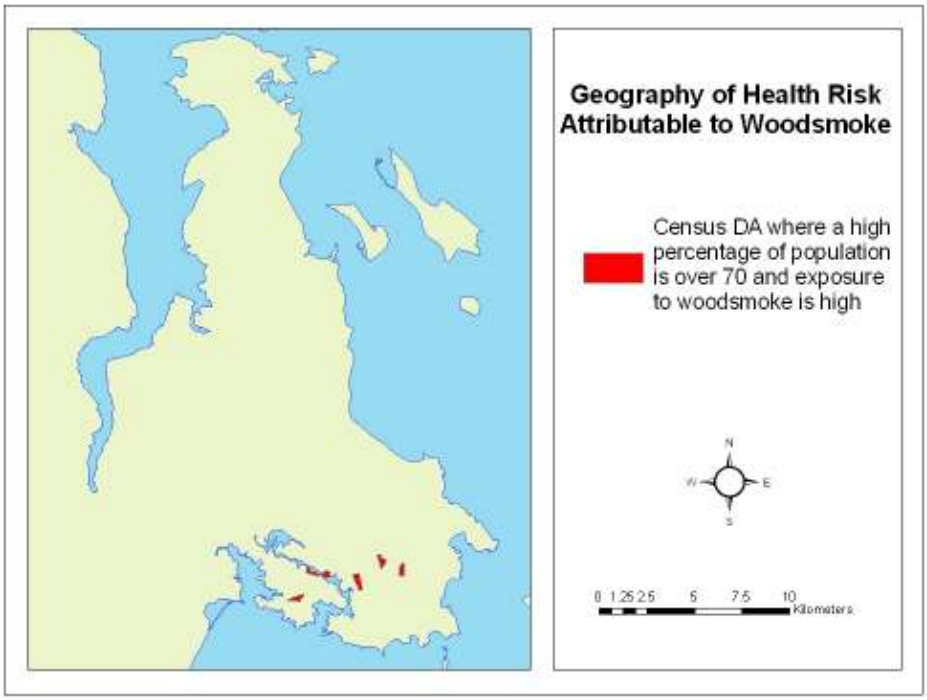


Figure 72. Geography of health risk for population over 70 by dissemination area

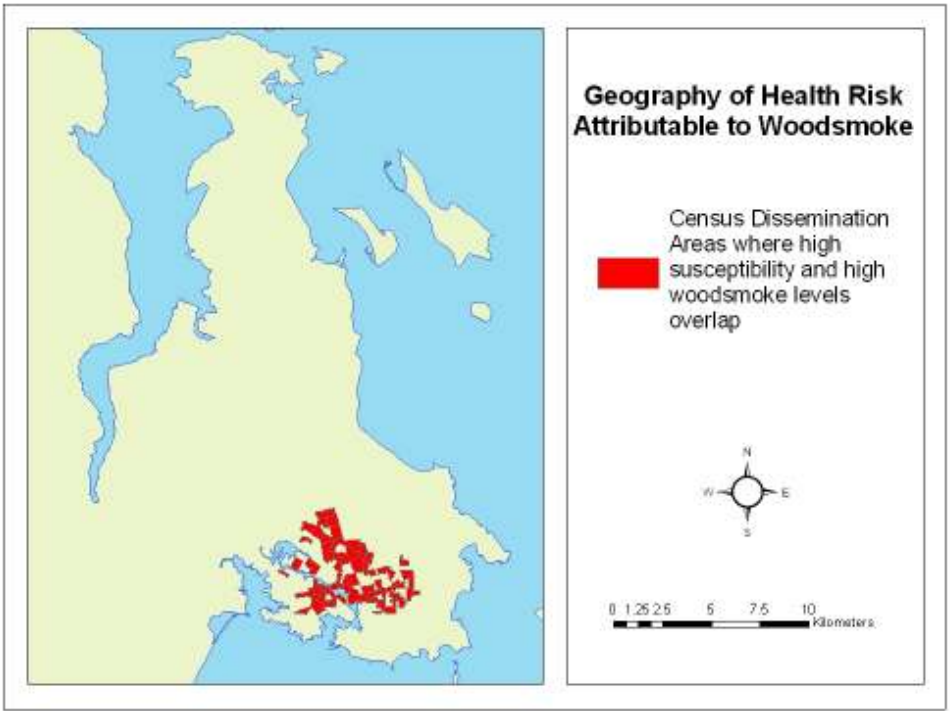


Figure 73. Geography of health risk attributed to woodsmoke for low income populations, children under the age of 5 and people over 70 by dissemination area

Williams and Paustenbach (2002) suggest including a discussion of cumulative or multiple exposures in a risk characterization. There are no studies relating long-term exposure to woodsmoke with health. The woodsmoke model developed here provides a tool that can be used to further this investigation because the model can be built to reconstruct past exposure to the extent that fixed monitoring or meteorological data are available. Although evidence is limited, there is biological plausibility to suggest that long term or multiple exposures are a health concern.

Confidence in the results from this risk characterization is low due to the limitations inherent in this risk characterization (discussed further in Section 5.4); however, this provides a useful starting point for air quality reduction strategies as well as identifying highly susceptible areas to target for health research.

5.4 Merits and Limitations of the Risk Characterization

There are a limited number of spatial approaches to health risk assessment and Jerrett and Finkelstein (2005) call for more comparisons between spatial and non-spatial approaches. Although the risk characterization is limited, it is the first attempt to characterize risk associated with woodsmoke and provides a starting point for future refinements.

All risk assessments have limitations due to the propagation of uncertainty as data sets of varying quality are integrated. For example, Figure 35 shows that the Langford/ Colwood area is exposed to higher than average levels of woodsmoke concentrations, but this area is under predicted by the woodsmoke model. Therefore, this error is propagated through to the Geography of Risk stage where this area is neglected as a high risk area.

Natural breaks were used to define high, medium and low susceptibility and exposure. The areas identified as high risk depend on how risk is defined. This definition, while based on variability in the data, imparts subjectivity into the analysis. If the data were divided based on equal intervals or areas below and above the mean, the number of high risk areas would be lower for equal intervals and higher using the mean. A way to address the effect of subjectively defining risk is to try several different approaches to

characterizing exposure and risk to determine if there are consistencies in the areas identified as high risk. Consistency would lend credibility to the findings.

No census data were used to construct the woodsmoke model; therefore, using the census data to define susceptible populations introduces the ecological fallacy and the MAUP to the risk characterization. Nevertheless, DA and census tract data can be useful in identifying what Jerrett and Finkelstein (2005) call contextual susceptibility. Contextual susceptibility refers to the influence of a person's place or environment on their health (cf. compositional susceptibility which refers to individual characteristics such as genetics). Contextual susceptibility, such as the socioeconomic status of a neighbourhood, have shown tremendous influence on health, producing odds ratios as high as 2 in health research (for comparison purposes, air pollution produces odds ratios of about 1.1-1.17).

The lack of evidence regarding the health effects of woodsmoke presents a significant limitation to this characterization. In addition, lack of available data relating to individual susceptibility and mobility at an individual level also limits the findings. In this risk characterization, health risk coincides with the exposure surface. Using ambient concentrations as a proxy for exposure fails to account for:

- personal mobility;
- variations in people's time-activity patterns;
- variations in infiltration rates of microenvironments;
- variations in breathing rate; and,
- variations in susceptibility (due to differences in genetics, immunity and age).

In addition, a full health risk characterization incorporates health end point data such as asthma or disease incidence and considers all potential routes for exposure such as ingestion, absorption, and inhalation (Table 3).

This risk characterization leaves much room for improvement but is included as a method to exhibit the potential use of the woodsmoke model to provide a flexible tool to use in future research to address these issues.

Chapter 6: Conclusions

The final chapter discusses conclusions from the research (Section 6.1), implications for spatial analysis, health research and policy (Section 6.2), and provides policy recommendations (Section 6.3). The chapter concludes by outlining areas for future research (Section 6.4).

6.1 Summary

Within the context of global climate change and soaring energy prices, people are searching for inexpensive and renewable sources of energy, and as a result, burning wood for home heating is on the rise. Human exposure to woodsmoke is so common that it is regarded as relatively benign and therefore little is done to address it. There is limited research into the health effects of woodsmoke and existing research suffers from methodological problems. Nevertheless, there is sufficient coherence among existing studies to suggest woodsmoke poses a serious threat to human health. With woodsmoke from residential heating being identified as a major contributor to air pollution in BC, there is growing interest in producing robust exposure estimates for health research and air quality management purposes.

Spatial analysis is relatively new to the field of health research. As investigators begin to see the influence of spatial processes on research findings, the importance of adopting a spatial approach to characterizing exposure and risk is becoming apparent. Since the associations between health outcomes and exposure to air pollution are dependent upon the exposure metric, there is a requirement to produce robust estimates as uncertainty could thwart efforts to address the issue (Jerrett and Finkelstein 2005).

This thesis contributes to the fields of woodsmoke and health research and it makes methodological contributions to the fields of exposure assessment and health geomatics by developing a spatial statistical model characterizing the spatial distribution of woodsmoke particulates throughout the CRD during the winter heating season.

Four research questions were outlined in the introduction that will now be addressed:

1. ***Is the current fixed monitoring network representative of the spatial distribution of woodsmoke throughout the CRD?***

The results of the baseline scenario using the closest of three fixed monitors to characterize air pollution indicate that the fixed monitoring network does not capture woodsmoke hotspots; therefore, any research employing this technique is subject to exposure misclassification.

Spatial dependence in the woodsmoke particulate data occurs up to approximately 2.7 km, providing an estimate of the area captured by each monitor. Figure 20 displays the approximate area represented by each monitor demonstrating that it is not representative of the spatial distribution of woodsmoke particulates.

2. How does the spatial distribution of woodsmoke differ from the Larson model predicting woodsmoke levels in the CRD?

The Larson Model is an improvement over the baseline scenario for predicting the spatial distribution of woodsmoke throughout the CRD. Nevertheless, limitations include:

- The potential for moderate exposure misclassification;
- It predicts a narrow range of seasonal values;
- It is dependent on adjusting the data for seasonality;
- It contains collinear variables artificially increasing performance metrics; and,
- It assumes exposure is negligible within catchments near the coast.

Drawing upon the lessons learned from the Larson Model, a new woodsmoke model was created with a finer spatial resolution. Where possible, the data were left at their original resolution with no seasonal adjustments thereby reducing the potential for exposure misclassification. This model was developed based on an understanding of spatial scale for analysis; addressing a limitation in typical exposure models that neglect the effect of spatial scale on research results. Several different methods were examined for dealing with spatial and temporal dependence to meet the requirements for OLS regression.

M6 (LUR model built by removing spatial and temporal dependence prior to modelling) sought to explain, rather than simply predict, the pattern of woodsmoke observed. It is a robust model applying to all meteorological conditions and it is flexible because it allows the user to construct the period of exposure to be estimated by the

model. The model was built using data available for BC, and as a result, is transferable to other areas (however, a measurement campaign must be conducted to evaluate and refine the model).

3. Is exposure to woodsmoke negligible on evenings that are not cool, calm and clear?

Levoglucosan measurements indicate that woodsmoke is present, although at lower concentrations, during unstable conditions. In addition, stability does not apply to the entire CRD at one time and therefore, an evening cannot be characterized as having entirely stable or entirely unstable conditions.

4. What is the spatial distribution of health risk attributable to woodsmoke throughout the CRD?

M6 was input to characterize health risk associated with $PM_{2.5}$ attributable to woodsmoke from residential heating during the winter heating season throughout the CRD. The risk characterization only depicted contextual susceptibility because data with respect to individual susceptibility were not available. Contextual susceptibility investigated in this characterization included the percent of the population that is low income, the percent of the population under the age of five and the percent of population over the age of 70, all of which show increased susceptibility to health outcomes associated with air pollution. The number of people exposed to Low, Medium and High woodsmoke concentrations were also estimated.

The spatial distribution of health risk is fairly congruent with the exposure surface with areas of higher risk occurring in the Gorge/Marigold area (large percentage of low income and children under 5) and the Langford/Colwood area (a high risk area for children under the age of 5). However, the Langford/Colwood area was not identified by the risk characterization process because the exposure surface underestimated concentrations in that area.

The notion that risk and exposure are not equivalent is imperative for understanding the limitations inherent in the woodsmoke model and subsequent risk characterizations. Several determinants of risk are neglected when using ambient concentrations to estimate exposure including personal mobility, microenvironments having different infiltration rates and breathing rate variability (Marshall et al. 2006). Nevertheless, individuals typically spend 60-70% of out of work time at home (Zelikoff

et al. 2002), when residential heating is more likely to occur; therefore, using residential addresses may be an appropriate assumption. It was difficult to characterize susceptible populations accurately because they are always changing and they are difficult to identify due to differences in genetics, immunity, and age.

6.2 Research and Policy Implications

This research is built on a framework for the spatial analysis of exposure to woodsmoke; the principles of which can be applied to other fields interested in spatial analysis. The framework followed Anselin's (2006) approach beginning with ESDA, followed by visualization, and spatial modelling. The exploratory analysis is summarized in Chapter 3 and included characterizing the spatial and temporal dependence in woodsmoke data to identify the spatial scale of analysis. This informed the development of various methods to deal with autocorrelation in the data set. For the woodsmoke model, the best method for dealing with spatial and temporal autocorrelation as a side effect, was to remove it prior to modelling. GIS layers were created for independent variables and associations with measured PM were investigated. Kriging provided a visualization tool to depict measured woodsmoke patterns. Kriging did not provide a robust surface for use in exposure assessment; therefore, the next step incorporated measurements and modelling to predict and explain the observed patterns.

The SPAD data provides a viable alternative to census data for improving the spatial resolution of exposure estimates, cited as a limitation in exposure assessment literature. Although census variables showed significant associations with woodsmoke, they were rendered insignificant when similar SPAD variables were included in the modelling process. Therefore, SPAD not only improves the spatial resolution of exposure models, it also improves model performance.

Another limitation in the woodsmoke and health literature addressed in this thesis is confirming that the source of $PM_{2.5}$ is attributable to woodsmoke. A simple and inexpensive monitoring method to measure levoglucosan was outlined and can be used to validate any future woodsmoke measurements.

A high resolution woodsmoke model was produced to estimate people’s exposure to woodsmoke during the winter heating season. The implications of these improved estimates for policy and research purposes include:

- More certainty in exposure estimates lends credibility to health research findings;
- Knowing the spatial scale of woodsmoke for future analysis and modelling since little is known about the spatial scale at which air pollutants vary;
- Identifying the optimal sampling interval for measuring woodsmoke;
- Identifying woodsmoke hot spots undetected by fixed monitors to improve or design air quality monitoring strategies;
- Identifying locations for targeted political intervention;
- Identifying the underlying causes of woodsmoke pollution;
- Monitoring the effectiveness of policy interventions; and,
- Providing a model that is transferable to small communities lacking in resources where wood burning is a health concern.

6.3 Policy Recommendations

The woodsmoke model requires some fixed monitoring be in place. Where no fixed monitoring occurs, temperature and high resolution wind speed data become predictors of woodsmoke (although, if no fixed monitoring is occurring, it seems unlikely that high resolution wind speed data would be available). The mobile monitoring campaign outlined in Chapter 3 provides an inexpensive, high resolution measurement strategy for identifying ideal locations for a fixed monitor. In the CRD it is recommended that fixed monitors be placed in Sydney, Langford and Metchosin. This would improve M6 performance in the modelling of $PM_{2.5}$ attributable to woodsmoke from residential heating.

In addition to having fixed monitoring in place, the size of measurement is also important to reduce exposure misclassification and measurement error. Most research relating to woodsmoke and health measure PM_{10}; however, woodsmoke particles are typically smaller than 1μm. Therefore, if fixed monitoring is in place, there needs to be a shift towards measuring particles smaller than 2.5 μm.

This model was developed for the CRD, a coastal area experiencing relatively good air quality. Although the model highlights areas of higher risk, health effects of woodsmoke are more of an issue in rural parts of BC. Given the likelihood of increased

wood burning for residential heating particularly in rural BC, the higher incidence of respiratory illness in these areas and the adverse socio-economic conditions facing many communities, there is a need for woodsmoke modelling in these communities both from an air quality management and a health perspective.

Burning wood for residential heating is associated with adverse health impacts; however, wood provides an inexpensive, renewable energy resource. Therefore, strategies are required to use wood as a safe and efficient fuel. Fortunately, the existence of efficient woodstoves, pellet stoves, fireplaces and wood furnaces makes this a viable option. These technologies create the capacity for a win-win situation. They are more efficient because they burn less wood and require less maintenance (thereby lowering operating costs and labour) and they reduce indoor and outdoor emissions by 90% over a conventional wood stove (Natural Resources Canada 2002).

If a policy is implemented to exchange old wood stoves, it must include the installation of a US EPA certified[27] low-emission stove by a certified professional because an improperly installed efficient woodstove can be as inefficient as a conventional woodstove. Woodstoves also need to be inspected by a professional on an annual basis. In addition, informing the public on the types of wood to burn (i.e., well-seasoned, split and dried wood) and the types of wood to avoid (i.e., "green" wood, driftwood, particle board, plywood, painted/treated wood) to prevent the release of more toxic chemicals will help reduce health risk (Natural Resources Canada 2002).

The biggest hurdle for installing a new wood burning appliance is cost. In a market survey conducted by the Ministry of Environment, people who burn wood for heat expressed no concern for health effects and will continue to burn as long as it is cost effective. This suggests that a cost incentive program may be required to support a stove exchange. In the Skeena-Bulkley Valley, a woodstove exchange has been legislated for all inefficient woodstoves by the year 2010 to deal with the air quality problems associated with burning in the valley. In addition, Local and Provincial Governments are providing financial compensation for the exchange. Monitoring the success of this program will provide insight into the benefit of expanding the program province-wide.

[27] These standards are endorsed by the Government of Canada

6.4 Future Research

The woodstove exchange study mentioned in the previous section provides a unique scientific opportunity to study health prior to, during and post implementation of a policy intervention designed to improve environmental quality and population health. In addition to providing insight into health research questions related to woodsmoke and health, it also provides an opportunity to assess the success of a public sector policy and to inform rural health policy. The study population includes residents from five communities in the Skeena/Bulkley Valley (Burns Lake, Houston, Telkwa, Smithers and Terrace). Community rates of respiratory and cardiovascular disease, and school absenteeism can be followed from 2002 (five years prior to the initiation of the woodstove exchange), until 2012, two years after the completion of the intervention program and the implementation of bylaws prohibiting the use of old wood stoves. Some research questions to address include:

1. What (if any) social processes make people more susceptible to adverse health effects attributable to woodsmoke in rural BC?
2. How important are the adverse health effects attributable to woodsmoke in the context of determinants of health in rural communities?
3. Do inequities[28] in environmental risk exist in rural BC?
4. Has the woodstove exchange program succeeded in reducing air pollution and the negative health outcomes associated with woodsmoke?
5. Can similar initiatives be implemented across BC? If so, how can it be improved?

Despite the ubiquitous nature of woodsmoke, little is known about the biological mechanisms leading to health effects, the effects of long term exposure and carcinogenic effects, leaving the area open for research. Model M6 provides high resolution estimates of concentrations to estimate people's exposure to woodsmoke at the residential level that can be used to advance the understanding of the impact of woodsmoke on health. The model is developed so that refinements such as time-activity patterns, indoor infiltration, personal monitoring, and intake fraction can be incorporated to further refine estimates.

[28] Environmental justice refers to the phenomenon where people in low socioeconomic positions (SEP) are more exposed to environmental hazards than those with higher SEP. For example, are lower income individuals more exposed to woodsmoke than higher income individuals in rural areas?

Spatial regression showed improvements over OLS models; however, the model specification order was not met indicating model specification problems. Since available census, meteorological, SPAD and geographical data were exploited; this presents an area for improving the model. One option includes improving elevation data using LIDAR which has the potential to provide elevations with a high degree of accuracy and spatial resolution (accurate to the sub-metre level). Nonetheless, acquiring project specific LIDAR imagery is prohibitively expensive if it is unavailable via other means.

It is concluded that model M6 is an improvement over the Larson Model, even though quantitatively, the Larson Model performs better. As data are aggregated, model performance improves as an artifact of the MAUP. As the spatial resolution increases, the potential for exposure misclassification declines; however, it is difficult to predict accurately at a fine resolution. Therefore, the trade off between spatial resolution and model performance needs to be investigated to determine if a lower performance model, although more robust, provides better estimates of exposure in health research.

This research employed OLS regression because it is a relatively simple, inexpensive, and robust method for modeling air pollution. Other methods to explore include GWR, neural network modeling meanwhile research is beginning on a Bayesian entropy approach to kriging as well as a Bayesian approach to spatially varying estimates of regression parameters. Regardless of the approach employed, any exposure modeling requires validation through a personal monitoring campaign.

Bibliography

Anselin, L. (2005). Exploring Spatial Data with GeoDa: A Workbook. Urbana, Il, Centre for Spatially Integrated Social Science.

Anselin, L. (2006). "How (not) to lie with spatial statistics." American Journal of Preventive Medicine **30**(2): S3-S6.

Boman, B. C., A. B. Forsberg and B. G. Jarvholm (2003). "Adverse health effects from ambient air pollution in relation to residential wood combustion in modern society." Scandinavian Journal of Work, Environment & Health **29**(4): 251-260.

Boman, C., B. Forsberg and T. Sandstrom (2006). "Shedding new light on wood smoke: a risk factor for respiratory health." European Respiratory Journal **27**(3): 446-447.

Boots, B. (2002). "Local measures of spatial association." Ecoscience **9**(2): 168-176.

Boudet, C., D. Zmirou and V. Vestri (2001). "Can one use ambient air concentration data to estimate personal and population exposures to particles? An approach within the European EXPOLIS study." Science of the Total Environment **267**(1-3): 141-150.

Boulos, M. N. K., A. V. Roudsari and E. R. Carsen (2001). "METHODOLOGICAL REVIEW Health Geomatics: An Enabling Suite of Technologies in Health and Healthcare." Journal of Biomedical Informatics(34): 195-219.

Brauer, M. (2002). Source, emissions, concentrations, exposures and doses. A Citizen's Guide to Air Pollution. D. V. Bates and R. B. Caton. Vancouver, David Suzuki Foundation.

Brauer, M., G. Hoek, P. VanVliet, K. Meliefste, P. Fischer, U. Gehring, J. Heinrich, J. Cyrys, T. Bellander, M. Lewne and B. Brunekreef (2003). "Estimating Long-Term Average Particulate Matter Air Pollution Concentrations: Application of Traffic Indicators and Geographic Information Systems." Epidemiology **14**(2): 228-239.

Briggs, D., S. Collins, P. Elliot, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. Van Reeuwijk, K. Smallbone and A. Van Der Veen (1997). "Mapping Urban Air Pollution Using GIS: A Regression Based Approach." Geographical Information Science **11**(7): 699-718.

Briggs, D. J., C. de Hoogh, J. Guiliver, J. Wills, P. Elliott, S. Kingham and K. Smallbone (2000). "A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments." Science of the Total Environment **253**(1-3): 151-167.

Brunekreef, B. and S. T. Holgate (2002). "Air Pollution and Health." Lancet **360**(9341): 1233-1242.

Buckeridge, D. L., H. Burkom, M. Campbell, W. R. Hogan and A. W. Moore (2005). "Algorithms for rapid outbreak detection: a research synthesis." Journal of Biomedical Informatics **38**(2): 99-113.

Burrough, P. A. and R. A. MCDonnell (1998). Principles of Geographical Information Systems. New York, Oxford University Press.

Cakmak, S., R. T. Burnett, M. Jerrett, M. S. Goldberg, C. A. Pope and R. J. Ma (2003). "Spatial regression models for large-cohort studies linking community air pollution and health." Journal of Toxicology and Environmental Health-Part A **66**(16-19): 1811-1823.

Clark, W. C. (1985). "Scales of Climate Impacts." Climatic Change **7**(1): 5-27.

Cockings, S., C. E. Dunn, R. S. Bhopal and D. R. Walker (2004). "Users' perspectives on epidemiological, GIS and point pattern approaches to analysing environment and health data." Health & Place **10**(2): 169-182.

Comrie, A. C. (1997). "Comparing neural networks and regression models for ozone forecasting." Journal of the Air & Waste Management Association **47**(6): 653-663.

Cupitt, L. T., W. G. Glen and J. Lewtas (1994). "Exposure and Risk from Ambient Particle-Bound Pollution in an Airshed Dominated by Residential Wood Combustion and Mobile Sources." Environmental Health Perspectives **102**: 75-84.

Cyrys, J., M. Hochadel, U. Gehring, G. Hoek, V. Diegmann, B. Brunekreef and J. Heinrich (2005). "GIS-based estimation of exposure to particulate matter and NO2 in an urban area: Stochastic versus dispersion modeling." Environmental Health Perspectives **113**(8): 987-992.

Davis, J. H., R. W. Howe and G. J. Davis (2000). "A multi-scale spatial analysis method for point data." Landscape Ecology **15**(2): 99-114.

Diem, J. E. (2003). "A critical examination of ozone mapping from a spatial-scale perspective." Environmental Pollution **125**(3): 369-383.

Elliott, P., J. C. Wakefield, N. G. Best and D. J. Briggs (2000). Spatial Epidemiology: Methods and Applications. New York, Oxford University Press.

Englund, E. and A. Sparks (1991). Geostatistical Environmental Assessment Software User's Guide. Las Vegas, United States Environmental Protection Agency.

EPA (1995). Guidance for Risk Characterization, Environmental Protection Agency.

EPA (2005). TRIM: Total Risk Integrated Methodology, United States Environmental Protection Agency.

Flaherty, M. (2007). Professor. Victoria, University of Victoria.

Fotheringham, A. S., C. Brunsdon and M. Charlton (2000). Quantitative geography : perspectives on spatial data analysis. London

Thousand Oaks, Calif, Sage Publications.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996). Markov Chain Monte Carlo In Practice. Cambridge, UK, Chapman & Hall.

Greenberg, M. M. (1997). "The central nervous system and exposure to toluene: A risk characterization." Environmental Research **72**(1): 1-7.

Grivas, G. and A. Chaloulakou (2006). "Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece." Atmospheric Environment **40**(7): 1216-1229.

Hair, J. F., R. E. Anderson, R. L. Tatham and W. C. Black (1984). Multivariate Data Analysis. New Jersey, Prentice Hall.

Henderson, S. and M. Brauer (2005). Measurement and modeling of traffic-related air pollution in the British Columbia Lower Mainland for use in health risk assessment and epidemiological analysis. Vancouver, BC, School of Occupational and Environmental Hygiene and Centre for Health and Environment Research, The University of British Columbia.

Hoek, G., K. Meliefste, J. Cyrys, M. Lewne, T. Bellander, M. Brauer, P. Fischer, U. Gehring, J. Heinrich, P. van Vliet and B. Brunekreef (2002). "Spatial variability of fine particle concentrations in three European areas." Atmospheric Environment **36**(25): 4077-4088.

Jelinski, D. E. and J. G. Wu (1996). "The modifiable areal unit problem and implications for landscape ecology." Landscape Ecology **11**(3): 129-140.

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

Jerrett, M., R. T. Burnet, R. Ma, C. A. Pope, D. Krewski, B. Newbold, G. Thurston, Y. Shi, N. Finkelstein, E. E. Calle and M. J. Thun (2006). "Spatial analysis of air pollution and mortality in Los Angeles." Epidemiology **17**(6): S69-S69.

Jerrett, M., R. T. Burnett, M. S. Goldberg, M. Sears, D. Krewski, R. Catalan, P. Kanaroglou, C. Giovis and N. Finkelstein (2003a). "Spatial analysis for environmental health research: Concepts, methods, and examples." Journal of Toxicology and Environmental Health-Part A **66**(16-19): 1783-1810.

Jerrett, M., R. T. Burnett, R. J. Ma, C. A. Pope, D. Krewski, K. B. Newbold, G. Thurston, Y. L. Shi, N. Finkelstein, E. E. Calle and M. J. Thun (2005b). "Spatial analysis of air pollution and mortality in Los Angeles." Epidemiology **16**(6): 727-736.

Jerrett, M., R. T. Burnett, A. Willis, D. Krewski, M. S. Goldberg, P. DeLuca and N. Finkelstein (2003b). "Spatial analysis of the air pollution-mortality relationship in the context of ecologic confounders." Journal of Toxicology and Environmental Health-Part A **66**(16-19): 1735-1777.

Jerrett, M. and M. Finkelstein (2005). "Geographies of risk in studies linking chronic air pollution exposure to health outcomes." Journal of Toxicology and Environmental Health-Part A-Current Issues **68**(13-14): 1207-1242.

Johnston, K., J. M. Ver Hoef, K. Krivoruchko and N. Lucas (2003). Using ArcGIS Geostatistical Analyst. Redlands, ESRI.

Keitt, T. H., O. N. Bjornstad, P. M. Dixon and S. Citron-Pousty (2002). "Accounting for spatial pattern when modeling organism-environment interactions." Ecography **25**(5): 616-625.

Larson, T., J. Su, A. M. Baribeau, M. Buzzelli, E. Setton and M. Brauer (2007). "A spatial model of urban winter woodsmoke concentrations." Environmental Science & Technology **41**(7): 2429-2436.

Larson, T. V. and J. Q. Koenig (1994). "Wood Smoke: Emissions and Noncancer Respiratory Effects." Annual Review of Public Health **15**(1): 133-156.

Lepage, M. and J. W. Boulton (2000). Source Apportionment of Particulate Matter in Canada. Guelph, Ontario, RWDI Group.

Lippmann, M. and R. B. Schlesinger (2000). "Toxicological bases for the setting of health-related atr pollution standards." Annual Review of Public Health **21**: 309-333.

Mann, C. J. (1987). Misuses of Linear Regression in Earth Sciences. Use and Abuse of Statistical Methods in the Earth Sciences. W. Size. New York, Oxford University Press**:** 74-106.

Marshall, J. D., P. W. Granvold, A. S. Hoats, T. E. McKone, E. Deakin and W. W. Nazaroff (2006). "Inhalation intake of ambient air pollution in California's South Coast Air Basin." Atmospheric Environment **40**(23): 4381-4392.

Marven, C. (2006). Methodologies for Identifying Marine Spill Risk Areas on Canada's Pacific Coast. Geography Department, University of Victoria.

Mavroulidou, M., S. J. Hughes and E. E. Hellawell (2004). "A qualitative tool combining an interaction matrix and a GIS to map vulnerability to traffic induced air pollution." Journal of Environmental Management **70**(4): 283-289.

McKendry, I. G. (2002). "Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM10 and PM2.5) Forecasting." Journal of Air and Waste Management Association **52**: 1096-1101.

Meyn, S. (2006). Air Quality Meteorologist. Victoria, BC, Air & Waste Management Association Pacific Northwest International Section**:** PNWIS 2006 Annual Conference.

Miller, K. A., D. S. Siscovick, L. Sheppard, K. Shepherd, J. H. Sullivan, G. L. Anderson and J. D. Kaufman (2007). "Long-term exposure to air pollution and incidence of cardiovascular events in women." New England Journal of Medicine **356**(5): 447-458.

Ministry of Environment (2007). Reducing Wood Stove Smoke: A Burning Issue Government of British Columbia July 25, 2007 http://www.env.gov.bc.ca/air/particulates/rwssabi.html

Moore, G. S. (2002). Assessing Human Risk. Living with the Earth: Concepts in Environmental Health Science, Lewis Publishers**:** 497-518.

Moral, F. J., P. Alvarez and J. L. Canito (2006). "Mapping and hazard assessment of atmospheric pollution in a medium sized urban area using the Rasch model and geostatistics techniques." Atmospheric Environment **40**(8): 1408-1418.

Naeher, L. P., M. Brauer, M. Lipsett, J. T. Zelikoff, C. D. Simpson, J. Q. Koenig and K. R. Smith (2007). "Woodsmoke health effects: A review." Inhalation Toxicology **19**(1): 67-106.

Natural Resources Canada (2002). Buying a High-Efficiency Wood-Burning Appliance Government of Canada August 28, 2007 http://www.canren.gc.ca/prod_serv/index.asp?CaId=126&PgId=719#Safety

Nelson, T. (2007). Professor of Quantitative Geography. Victoria.

Nelson, T., B. Boots and M. A. Wulder (2005). "Techniques for accuracy assessment of tree locations extracted from remotely sensed imagery." Journal of Environmental Management **74**(3): 265-271.

Nuckols, J. R., M. H. Ward and L. Jarup (2004). "Using geographic information systems for exposure assessment in environmental epidemiology studies." Environmental Health Perspectives **112**(9): 1007-1015.

O'Neill, M. S., M. Jerrett, L. Kawachi, J. L. Levy, A. J. Cohen, N. Gouveia, P. Wilkinson, T. Fletcher, L. Cifuentes and J. Schwartz (2003). "Health, wealth, and air pollution: Advancing theory and methods." Environmental Health Perspectives **111**(16): 1861-1870.

O'Sullivan, D. and D. J. Unwin (2003). Geographic Information Analysis. New Jersey, John Wiley & Sons, Inc.

Paustenbach, D. J. (2002a). Chapter 1 Primer on Human and Ecological Risk Assessment. Human and Ecological Risk Assessment: Theory and Practice. D. J. Paustenbach, John Wiley & Sons, Inc.

Paustenbach, D. J. (2002b). Chapter 4 Exposure Assessment. Human and Ecological Risk Assessment: Theory and Practice. D. J. Paustenbach, John Wiley & Sons, Inc.

Paustenbach, D. J., Ed. (2002c). Human and Ecological Risk Assessment: Theory and Practice, Wiley-Interscience.

Pierson, T. K., R. G. Hetes and D. F. Naugle (1991). "Risk Characterization Framework for Noncancer End-Points." Environmental Health Perspectives **95**: 121-129.

Reif, J. S., J. B. Burch, J. R. Nuckols, L. Metzger, D. Ellington and W. K. Anger (2003). "Neurobehavioral effects of exposure to trichloroethylene through a municipal water supply." Environmental Research **93**(3): 248-258.

Ricketts, T. C. (2003). "Geographic Information Systems and Public Health." Annual Review of Public Health(24): 1-6.

Ross, Z., P. B. English, R. Scalf, R. Gunier, S. Smorodinsky, S. Wall and M. Jerrett (2006). "Nitrogen dioxide prediction in Southern California using land use

regression modeling: potential for environmental health analyses." Journal of Exposure Science and Environmental Epidemiology **16**(2): 106-114.

Rossi, R. E., D. J. Mulla, A. G. Journel and E. H. Franz (1992). "Geostatistical Tools for Modelling and Interpreting Ecological Spatial Dependence." Ecological Monographs **62**(2): 277-314.

Schwartz, J., F. Laden and A. Zanobetti (2002). "The Concentration-Response Relation between PM2.5 and Daily Deaths." Environmental Health Perspectives **110**(10): 1025-1029.

Scoggins, A., T. Kjellstrom, G. Fisher, J. Connor and N. Gimson (2004). "Spatial analysis of annual air pollution exposure and mortality." Science of the Total Environment **321**(1-3): 71-85.

Setton, E., P. Hystad and C. P. Keller (2005). "Opportunities for using spatial property assessment data in air pollution exposure assessments." International Journal of Health Geographics **4**(1): 26.

Shaddick, G. and J. Wakefield (2002). "Modelling daily multivariate pollutant data at multiple sites." Journal of the Royal Statistical Society Series C-Applied Statistics **51**: 351-372.

Slini, T., K. Karatzas and N. Moussiopoulos (2003). "Correlation of air pollution and meteorological data using neural networks." International Journal of Environment and Pollution **20**(1-6): 218-229.

Tian, Y. Q., J. D. Radke, P. Gong and Q. Yu (2004). "Model development for spatial variation of PM2.5 emissions from residential wood burning." Atmospheric Environment **38**(6): 833-843.

Wheeler, D. C. and C. A. Calder (2007). "An assessment of coefficient accuracy in linear regression models with spatially varying coefficients." Journal of Geographical Systems **9**(2): 145-166.

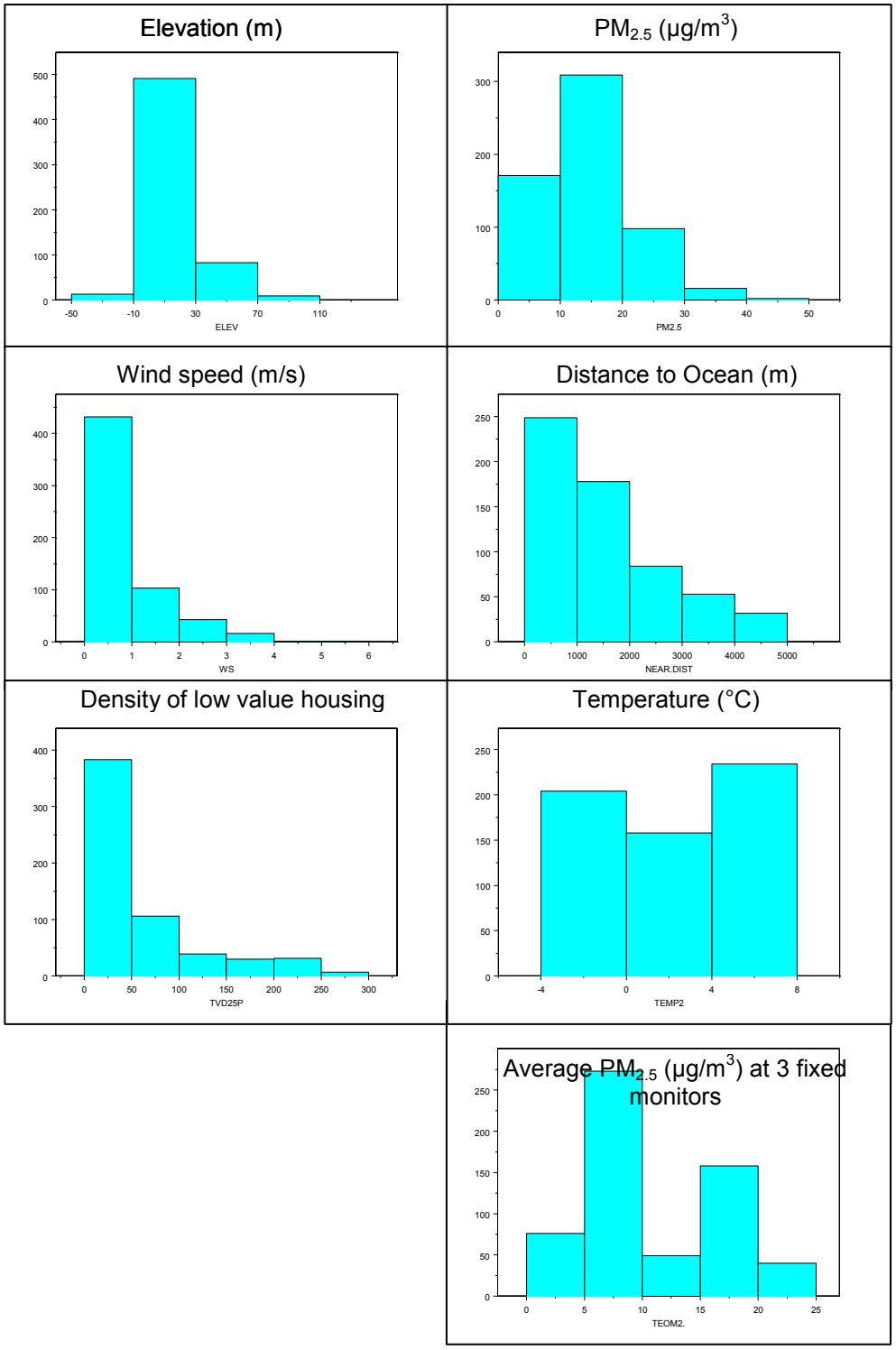
Wiens, J. A. (1989). "Spatial Scaling in Ecology." Functional Ecology **3**(4): 385-397.

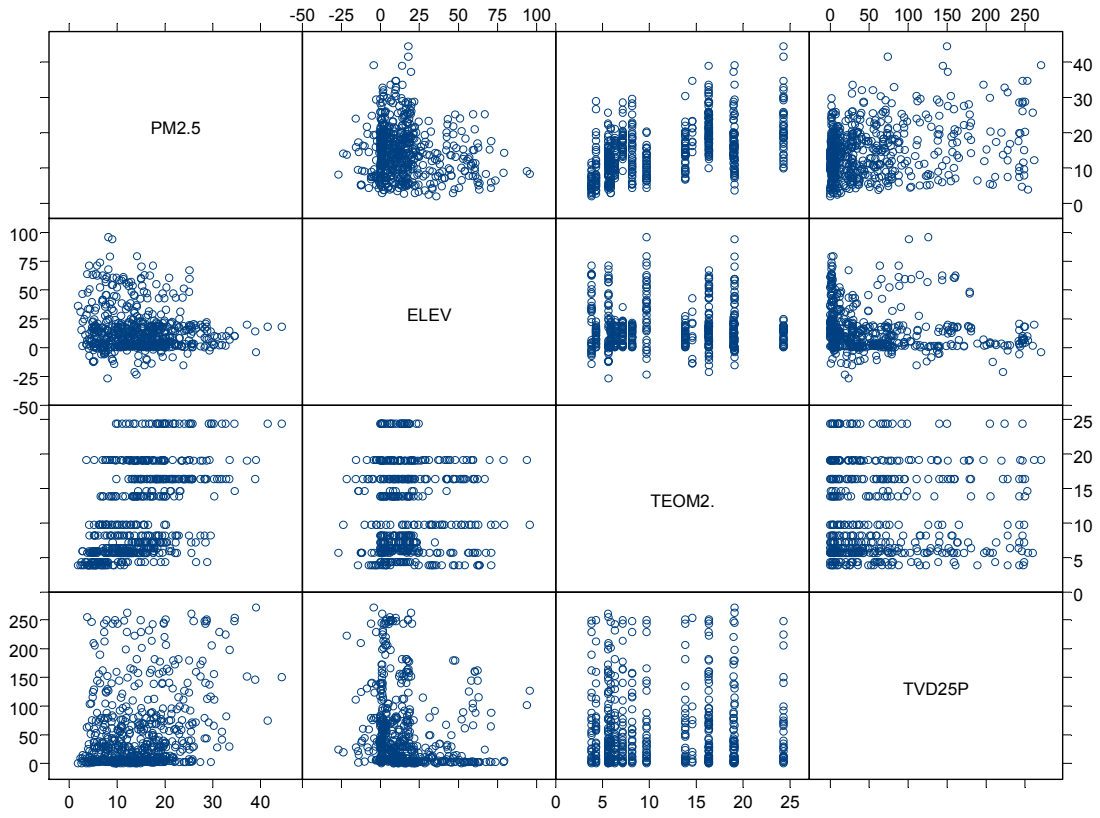
Williams, P. R. D. and D. J. Paustenbach (2002). Chapter 5 Risk Characterization. Human and Ecological Risk Assessment: Theory and Practice. D. J. Paustenbach, John Wiley & Sons, Inc.

Zelikoff, J. T., L. C. Chen, M. D. Cohen and R. B. Schlesinger (2002). "The toxicology of inhaled woodsmoke." Journal of Toxicology and Environmental Health-Part B-Critical Reviews **5**(3): 269-282.

Appendix A:

Model variable distributions





Matrix scatter plot of $PM_{2.5}$ (in $\mu g/m^3$), the dependent model variable, and the independent model variables

Appendix B:
Bootstrap for Resampling M1 Dataset 1000 Times for a Distribution of R^2

```
mystatistic <- function(m1){ #this takes M1 as input, will usually have 449
    fit <- lm(PM2.5~ELEV+TEOM2.+TVD25P, data=m1)
    sumfit <- summary (fit) #calls the summary
    c(coef(fit), summary(fit)$r.squared)}#produces a table of model coefficients
    and r.squared in mc.results$replicates

mcresults <- bootstrap(m1, mystatistic, B=1000, group=Grid500,
    sampler.args=list(size=1), assign.frame1 = T, save.indices = T, save.group =
    T,
    group.order.matters = F)

m1mcresults <- mcresults$replicates #open m1modelresults in object explorer
m1indices <- mcresults$indices #saves indices, shows the row numbers that were
    selected for each iteration, use this to ensure it sampled properly
m1group <- mcresults$group

qqnorm (mcresults)

plot (mcresults) #there is a problem with weights

#plot(sumfit) or c
#summary(mcresults) 1000 rows of results are in mcresults$replicates
```

Appendix C:
Bayesian Model Specification for WinBUGS

```
Model
{

        for (i in 1: n)
{
        pm[i] ~ dnorm(mu[i], tau.model)

        mu[i] <-  alpha0 + beta1 * (TEOM2.[i] – TEOM.bar) +
        beta2 * (ELEV[i] – ELEV.bar) + beta3 * (TVD25P[i] – TVD25P.bar) +
        beta4 * (WS[i] – WS.bar) + beta5 * (TEMP2[i] – TEMP2.bar) +
        beta6 * (NEAR.DIST[i] – NEAR.DIST.bar)
}

#set priors ----------------------------------------------------

alpha0 ~ dnorm(0.0, 1.0E-3)
beta1 ~ dnorm(0.0, 1.0E-3)
beta 2 ~ dnorm(0.0, 1.0E-3)
beta 3 ~ dnorm(0.0, 1.0E-3)
beta 4 ~ dnorm(0.0, 1.0E-3)
beta 5 ~ dnorm(0.0, 1.0E-3)
beta 6 ~ dnorm(0.0, 1.0E-3)

sigma.model ~ dunif(0, 10)
tau.model <- 1/pow(sigma.model,2)

#compute constants ----------------------------------------------

TEOM2.bar       <- mean(TEOM2[])
ELEV.bar        <- mean(ELEV[])
TVD25P.bar      <- mean(TVD25P[])
WS.bar          <- mean(WS[])
TEMP2.bar       <- mean(TEMP2[])
NEAR.DIST.bar   <- mean(NEAR.DIST[])

}
```

Appendix D:
OLS Models for Individual Sample Evenings

Date	(Intercept)	ELEV	WS	TEMP	TVD25P	NEAR.DIST	R-squared	Average windspeed (m/s)	Average temperature (°C)	Below average windspeed	Full route
January 14, 2007	37.4871	-0.5428	*8.2703 *-	*0.6475	0.1152	0.0033	0.3	0.17	-2	yes	No
February 7, 2005	11.2386	0.14	0.0065	1.2019	0.0473	*0.0003	0.3	0.52	0.12	yes	Yes
December 8, 2005	13.8882	0.0154	-2.1102	0.517	0.025	0.0011	0.19	3.2	0.19	no	Yes
November 30, 2006	13.4477	-0.0634	2.1319 *-	-1.1091	0.067	*-0.0003	0.5	0.26	0.3	yes	Yes
February 9, 2005	17.5149	-0.1993	1.1487	*0.6833	0.0566	0.0028	0.45	0.41	0.37	yes	Yes
December 14, 2005	15.3221	-0.0448	3.7458	-2.4195	0.0434	-0.001	0.24	0.32	0.52	yes	No
March 12, 2006	13.4964	-0.0293	*6.2206	0.5177	0.0004	*0.0006	0.07	0.02	0.7	yes	No
January 12, 2005	3.9642	-0.1591	6.4155	1.2703	0.0541	0.0029	0.50	0.58	1.16	yes	Yes
January 15, 2007	9.096	0.0146	-0.4279	0.0455	0.0198	-0.0007	0.17	0.75	1.3	no	No
December 13, 2005	34.7679	-0.3886	-6.8746	*0.6348	0.0276	-0.0025	0.51	0.55	1.6	yes	No
December 28, 2006	14.0665	0.0307	*-0.044	-2.4637	*-0.0041	-0.0006	0.6	1.56	1.7	no	Yes
January 20, 2006	15.0459	-0.0973	-1.8493 *-	*0.2789	0.0003	0.0026	0.39	1.09	2.22	no	No
February 8, 2006	13.7961	-0.1765	2.6295	0.9422	0.0313	0.0011	0.4	0.15	2.65	yes	No
December 16, 2006	8.9345	*-0.028	-2.4761 *-	*1.3203	*-0.0002	0.0037	0.13	0.83	2.9	no	No
February 20, 2006	11.2042	0.0104	0.2751	*0.7585	0.0176	-0.0005	0.27	0.41	3.53	yes	No
December 10, 2005	27.6035	0.0132	0.8781	-2.2723	0.0168	0.0009	0.48	0.99	4.05	no	No
January 16, 2005	10.2791	-0.2603	1.2434 *-	0.5493	0.0089	*-0.0001	0.24	1.55	4.2	no	Yes
February 5, 2005	-12.5809	0.279 *-	1.3706	3.8003 *-	0.0223	0.002	0.27	0.2	4.4	yes	Yes
November 10, 2006	7.5939	0.0125	*0.4951	0.3751	*-0.0002	*-0.0001	0.03	0.42	4.5	yes	No
December 26, 2006	16.1258	-0.0387	-0.1915	-2.1378	0.0129	*-0.0001	0.48	0.87	4.57	no	Yes
March 14, 2007	11.059	*0.0042	*0.4213	-1.2092	0.0001	*-0.0002	0.24	0.67	4.9	yes	No
March 14, 2006	9.0556	-0.046	0.9398	-0.5156	0.0091	0.0001	0.18	0.24	5.8	yes	No
March 20, 2006	8.0919	*0.0082	-3.0599	*0.2559	0.0289	0.0019	0.28	0.24	5.9	yes	No

Date	(Intercept)	ELEV	WS	TEMP	TVD25P	NEAR.DIST	R-squared	Average windspeed (m/s)	Average temperature (°C)	Below average windspeed	Full route
February 22, 2006	28.3221	*-0.0025	0.225	-3.526	*-0.0001	-0.0005	0.28	1.5	6	no	No
January 18, 2006	57.8224	-0.194	*0.0533	-6.7576	0.0397	*0.0015	0.35	0.11	6.5	yes	No
February 4, 2007	12.3405	-0.0118	-0.7085	0.0038	-0.0051	-0.0003	0.05	0.76	6.6	no	Yes
February 6, 2007	*4.0189	3.608	-1.4682	*0.6429	0.0091	0.0011	0.18	0.93	6.6	no	Yes
January 26, 2005	16.5772	*-0.0484	*6.2354	*0.4714	0.0314	0.0013	0.22	0.02	6.8	yes	Yes
January 27, 2005	*6.311	*0.0259	-0.5778	1.7748	*0.0013	-0.001	0.12	0.47	6.9	yes	Yes
January 31, 2006	12.031	*-0.0001	*0.0301	*-0.6063	0.0039	*-0.0001	0.04	4.12	8.66	no	No
January 20, 2005	*-1.106	*0.0137	*0.8616	1.0224	0.0085	0.0006	0.04	0.18	9.42	yes	Yes
February 8, 2005	15.701	-0.130	-0.614	3.535	0.001	0.025	0.34	0.71	0.48	yes	Yes
Average**	15.63	0.075	-0.27	-0.34	0.026	0.0019	0.28	0.78	3.55		

l PM2.5 ~ ELEV + WS + TEMP + TVD25P + NEAR.DIST
*Indicates variable not significant
**Does not include insignificant variables