Population-Level Ambient Pollution Exposure Proxies

A Thesis Submitted to the Committee on Graduate Studies
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in the Faculty of Arts and Science

TRENT UNIVERSITY
Peterborough, Ontario, Canada

Applied Modelling and Quantitative Methods (AMOD) MSc Graduate Program
May 2019

Abstract

Population-Level Ambient Pollution Exposure Proxies

Carlone Scott

The Air Health Trend Indicator (AHTI) is a joint Health Canada / Environment and Climate Change Canada initiative that seeks to model the Canadian national population health risk due to acute exposure to ambient air pollution. The common model in the field uses averages of local ambient air pollution monitors to produce a population-level exposure proxy variable. This method is applied to ozone, nitrogen dioxide, particulate matter, and other similar air pollutants. We examine the representative nature of these proxy averages on a large-scale Canadian data set, representing hundreds of monitors and dozens of city-level populations. The careful determination of temporal and spatial correlations between the disparate monitors allows for more precise estimation of population-level exposure, taking inspiration from the land-use regression models commonly used in geography. We conclude this work with an examination of the risk estimation differences between the original, simplistic population exposure metric and our new, revised metric.

Keywords: *Time Series, Spatial Process, Temporal Process, Spatio-Temporal, Nugget effect, Air Pollution, Population Health Risk*

Acknowledgements

I would like to thank:

- God for giving me mental strength throughout this work.
- My beloved parents Claudette and Livingston for their emotional and mental support. My siblings, in particular Ceraphia and Oshane, for their encouragement; Naudia, for her unwavering much needed support for which I will always be grateful.
- My supervisor Dr. Wesley Burr for imparting his statistical knowledge that sparked my love for statistics, his push that led me to improve my programming skills through this thesis.
- Professors Dr. Kenzu Andela, Dr. Wenying Feng and Dr. Stefan Bilaniuk for their guidance and willingness to always help where ever possible.
- My friend Sherine for her help in proof reading this work.

Table of Contents

Appendix B:

List of Figures

List of Tables

1. *Introduction*

Acute and long term health effects that air quality have on human morbidity and mortality have been studied by many researchers [3, 11, 21, 31]. These health effects are modelled from reported data for pollutants based on some metric which may be a twenty-four or eight-hour mean, or some other daily metric. Furthermore the standard population air pollution “‘risk” model also uses these averaged local ambient air pollution monitors to produce a population-level exposure proxy variable, and often applied to ozone, nitrogen dioxide, particulate matter, and other similar air pollutants. We examine the representative nature of the proxy averages and compare with variations of other averaging methods.

The paper makes use of Toronto data to examine the representative nature of the proxy using monitors and city-level populations. Our model will use observed data from this data set for the pollutants ozone, nitrogen dioxide, and particulate matter. The thesis is set out as follows: we begin with an introduction and an examination of some previous work. We then present background information that is required for our model development. The data set being used is then discussed along with an overview of how monitoring occurs for Canadian air pollution records. The space-time model employed and the theory behind it is then presented, along with the different metrics used in practice for air pollution. We then present the results of our modelling followed by a discussion of these results and their application to risk metrics. This work then concludes with findings and a scope of future work to be done.

2. *Background*

We begin this chapter by examining previous work, we then provide some information on the chemistry of the pollutants we are studying. We then review some basic definitions to aid with clarity of the paper. It is key to have an understanding of Temporal and Spatial Processes in order to fully understand the Spatio-Temporal models used in this project. This chapter is then closed by presenting these fundamentals for Spatial and Temporal Processes.

2.1 Previous Work

The following review of literature presents associated work that have been done in predicting pollutant concentrations. It will be noted that significant work has been done applying land use regression (LUR) models in particular for prediction of pollutant concentrations.

Henderson *et al.* [15] applied the method of Land Use Regression (LUR) to estimate long-term concentrations of traffic-related Nitrogen Oxides and Fine Particulate Matter. The authors used integrated 14-day mean concentrations that were measured with passive samplers at 116 sites in Vancouver, British Columbia for spring and fall 2003. They estimated annual mean NO, and NO_2 concentrations; the range for these estimates achieved greater spatial variability than the reported regulatory range for the region. They also investigated concentrations using LUR for fine particulate matter ($PM_{2.5}$) that was measured at a subset of 25 sites. The authors used some fifty-five variables that described each sampling site, with these variables generated in a Geographic Information System (GIS). The authors' use of LUR presents a good technique for predicting ambient air pollution concentrations. These predictions however were a yearly predicted concentration for each of the pollutants of study.

Ramos *et al.* performed two studies on this topic. In the first [25] they developed a hybrid interpolation technique that combined the inverse distance-weighted (IDW) method with Kriging with external drift (KED), and applied it to daily $PM_{2.5}$ levels observed at 10 monitoring stations. The authors achieved a down-scaled high-resolution map for $PM_{2.5}$ for the Island of Montreal. For the KED interpolation, the authors used spatio-temporal daily meteorological estimates and spatial covariates as land use and vegetation density. They developed different KED and IDW daily estimation models for the year 2010 for each of the six synoptic weather classes. They developed the clusters using principal component analysis and unsupervised hierarchical classification. They found that the performance of the hybrid model was better than that of the KED or the IDW alone.

In their second study [26] they looked at spatial modelling of daily concentrations of ground level ozone (O_3) in Montreal for the year 2010. The authors assessed the kriging with external drift (KED) model to estimate O_3 concentrations by synoptic weather classes for 2010. They compared these results with ordinary kriging (OK), and a simple average of 12 monitoring stations. They also compared the estimates obtained for the 2010 summer with those from a Bayesian maximum entropy (BME) model. They found that the KED model with road and vegetation density as covariates showed good performance for all six synoptic classes. They did note that future work needed to be done to integrate the temporal dependency in the data in order to not overstate the performance of the KED model.

Sabaliauskas *et al.* [29] applied land-use regression (LUR) to characterise the spatial distribution of ultra-fine particles (UFP) in Toronto. They used measured Particle Number (PN) concentrations from residential areas around Toronto, Canada, between June and August 2008. They used combination of fixed and mobile monitoring to assess spatial gradients between and within communities. Their mobile data included average PN concentrations collected on 112 road segments from 10 study routes that were repeated on three separate days. They used the mobile data to create the land-use regression model while the fixed sites were used for validation purposes. The authors found that the predictor variables that best described the spatial variation of PN concentration included population density, total resource and industrial area within, total residential area, and major roadway and highway length. They found that the LUR model successfully predicted the afternoon peak PN concentration but over-predicted the 24-h average PN concentration.

Weichenthal *et al.* [32] applied a LUR model for characterising the spatial distribution of ambient ultra-fine particles (UFPs). The authors developed a land use regression model for ambient UFPs in Toronto, Canada using mobile monitoring data collected during summer/winter 2010–2011. They included 405 road segments in the analysis. Their final model explained 67% of the spatial variation in mean UFPs and included terms for the logarithm of distances to highways, major roads, the central business district, Pearson airport, and bus routes as well as variables for the number of on-street trees, parks, open space, and the length of bus routes within a 100 m buffer. They found that there was no systematic difference between measured and predicted values when the model was evaluated in an external data set. They developed this model to be used in the evaluation of the chronic health effects of UFPs using population-based cohorts in the Toronto area.

Though there have been multiple studies associated with the prediction of pollutant concentrations, we have noted that the main concentration of previous work has been done employing land use regression models. These studies mainly accomplished these predictions at the yearly level, and some of the studies did not fully employ the temporal dependency. We will therefore set out in this study to employ the temporal dependency and predict a daily mean for pollutant concentrations.

2.2 The Pollutants

As stated in the introduction, we are concerned with acute and long term health effects that air quality has on morbidity and mortality. It is critical that we have an understanding of the structure and chemistry of the pollutants we are studying. We are focusing on three pollutants, *nitrogen dioxide* (NO_2), *ozone* (chemically known as trioxide, O_3) and *particulate matter* ($PM_{2.5}$).

Nitrogen Dioxide (chemical formula NO_2) in its natural form is a pungent gas with a light brown appearance that is produced both naturally and artificially. Nitrogen Dioxide is a natural gas in the atmosphere, produced from the stratosphere, from volcanoes, lightning and also from bacterial respiration. This naturally occurring gas is critical in the process of absorbing sunlight and regulating the chemistry of the troposphere. It plays a significant role in determining ozone concentrations. Artificially NO_2 is a produced as a by-product from power plants, industrial processes, cigarette smoking, butane and also from fertilisers. Chronic exposure to NO_2 can cause respiratory effects in humans and can also worsen respiratory illness in people already suffering from respiratory related illnesses such as asthma. Health Canada recommends [6] a residential maximum exposure short term limit of 170 $\mu g/m^3$ (90 ppb) NO_2 and long term limit of 20 $\mu g/m^3$ (11 ppb). The critical effects are decreased lung function and increased airway responsiveness in asthmatics.

In 2014, the annual average concentrations of NO_2 in the air varied from 4.4 ppb in Atlantic Canada to 10.5 ppb in British Columbia. Concentrations of NO_2 were lower in Atlantic Canada, southern Quebec and the Prairies and northern Ontario region than in 2013. However, concentrations were higher in southern Ontario and British Columbia in 2014 compared to the previous year. Since 2000, decreasing trends in NO_2 concentrations were observed for all regions in Canada. Southern Ontario had the largest decreasing trend at 0.7 ppb per year, followed by southern Quebec and British Columbia with 0.5 ppb per year and 0.4 ppb per year, respectively. Atlantic Canada had a decreasing trend of 0.3 ppb per year, while the Prairies and northern Ontario region had a decreasing trend of 0.2 ppb per year [6].

Ozone or trioxide (chemical formula O_3) is an inorganic molecule. It presents as a pale blue gas with a distinctively pungent smell. Ozone is formed naturally from dioxygen (O_2) by the action of ultraviolet light and also atmospheric electrical discharges, and is present in very low concentrations throughout the Earth's atmosphere. Ozone is also produced artificially by the burning of fossil fuels, methane, and from ozone generators used to produce ozone for cleaning air or removing smoke odours in unoccupied rooms. Ozone in the upper atmosphere (10 to 50 kilometres above the earth's surface) protects the earth from the sun's harmful ultraviolet radiation. In the lower atmosphere and at ground level, ozone is harmful to human health. It can cause breathing problems, reduce lung function and aggravate asthma and other lung diseases. Ozone is not directly emitted by anthropogenic sources, but is formed in the lower atmosphere when precursor gases such as nitrogen oxides (NO_x) and volatile organic compounds ($VOCs$) react in sunlight. Ozone (ground-level) can damage vegetation and it is one of the leading causes for summertime smog. Ground-level ozone can harm lung function and irritate the respiratory system. Exposure to ozone (and the pollutants that produce it) is linked to premature death, asthma, bronchitis, heart attack, and other cardiopulmonary problems [13]. Health Canada recommends a residential maximum exposure limit of 40 $\mu g/m^3$ (20 ppb) ozone, based on an averaging time of 8-hours [4]. Above this limit there is a decrease in pulmonary function and increases in subjective respiratory symptoms.

In 2014, the national annual average concentration of ground-level O_3 was 32.9 parts per billion (ppb), or 0.1% higher than in 2013. The annual peak (4th-highest) 8-hour O_3 concentration in 2014 was 53.4 ppb, or 3.5% lower than in 2013. Although the annual peak concentration of O_3 was frequently above the 2015 Standard before 2008, it has consistently been below for the last seven years. In Canada, there are two national ozone (O_3) indicators:

1. An annual average concentration indicator that is based on the annual average concentrations (of the daily maximum 8-hour averages); it is used to capture prolonged or repeated exposures over longer periods or chronic exposure.

2. An annual peak (4^{th}-highest) 8-hour indicator that is based on the annual 4^{th}-highest daily maximum 8-hour average concentrations; it is used to capture immediate or acute short-term exposure.

The peak O_3 indicator is calculated using an approach that is aligned with the 2015 Canadian Ambient Air Quality Standards (the standards) [4].

Particulate Matter ($PM_{2.5}$) or fine particulate matter is a general term for all small particles found in air measuring equal to or less than 2.5 micrometres in aerodynamic diameter. It is a complex mixture with its constituents varying in shape, size, surface area, density, and chemical composition. Indoor $PM_{2.5}$ is composed of sources such as smoking, cooking, cleaning and from external $PM_{2.5}$ sources such as traffic, heating (in winter), industrial processes, among others, that have infiltrated from outside. Average indoor $PM_{2.5}$ concentrations in different Canadian cities were less than 15 $\mu g/m^3$ in homes without smokers and 35 $\mu g/m^3$ in homes with smokers [5]. In general, indoor $PM_{2.5}$ levels were lower than outdoor concentrations measured directly outside the home, except in homes with smokers. Studies have investigated the relationship between indoor $PM_{2.5}$ and health, presenting evidence for a relationship between indoor $PM_{2.5}$ levels and declines in lung function and increases in exhaled nitric oxide, a marker of airway inflammation in asthmatic children [10, 17, 18]. Changes in exhaled nitric oxide were however more strongly associated with outdoor $PM_{2.5}$ than indoor $PM_{2.5}$. Other studies showed that the associations between indoor $PM_{2.5}$ and subtle changes in markers of cardiovascular disease have also been observed in older adults [21, 1, 31].

Indoor levels of $PM_{2.5}$ should be kept as low as possible, as there is no apparent threshold for the health effects of $PM_{2.5}$. It is impossible to entirely eliminate $PM_{2.5}$ indoors, as among its sources are essential and everyday activities, such as cooking and cleaning mentioned above, as well as infiltration from outdoor sources. Any significant reduction in $PM_{2.5}$ would be expected to result in health benefits, especially for the elderly or children and other people with underlying health conditions. The Canadian Ambient Air Quality Standards has set objectives for outdoor air quality in Canada as 10 $\mu g/m^3$ (annual) and 28 $\mu g/m^3$ (24 hour) [12]. While Health Canada has not set a maximum limit for indoor $PM_{2.5}$ levels, they have advised that indoor levels be kept lower than outdoor levels.

2.2.1 Pollutant units of measurement

For clarity, we will explain the units of measurements used for each pollutant. As mentioned above NO_2 and O_3 are reported in ppb (parts per billion), while $PM_{2.5}$ is reported in $\mu g/m^3$ (micro-grams per cubic meter). Furthermore PM sizes are measured in microns (micrometers), so for $PM_{2.5}$ the reported measurement unit is associated with particulate matter of size 2.5 microns. The unit $\mu g/m^3$ is generally known as *mass per unit volume*, while ppb is generally known as *volume mixing ratio*.

In *mass per unit volume*, the mass of pollutant is expressed as a ratio to the volume of air. Since the volume of a given packet of air is dependent upon the temperature and pressure at the time of sampling, the pollutant concentration expressed in these units is dependent on the conditions at the time of sampling. In *volume mixing ratio*, the unit expresses the concentration of a pollutant as the ratio of its volume if segregated pure (no air), to the volume of the air in which it is contained. Ideal gas (a hypothetical gas whose molecules occupy negligible space and have no interactions, and that consequently obey the gas laws exactly) [16] behaviour is assumed and thus the concentration is not dependent upon temperature and pressure as these affect both the pollutant and the air to the same extent. As a consequence of the gas laws [30], a gas present at a volume mixing ratio of 1 ppb is not only $1cm^3$ per 10^{-9} cm^3 of polluted air, it is also 1 molecule per 10^{-9} molecules and has a partial pressure of one billionth of the atmospheric pressure.

2.2.1.1 Unit conversions

It is useful to know the unit conversion between $\mu g/m^3$ and ppb, so we give a brief derivation of the conversion formula in this section. Under standard conditions (0 degrees Centigrade, 1013.25 hectopascals-hPa), one mole of an ideal gas occupies 22.414 litres(l). The mass of a pollutant p, M_p in grams (g) can therefore be converted to its equivalent molecular volume V_p in litres:

$$V_p = \frac{M_p}{MW} (22.414\ l)$$

where MW denotes the molecular weight of the pollutant (measured in g/mol). For measurements at pressure and temperature other than the standard conditions, corrections to the standard volume must be applied, based on the ideal gas law:

$$22.414\ l \left(\frac{T}{273.15\ K} \right) \left(\frac{1013.25\ hPa}{P} \right)$$

where T and P are the ambient temperature (measured in Kelvins) and pressure (measured in hectopascals-hPa) at the time of measurement, respectively. Therefore,

$$ppb = V_p/V_a$$

where V_a and V_p are the air and pollutant volume, respectively. Combining the equations gives the conversion formula:

$$ppb = \left(\frac{\frac{M_p}{MW} (22.414\ l) \left(\frac{T}{273.15\ K}\right) \left(\frac{1013.25\ hPa}{P}\right)}{V_a} \right) \left(1,000,000\ l/m^3\right).$$

2.3 Temporal Processes

The data of interest in this study is generated from observations over time (time series) in a given space. We need to investigate the change in time that occurs in the data of interest, being the observations at any given location ("constant location"). We are interested in the temporal processes associated with the observations, noting that these observations come with measurement error. We begin with some definitions that we will use throughout the project.

We denote $[A]$ and $[A|B]$ to represent the marginal and conditional probability distributions, respectively. Then the joint distribution of A and B can be written as

$$[A, B] = [A|B][B],$$

and the law of total probability can be written as

$$[A] = \int [A|B][B]dB,$$

where $\int g(B)\,[B]\,dB$ denotes the expectation of some function $g(B)$ of B. In terms of this notation, Bayes' Theorem can be written as

$$[B|A] = \frac{[A|B][B]}{\int [A|B][B]dB} = \frac{[A|B][B]}{[A]}.$$

The following definitions as set out in [8] are critical for the understanding of Spatio-Temporal Modelling.

Bayesian Hierarchical Modelling (BHM)

If we let Z represent the data, Y the hidden process that we need to predict and θ the unknown parameters, then the basic representation of a Bayesian Hierarchical Model is obtained by

$$\text{Data model:} \qquad [Z|Y,\theta]$$

$$\text{Process model:} \qquad [Y|\theta]$$

$$\text{Parameter model:} \qquad [\theta]$$

Note that sometimes $[Z|Y,\theta_D]$ and $[Y|\theta_P]$ is written to emphasise the data-model parameter θ_D and the process-model parameters θ_P. Then $\theta = \{\theta_D,\theta_P\}$, and the parameter model is $\{\theta_D,\theta_P\}$.

The joint distribution is therefore given as

$$[Z,Y,\theta] = [Z|Y,\theta][Y|\theta][\theta]$$

From Bayes' Theorem the conditional distribution of Y and θ, given the data Z (which is called the posterior distribution) is obtained as

$$[Y,\theta|Z] = \frac{[Z|Y,\theta][Y|\theta][\theta]}{[Z]},$$

where

$$[Z] = \int\int [Z|Y,\theta][Y|\theta][\theta]dYd\theta.$$

All inference on Y and θ in the BHM depends on this distribution (within the framework of Bayesian decision theory).

Empirical Hierarchical Modelling (EHM)

An EHM results from estimating the parameters directly from the data and "plugging" them back into the model. The representation of EHM is obtained by

$$\text{Data model:} \qquad [Z|Y,\theta]$$

$$\text{Process model:} \qquad [Y|\theta]$$

where it is assumed the parameter θ is fixed, but unknown. Formally a third level could still be considered, but where the parameter model $[\theta]$ concentrates all its probability at the fixed θ. We can write the data-model parameters as θ_D and the process-model parameters as θ_P by writing the two-level model as $[Z|Y, \theta_D]$, $[Y|\theta_P]$, and $\theta = \{\theta_D, \theta_P\}$. In an EHM, all probability distributions are conditional on θ. Inference on Y depends on the distribution

$$[Y, \theta|Z] = \frac{[Z|Y, \theta][Y|\theta]}{[Z|\theta]},$$

where $[Z|\theta] = \int [Z|Y, \theta][Y|\theta]dY$. The "Empirical" part of the EHM arises from replacing $[Z|Y, \theta]$ with $[Z|Y, \widehat{\theta}]$, where $\widehat{\theta}$ is an estimator of θ (that is, depends only on the data Z). It is also possible that θ is estimated from an independent study.

2.3.1 Characterisation of Temporal Processes

A temporal process can be written as $Y(\cdot)$ or more completely as

$$\mathbf{Y}(r) : r \in D_t$$

where r indexes the time of the possibly multivariate process $\mathbf{Y}(\cdot)$ and D_t is a subset of $\mathbb{R}^1$. The process $\mathbf{Y}$ may be deterministic or stochastic. The model is quite general if the possibility that the index set D_t can be a random set. For a *continuous-time process*, it is assumed that D_t is fixed and has nonzero length in the continuous interval $(-\infty, \infty)$, generally assuming that $D_t = [0, \infty)$. For a *discrete-time process* (time series), a fixed index set of finite or countable set of times, $D_t = \{0, \pm 1, \pm 2, ...\}$ is assumed; generally the set is limited to $D_t = \{0, 1, 2, ...\}$. A third type of process is a *temporal point process*, where D_t is assumed to be a random set made up of randomly occurring points (events) in $\mathbb{R}^1$. For example, the time of occurrence of a tornado in a given country could be represented as a Poisson point process in time.

For clarity, a continuous-time process is denoted $\mathbf{Y}(t)$ and a discrete-time process is denoted $\mathbf{Y}_t$.

When working with real-world temporal processes, typically we have only one realisation of that process and the associated time series can be viewed as just one sample from a *population.* This population is characterised by its *joint distribution.* As an example, consider a time series given by $\{Y_t : t = 0, ..., T\}$. Its joint distribution is denoted by

$$[Y_0, Y_1, ..., Y_T]$$

From our basic results earlier we can write

$$[Y_0, ..., Y_T] = [Y_T | Y_{T-1}, ..., Y_0][Y_{T-1} | Y_{T-2}, ..., Y_0]...[Y_1 | Y_0][Y_0]. \tag{2.1}$$

For practical modelling additional assumptions is needed about the components of Equation (2.1). Equation (2.1) can be assumed to be modelled by a first-order Markov property.

$$[Y_t | Y_{t-1}, ..., Y_0] = [Y_t | Y_{t-1}], \qquad \text{for all} \;\; t = 1, 2, ... \tag{2.2}$$

This property suggests a "lack of memory", so only the most recent past determines the conditional probabilities about the present given the whole past, so Equation (2.1) becomes

$$[Y_0, ..., Y_T] = [Y_0] \prod_{t=1}^{T} [Y_t | Y_{t-1}]. \tag{2.3}$$

A *continuous-time process* $Y(t)$ might be described by the simple differential equation in which the rate of change of the process Y with time is simply related to a function of the process at time t:

$$\frac{dY(t)}{dt} = f(Y(t)), \qquad t \geq 0, \tag{2.4}$$

where the function f may be linear or nonlinear in $Y(t)$. Thus given some *initial condition* $Y(0)$, the evolution of the process is completely determined by the function

f. If we are only limited to processes at discrete times then an analogous equation could be written as,

$$Y_t = \mathcal{M}(Y_{t-1}), \qquad t = 1, 2, ..., \tag{2.5}$$

where $\mathcal{M}$ maps the process from the previous time $t-1$ to the current time t. Again, given an initial condition Y_0, the process $\{Y_t : t = 0, 1, 2, ...\}$ is completely determined by the function $\mathcal{M}$.

For a *stochastic* or *random* processes, it can be said that the future is only partially determined from the past. If we added a random "noise" component to Equation (2.5):

$$Y_t = \mathcal{M}(Y_{t-1}) + \eta_t, \quad t = 1, 2, ..., \tag{2.6}$$

where $\{\eta_t : t = 1, 2, ...\}$ is a mean-zero random process and η_t is statistically independent of Y_{t-1}, then this random-noise term implies that the process $\{Y_t\}$ is random as well. The next section gives some fundamental principles on time series, which will assist in framing the background for this thesis.

2.3.2 Time Series Fundamentals

We begin by defining some basic functions of a discrete-time sequence of real-valued random variables $\{Y_t : t \in D\}$. Given $D_t = \{0, 1, ...\}$ and that $\{Y_t : t = 0, 1, ..., \}$ is a time series then the *mean function* is defined as

$$\mu_t \equiv E_t(Y_t), \qquad t \in D_t, \tag{2.7}$$

which is simply the mean of the process relative to the underlying probability space. A simple model for μ_t is normally chosen with the uncertainty placed in the remaining stochastic component, $\{Y_t - \mu_t : t = 0, 1, ...\}$. This uncertainty is described through the *autocovariance function*,

$$C_Y(t, r) = \text{cov}(Y_t, Y_r) = E\{(Y_t - \mu_t)(Y_r - \mu_r)\}, \qquad t, r \in D_t \tag{2.8}$$

which describes how the process co-varies across different time lags, after accounting for the mean function. Here $C_Y(t,r) = C_Y(r,t)$ and also the variance is a special case of the autocovariance where $C_Y(t,t) = \mathrm{var}(Y_t) = \sigma_t^2$. The auto-correlation function is then obtained after normalising, and measures the linear statistical dependence between different members of the time series. Formally,

$$\rho_Y(t,r) = \frac{C_Y(Y_t, Y_r)}{\sqrt{C_Y(t,t)C_Y(r,r)}}, \qquad t, r \in D_t. \tag{2.9}$$

where $\rho_Y(t,r) \in [-1,1]$.

A sufficient condition for Equations (2.8) and (2.9) to exist is $\sigma_t^2 < \infty$, for all D_t. A common assumption that is normally made for time series models is that of *stationarity.* There are different forms of stationarity for d-dimensional space, we are interested in the special case of one dimension (time).

Definition 1. *Strong stationarity of $\{Y_t\}$ is an assumption that says that any finite collection, $\{Y_{t1}, ..., Y_{tm}\}$, of random variables from the time series has the same joint distribution as $\{Y_{t1+\tau}, ..., Y_{tm+\tau}\}$, for any $\tau \in \{0, \pm 1, \pm 2, ...\}$. Weak stationarity of $\{Y_t\}$ is an assumption that requires initially the existence of the second moment (then the first moment automatically exists); that is, assume var$(Y_t) \equiv \sigma_t^2 < \infty$, for all $t \in D_t$. Time series whose variance are finite are said to be second-order stationary if*

(i) $E(Y_t) \equiv \mu$*, for all* $t \in D_t$.

(ii) $\mathrm{cov}(Y_t, Y_r) \equiv C_Y(t-r)$*, for all* $t, r \in D_t$.

Second-order stationarity is an assumption that says any pair, Y_t, Y_r, has exactly the same first and second moments (including cross-moments which defines the autocovariance function) as the pair $Y_t + \tau, Y_r + \tau$, for any $\tau \in \{0, \pm 1, ...\}$. For processes with finite variance, strong stationarity implies second-order stationarity, but not vice versa.

2.3.2.1 White-Noise Process

A *white-noise* process is defined as a discrete-time random process $\{W_t : ..., -1, 0, 1, ...\}$ whose elements are mutually independent and have a common probability density function. Typically, its mean μ_w is assumed to be zero, and its autocovariance function is

$$C(\tau) = \begin{cases} \sigma_w^2, & \tau = 0 \\ 0, & \tau = \pm 1, \pm 2, ..., \end{cases} \tag{2.10}$$

where $\sigma_w^2 > 0$ is the white-noise variance.

2.3.2.2 Random-Walk Process

A time series $\{Y_t\}$ is said to be a random walk if

$$Y_t = Y_{t-1} + W_t, \qquad t = 1, 2, ..., \tag{2.11}$$

where W_t is the white-noise process with mean μ_w and variance σ_w^2.

2.3.2.3 Autoregressive Process

In this study the observations of environmental processes depend on one or more observations that immediately proceed it. A time series that models this structure is the *autoregressive* (or AR) process. The model that is introduced for the modelling of our pollutants in Chapter 3 employs the AR process, hence a formal definition is presented here. A time series $\{Y_t\}$ is said to be an autoregressive process of order p, AR(p), if

$$Y_t = m_1 Y_{t-1} + m_2 Y_{t-2} + ... + m_p Y_{t-p} + W_t, \qquad t = ..., -2, -1, 0, 1, 2, ..., \tag{2.12}$$

where W_t is a white-noise process with mean zero and variance σ_w^2, and where $\{m_i : i = 1, ..., p\}$, are fixed but unknown parameters. Specifically we will employ the AR(1) process, that is

$$Y_t = m_1 Y_{t-1} \qquad t = ..., -2, -1, 0, 1, 2, ..., \tag{2.13}$$

where W_t is a white-noise process with mean zero and variance σ_w^2, and we write $\alpha = \alpha_1$ for the notional simplicity. By back-substitution, we obtain

$$\begin{aligned} Y_t &= W_t + \alpha W_{t-1} + \alpha^2 W_{t-2} + \alpha^3 W_{t-3} + ... \\ &= \sum_{k=0}^{\infty} \alpha^k W_{t-k} \end{aligned} \tag{2.14}$$

where it is assumed that $|\alpha < 1|$. That is, the AR(1) process can be written as an infinite series of white-noise random variables. Since $E(W_t) = 0$ and $var(W_t) = \sigma_w^2$, it follows that $E(Y_t) = 0$ and

$$\mathrm{var}(Y_t) = \sigma_w^2 \left(1 + \alpha^2 + \alpha^4 + ...\right) = \frac{\sigma_w^2}{1 - \alpha^2} \tag{2.15}$$

which does not depend on t.

2.4 Spatial Processes

This section follows closely the work of Cressie and Wikle [8]. A spatial process can be considered as a temporal snapshot, a temporal aggregation or a temporally frozen state of space-time process. We will present some definitions followed by some key concepts to facilitate our understanding of spatial processes.

Definition 2. *The **variogram** is defined as the variance of the difference between field values at two locations (*$\mathbf{s}_1$ *and* $\mathbf{s}_2$*) across realisations of the field [9]:*

$$2\gamma(\mathbf{s}_1, \mathbf{s}_2) = \mathrm{var}\left(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)\right) = E\left[\left((Z(\mathbf{s}_1) - \mu(\mathbf{s}_1)) - (Z(\mathbf{s}_2) - \mu(\mathbf{s}_2))\right)^2\right].$$

If the spatial random field has constant mean μ*, this is equivalent to the expectation for the squared increment of the values between locations* $\mathbf{s}_1$ *and* s_2 *(where* $\mathbf{s}_1$ *and* $\mathbf{s}_2$

are points in space and possibly time):

$$2\gamma(\mathbf{s}_1, \mathbf{s}_2) = E\left[(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2\right].$$

2.4.1 Data, Measurement and Error

Data as we know comes with error: there are the obvious errors associated with measuring, manipulating, archiving, there are however other errors associated with the discrete spatial and temporal sampling of an inherently continuous system. There are scales of variability that are unresolved and that will further "contaminate" the observations. Variability results from two important sources: one from the sources due to measurement of the phenomenon of interest, and the other due to the incomplete knowledge of the phenomenon. These are called the "data process" and "process model" respectively. The nugget effect is defined as being equal to the discontinuity of the variogram at the origin. This discontinuity can be made up of both measurement error and spatial dependence at scales smaller (microscales) than the available distances between observations.

It must be noted that care must be taken when applying kriging as some statistical software may fail to filter out these variability due to measurement error. The model that we will introduce later in this thesis employs algorithms that does not filter out the variability mentioned above. Consider an unobserved air quality measurement such as NO_2 at a known location along with observed Nitrogen Dioxide (NO_2) readings at known locations throughout our region of interest (Toronto). The data process (D) at the location of our region (s) is given by $\mathbf{D}_s \subset \mathbb{R}^d$. Then $Y(\cdot) \equiv \{Y(\mathbf{s}) : \mathbf{s} \in \mathbf{D}_s\}$ is defined to be the true NO_2 value at location $\mathbf{s}_o$. The observations are "noisy" versions of the true NO_2 reading at known locations $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$; the observations can be written as (assuming additive measurement error):

$$Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i) \qquad i = 1, ..., m$$

where, independently of $Y(\cdot)$, $\varepsilon(\cdot) \equiv \{\varepsilon(\mathbf{s}) : \mathbf{s} \in \mathbf{D}_s\}$ is a white-noise process with mean zero and variance $\sigma_\varepsilon^2 \geq 0$.

The optimal spatial predictor of $Y(\mathbf{s}_0)$ at known location $\mathbf{s}_0$, based on squared error loss and data,

$$\mathbf{Z} \equiv (Z(s_1), ..., Z(\mathbf{s}_m))',$$

is given as

$$E(Y(\mathbf{s}_0)|\mathbf{Z}).$$

2.4.2 The Nugget Effect explained by simple Spatial Hierarchical Model

It is assumed that $Y(\cdot)$ and $\varepsilon(\cdot)$ in $Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i)$, are independent Gaussian processes, and therefore only the first two moments are needed to characterise completely the probabilistic distributions. By further assuming second-order stationarity the following is defined:

$$C_Y(\mathbf{h}) \equiv \text{cov}(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})), \qquad \text{for all } \mathbf{s}, \mathbf{s}+\mathbf{h} \in D_s.$$

Clearly $\sigma_Y^2 \equiv C_Y(\mathbf{0})$, and

$$\lim_{h \to 0}(C_Y(\mathbf{0}) - C_Y(\mathbf{h})) \equiv C_Y(\mathbf{0}+) \equiv \sigma_0^2 \geq 0,$$

with $\sigma_0^2 \leq \sigma_Y^2$. Then σ_0^2 represents the variance of the microscale component of the process $Y(\cdot)$. Since $\text{var}(\varepsilon(\mathbf{s})) = \sigma_\varepsilon^2$,

$$\mathbf{cov}(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})) \equiv C_Z(\mathbf{h}) = \begin{cases} \sigma_Y^2 + \sigma_\varepsilon^2, & \mathbf{h}{=}\mathbf{0} \\ C_Y(\mathbf{h}), & \mathbf{h}{\neq}\mathbf{0} \end{cases}$$

which implies that

$$\lim_{h \to 0}(C_Z(\mathbf{0}) - C_Z(\mathbf{h})) \equiv C_Z(\mathbf{0}+) = \sigma_0^2 + \sigma_\varepsilon^2 \equiv c_0 \geq 0$$

This quantity, c_0, is the "nugget effect" and is made up of two non-negative components: σ_0^2 (the microscale variance of $Y(\cdot)$) and σ_ε^2 (the measurement-error variance of $Z(\cdot)$).

The geostatistical model on $D_s \subset \mathbb{R}^d$ described above can be written hierarchically as:

Data model: Conditional on σ_ε^2, and for $i = 1, ..., m$,

$$Z(\mathbf{s}_i)|Y((\mathbf{s}_i), \sigma_\varepsilon^2 \sim iid\ Gaussian(Y(\mathbf{s}_i), \sigma_\varepsilon^2).$$

Process model: Conditional on μ and $C_Y, Y(\cdot)$ is a stationary Gaussian process with mean μ and covariance function $C_Y(\mathbf{h}); \mathbf{h} \in \mathbb{R}^d$.

For the optimal spatial prediction mentioned earlier in this section geostatistics uses an EHM approach, estimating unknown parameters of the HM (such as, σ_ε^2, μ, C_Y) from the empirical variogram.

The nugget effect is important in the sense of environmental modelling such as the pollutant modelling that we set out to accomplish in this work. If we were to ignore the nugget effect in our model accuracy would be affected as a result of the errors mentioned above that are associated with σ_ε^2 and σ_0^2 being ignored. This would result in inaccuracies in our predictions of the true pollutant values $(Y(\cdot))$.

3. *Spatio-temporal modelling*

3.1 The model

We look at the framework of the model in this chapter along with additional derivations that help to shape that framework. In this section we set out to explain the model [2] used in the project. The pollutants being modelled as stated in the background reading are: Ozone (O_3), Nitrogen Dioxide (NO_2) and Particulate Matter ($PM_{2.5}$). For observations of the pollutants of study we employ the following distribution:

$$y_{it} \sim \text{Normal}\left(\eta_{it}, \sigma_e^2\right), \tag{3.1}$$

Where y_{it} denotes the logarithm of the pollutant concentration measured at site $\mathbf{s}_i$ $(i = 1, ..., n)$ and day $t = 1, ..., T$, σ_e^2 the variance of the measurement error defined by a Gaussian white-noise process, both spatially and serially uncorrelated (that is there is no correlation (similarity) between observations at different sites and days), and

$$\eta_{it} = b_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \omega_{it}, \tag{3.2}$$

where b_0 is the intercept and $\beta_1, ..., \beta_M$ are the mixed effects related to meteorological and orographical covariates $x_1, ..., x_m$. ω_{it} refers to the latent or inferred spatio-temporal process (that is the true unobserved level of pollution) which changes in time with first-order autoregressive dynamics and spatially correlated innovations:

$$\omega_{it} = a\omega_{i(t-1)} + \xi_{it}, \tag{3.3}$$

with $t = 2, ..., T$, $|a| < 1$ and $\omega_{i1} \sim \text{Normal}\left(0, \frac{\sigma^2}{(1-a^2)}\right)$. ξ is a zero-mean Gaussian field, assumed to be temporally independent and characterised by the following spatio-temporal covariance function:

$$\text{Cov}\left(\xi_{it}, \xi_{ju}\right) = \begin{cases} 0, & \text{if t} \neq \text{u} \\ \text{Cov}\left(\xi_i, \xi_j\right), & \text{if t} = \text{u} \end{cases}, \tag{3.4}$$

for $i \neq j$, where $\text{Cov}\left(\xi_{it}, \xi_{ju}\right)$ is given by the Matèrn spatial covariance function:

$$\text{Cov}\left(\xi(\mathbf{s}_i), \xi(\mathbf{s}_j)\right) = \text{Cov}\left(\xi_i, \xi_j\right) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}} \left(\kappa \parallel \mathbf{s}_i - \mathbf{s}_j \parallel\right)^{\lambda} K_{\lambda}\left(\kappa \parallel \mathbf{s}_i - \mathbf{s}_j \parallel\right). \tag{3.5}$$

Equation 3.4 is separable and can be written as the product of a purely spatial and purely temporal covariance function.

3.2 Stochastic Partial Differential Equation (SPDE) approach

In this section we set out to explain the SPDE methodology employed in the INLA model used in Chapter 5. Rue and Tjelmeland, see [28], proposed approximating a continuous field using a Gaussian Markov Random Field (GMRF). We can have continuous random fields that are Markov, which present the case when the continuous field is a solution of a linear stochastic partial differential equation (SPDE). The SPDE approach used in INLA consists of representing a continuous spatial process (a Gaussian field) using a discretely indexed spatial random process (a GMRF). The starting point is the linear fractional SPDE:

$$\left(\kappa^2 - \Delta\right)^{\frac{\alpha}{2}} \left(\tau\xi(\mathbf{s})\right) = \mathcal{W}(\mathbf{s}), \tag{3.6}$$

where $\mathbf{s} \in \mathbb{R}^d$, Δ is the Laplacian, α controls the smoothness, $\kappa > 0$ is the scale parameter, τ controls the variance, and $\mathcal{W}(\mathbf{s})$ is the Gaussian spatial white noise process. The exact stationary solution to this SPDE is the stationary GF $\xi(\mathbf{s})$ given by

$$\operatorname{Cov}\left(\xi(\mathbf{s}_i), \xi(\mathbf{s}_j)\right) = \operatorname{Cov}\left(\xi_i, \xi_j\right) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}} \left(\kappa \parallel \mathbf{s}_i - \mathbf{s}_j \parallel\right)^{\lambda} K_{\lambda}\left(\kappa \parallel \mathbf{s}_i - \mathbf{s}_j \parallel\right), \quad (3.7)$$

where $\parallel \mathbf{s}_i - \mathbf{s}_j \parallel$ is the Euclidean distance between two geometric locations, $\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^d$, σ^2 is the marginal variance, and K_λ denotes a modified Bessel function [27] of the second kind, of order $\lambda > 0$

$$K_{\lambda}\left(x\right) = \int_0^{\infty} e^{-x \cosh t} \cosh \ \lambda t \ dt \qquad \lambda > 0 \quad (3.8)$$

Note the Bessel function is used as there is a need to find separable solutions for Laplace's equation. The order $\lambda > 0$ measures the degree of smoothness of the process and is usually kept fixed. $\kappa > 0$ on the converse is a scaling parameter related to the range r, the distance at which the spatial correlation becomes almost null.

The link between the SPDE in Eq. (3.5) and the Matèrn parameters is given by the following equations involving the smoothness parameter λ and the marginal variance σ^2:

$$\begin{cases} \lambda = \alpha - \frac{d}{2} \\ \sigma^2 = \frac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)^{\frac{d}{2}}\kappa^{2\lambda}\tau^2}. \end{cases}$$

For the case of $\mathbf{s} \in \mathbb{R}^2 \quad (d = 2)$, it follows that

$$\begin{cases} \lambda = \alpha - 1 \\ \sigma^2 = \frac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)\kappa^{2\lambda}\tau^2}. \end{cases}$$

For the INLA model used in this project the default value for the smoothness of the parameter is $\alpha = 2$ (corresponding to $\lambda = 1$). For $\alpha = 2$ the range r and the variance σ^2 are given by

$$r = \frac{\sqrt{8}}{\kappa} \quad (3.9)$$

$$\sigma^2 = \frac{1}{(4\pi)\kappa^2\tau^2} \tag{3.10}$$

The solution to the SPDE represented by the stationary and isotropic Matèrn Gaussian field $\xi(s)$, can be approximated using the finite element method (FEM) [20] through a basis function representation defined on a triangulation (see Section 3.3) of the domain $\mathcal{D}$.

$$\xi\left(\mathbf{s}\right) = \sum_{g=1}^{G} \varphi_g\left(\mathbf{s}\right)\tilde{\xi}_g. \tag{3.11}$$

Here

G is the total number of vertices of the triangulation,

$\{\varphi_g\}$ is the set of (deterministic) basis functions, and

$\tilde{\xi}_g$ are zero mean Gaussian distributed weights

In order to obtain a Markov structure, the basis functions are chosen to have a local support and to be piece-wise linear in each triangle, i.e., φ_g is 1 at vertex g and 0 at all other vertices. Using Neumann boundary conditions [22], it follows that (for the case $\alpha = 2$) the precision matrix Q for the Gaussian weight vector $\widetilde{\xi} = \{\widetilde{\xi}_1, ..., \widetilde{\xi}_G\}$ is given by

$$\mathbf{Q} = \tau^2\left(\kappa^4\mathbf{C} + 2\kappa^2\mathbf{G} + \mathbf{G}\mathbf{C}^{-1}\mathbf{G}\right) \tag{3.12}$$

where the generic element of the diagonal matrix $\mathbf{C}$ is

$\mathbf{C}_{ii} = \int \varphi_i(\mathbf{s})ds$

and the one of the sparse matrix $\mathbf{G}$ is

$\mathbf{G}_{ii} = \int \nabla\varphi_i(\mathbf{s})\nabla\varphi_j(\mathbf{s})ds$ (where ∇ denotes the gradient)

The precision matrix $\mathbf{Q}$, whose elements depend on τ and κ, is sparse and consequently ξ is a **GMRF** with distribution Normal$(\mathbf{0}, \mathbf{Q}^{-1})$ and represents the approximated solution to the SPDE. It is worth noting that the default internal representation for the SPDE parameters in the R-INLA package that was used for the project is $log(\tau) = \theta_1$ and $log(\kappa) = \theta_2$ where θ_1 and θ_2 are given a joint Normal prior distribution (by default independent).

3.2.1 Defining the projector (observation) matrix

The response mean (η_i) that correspond with the linear predictor from normally distributed observations ($y_i \sim \text{Normal}(\eta_i, \sigma_e^2)$) is defined as

$$\eta_i = b_0 + \xi_i, \tag{3.13}$$

where b_0 represents the intercept and ξ_i represents the random effect. This linear predictor can be represented by

$$\eta_i = b_0 + \sum_{g=1}^{G} A_{ig}\tilde{\xi}_g, \tag{3.14}$$

given the basis function representation of Equation (3.10). Here $\varphi_g(\mathbf{s}_i)$ is the gth basis function evaluated in $\mathbf{s}_i$, the linear predictor can be expressed as

$$\eta_i = b_0 + \sum_{g=1}^{G} \varphi_g(\mathbf{s}_i)\tilde{\xi}_g, \tag{3.15}$$

with $A_{ig} = \varphi_g(\mathbf{s}_i)$ being a common element of the sparse matrix $\mathbf{A}$ which maps the GRMF $\tilde{\xi}$ from the G triangulation vertices to the n observation locations.

3.3 Mesh pre-construction

The triangulation of the spatial domain mentioned in Section 3.2 is obtained by subdividing the spatial domain into a set of non-intersecting triangles, where any two triangles meet at most, a common edge or corner see Figure 3.1. There is a trade off between accuracy of the GMRF representation and computational costs, we will look at this trade off and look at a few meshes and discuss the mesh used for computation in the following.

3.3.1 Toronto mesh pre-construction

We now discuss the steps taken in constructing the mesh used for the Toronto study. The R-INLA package uses the function `inla.mesh.2d`. Among the arguments that `inla.mesh.2d` takes, we will look at the following: `loc, loc.domain, max.edge` and `offset.` The `loc` argument is a non-optional argument that gives the coordinates of the desired location of study. These coordinates may not necessarily include the coordinates of the location of the sites where the pollutants were observed. Another non-optional argument `max.edge` defines the required maximum edge for any given triangle in the triangulation, this may be a one or two argument assignment. The following code produces figure 3.1 which shows the mesh plot with `max.edge` defined to be 1000 *meters* (1 *kilometre*).

```
mesh_0 <- inla.mesh.2d(loc = t_coords, max.edge = 1000)
```

In the case where `max.edge` takes two arguments, the second sets the maximum edge for the triangles in the outer extension of the domain as seen in Figure 3.2. It can be seen that adjusting the second value of `max.edge` produces a change in the outer domain's triangles as seen in the comparison of the plots in Figure 3.2 which is produced by the following code. In `mesh_1` the spatial domain is extended to the outer. When we compare both plots in Figure 3.2 they would both yield equally accurate results for the inner bounds of the spatial domain. The top plot would produce more accurate results for the outer bound as opposed to the bottom plot. The latter plot would however be favourable as it accomplishes the task of reducing or eliminating boundary effects whilst not compromising too much on computing costs.

```
mesh_1 <- inla.mesh.2d(loc = t_coords, max.edge = c(1000, 1000))
mesh_2 <- inla.mesh.2d(loc = t_coords, max.edge = c(1000, 2000))
```

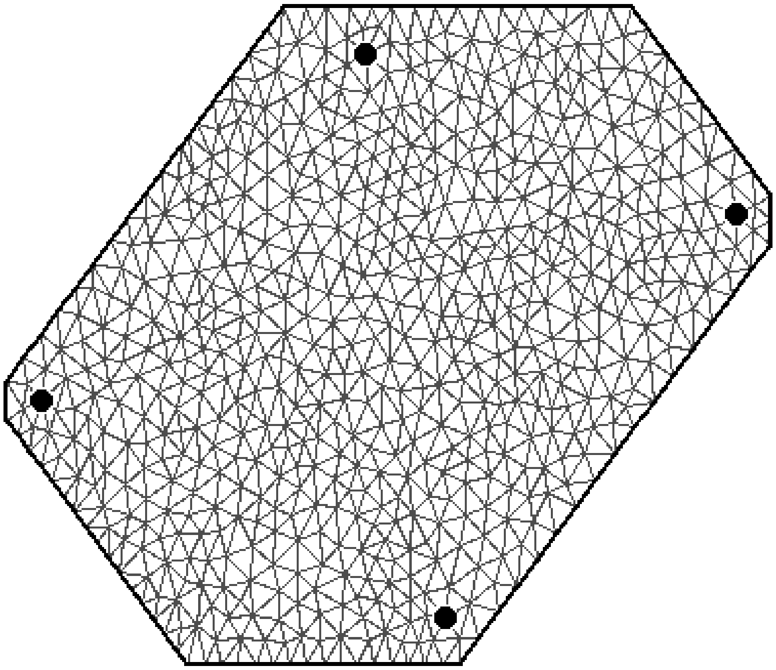


Figure 3.1: Triangulation for Toronto mesh, with maximum triangle edge defined as 1000 *meters*

Now let us examine the `loc.domain` arguments. Our previous plots assumed that our area of study was bounded by the coordinates supplied, which means predictions will only be calculated for the domain in the inner and outer sections which may not necessarily include the desired complete domain as is the case with Toronto. Since we are interested in the entire Toronto area and the coordinates (which in this case are the pollutant monitoring stations), our previous mesh will not suffice as some of the spatial domain is located outside this prescribed domain. To fix this we supply `loc.domain,` the intended domain, which in this case is the Toronto city borders. The following code produces Figure 3.3. It can be seen that once the Toronto border is specified, our spatial domain now extends to include the entire Toronto area with an outer domain specified to deal with the problem of boundary effect.

```
mesh_3 <- inla.mesh.2d(loc = t_cords, max.edge = c(1000, 2000),
```

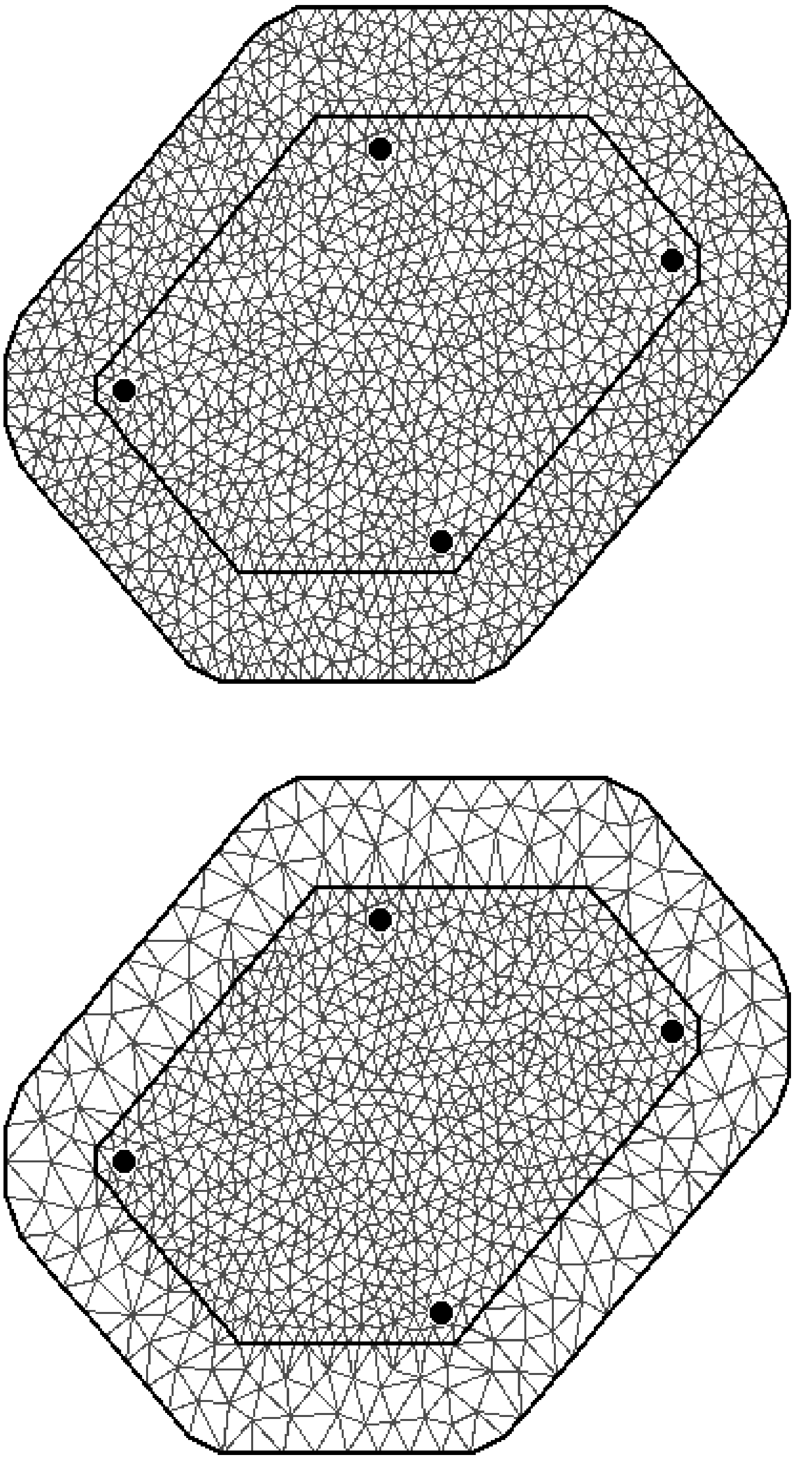


Figure 3.2: Outer triangle bounds defined, first mesh (top) inner and outer max edge similar, and second mesh (bottom) outer max edge twice inner max edge

```
                    loc.domain = TO_border)
```

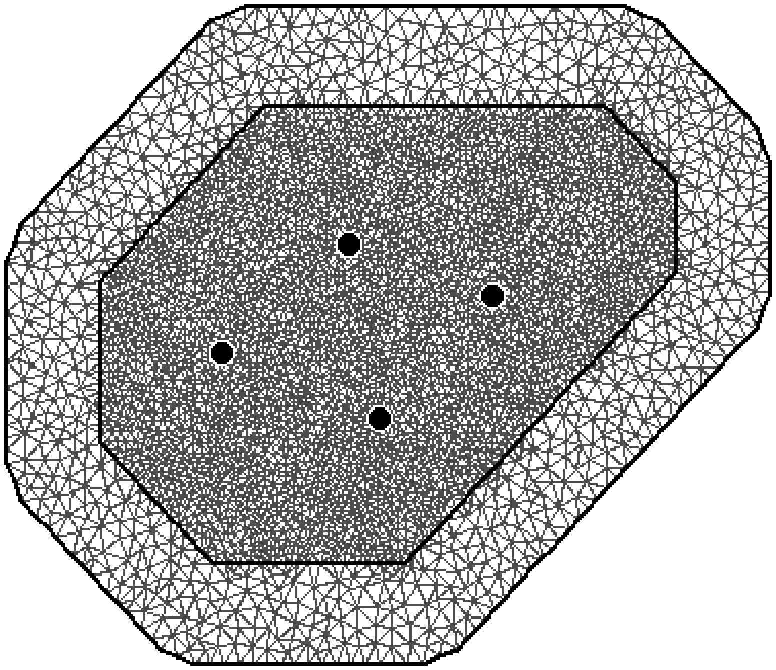


Figure 3.3: Triangulation of (third mesh) spatial domain with Toronto border defined

The `offset,` an optional argument, defines how much domain should be extended. We examine this by looking at `mesh_4`, and `mesh_5` with different `offset` as defined in the code included. Note the top plot in Figure 3.4 `offset` is defined with a larger offset for the outer, while in the bottom plot it is defined with a smaller offset for the inner. Like `max.edge`, `offset` can take one or two arguments. If one argument is supplied, it is automatically, by default, defined as the distance to which the domain is extended in the outer. For two arguments passed to `offset`, they are defined as: firstly, the distance to which the domain is extended in the outer, and secondly, as the distance to which the domain is extended in the inner. Here we have supplied two arguments to `offset` for the extent to which the domain is extended. We have also made minor adjustments to `max.edge.` For demonstration purposes, we will revisit this in the Section 3.3.2.

```
mesh_4 <- inla.mesh.2d(loc = t_cords, max.edge = c(1500, 6000),
                     loc.domain = TO_border, offset = c(600, 18000))
mesh_5 <- inla.mesh.2d(loc = t_cords, max.edge = c(1500, 6000),
                       loc.domain = TO_border, offset = c(1000, 1500))
```

3.3.2 Toronto mesh and grid

In an ideal world where computational costs do not exist and time is not a constraint, `mesh_5` as constructed in Section 3.3, would be ideal for extremely precise calculations/predictions. Since we do not exist in the ideal world we will adjust our mesh construction to reflect this reality. This process is a meticulous task and can be a tedious one if not carefully thought out and executed. Before we can utilise a constructed mesh for $INLA$ calculations we require a grid on which our predictions are projected; this is required because we are working with data that is associated with area and we have to take into account the spatial dependency through the neighbourhood structure.

From the background, our study observations tend to be more similar the closer they are spatially, because they are influenced by similar conditions. Given $1, ..., n$ where n denotes the number of locations at which the observations are recorded, then for the given area i, its neighbour $\mathcal{N}(i)$ are defined as the areas which share its borders. A first-order neighbour is a neighbour that shares the immediate border with area i, a second-order neighbour shares its immediate border with a first-order neighbour of area i, see [2], and Figure 3.5 demonstrates this concept graphically. Given the set of neighbours $\mathcal{N}(i)$ and using the local Markov property that the parameter θ_i for the ith area is independent of all other parameters, then the `makegrid` function is used to construct the desired grid for Toronto. This function takes a spatial object. This spatial object was retrieved from the City of Toronto [23] as a shapefile and

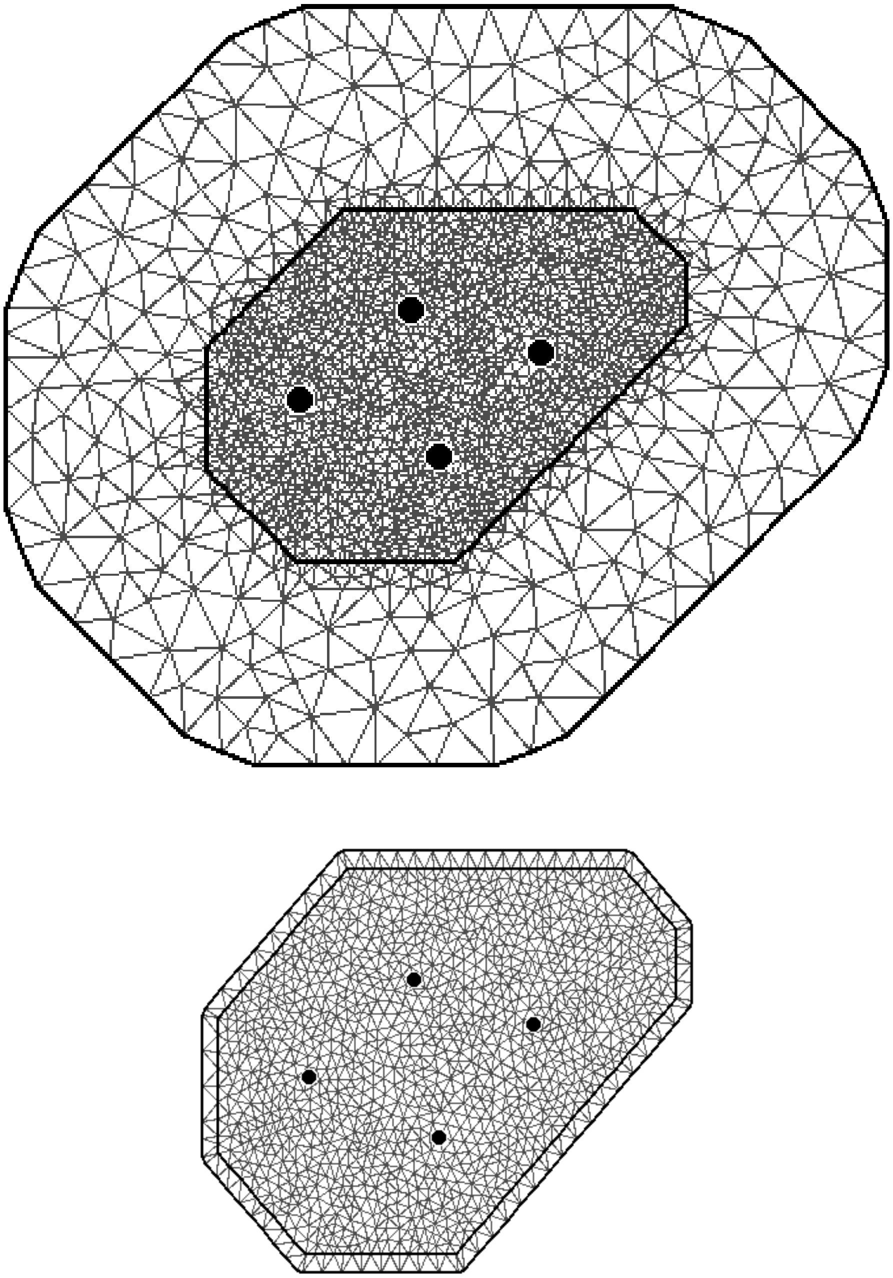


Figure 3.4: Triangulation for Toronto pollutant stations with offset defined for comparison, fourth mesh (top) offset defined larger than fifth mesh (bottom)

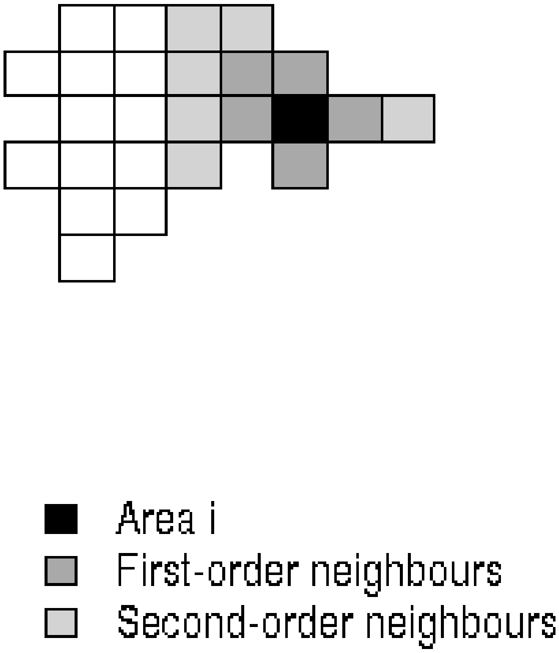


Figure 3.5: Neighbouring structure demonstrating first-order and second-order neighbours

converted for use in `R`. The grid size that was used consisted of 500 *metre* grids. This high resolution grid was chosen because we wanted to capture points in every *census tract area* (we will define this in Chapter 4), our desired resolution being 1000 *metre* to 2000 *metre* grids. As mentioned earlier, higher `max.edge` defined in `inla.mesh.2d` will result in less accuracy. Given that the grid used had a higher resolution than desired, increasing the value of `max.edge` within reason does not compromise our results, while reducing computing costs. The triangulation of the spatial domain was adjusted until a balance of accuracy and speed was attained.

On average it would take forty seconds to run a month's calculation with R-INLA (or four hours per year per pollutant), which resulted in a total run-time of approximately 134 hours (for all three pollutants for for the eleven year period). This was improved by using cloud computing and parallel programming to cut processing times by half. Using the original mesh constructed in Section 3.3 would result in an even slower run time and crashes as was experienced during the testing and adjusting stages of mesh construction.

The final mesh, Figure 3.6, that was employed given the constraints was sufficient for this study. The offset was extended beyond the "ideal" mesh boundaries to increase accuracy lost due to increase in triangle size. The Toronto border has also been included in the plot in Figure 3.6 to give a clearer picture of the final mesh construction and relative dimensions.

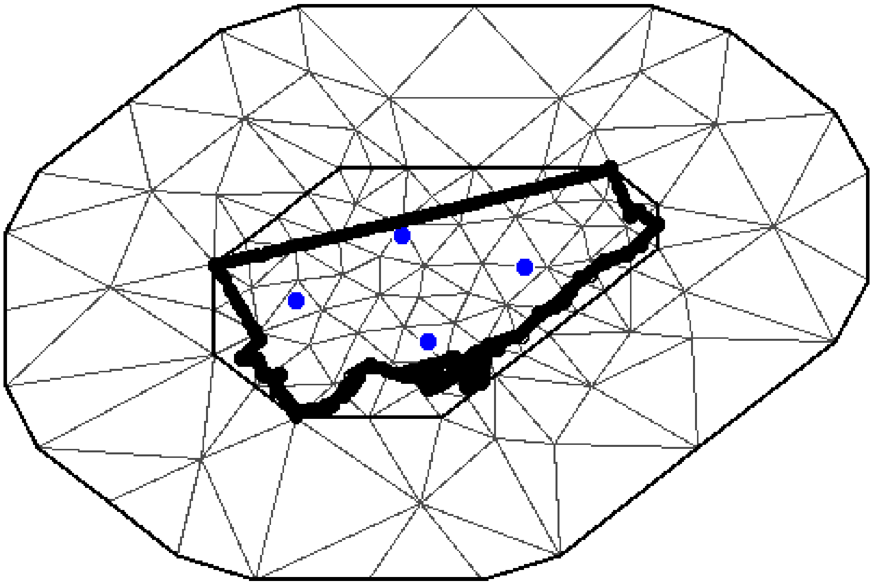


Figure 3.6: Mesh used for Toronto study (adjusted to balance computational costs and accuracy)

4. *Data Pre-processing and Metrics*

4.1 Data Retrieval

The data set used in this study is drawn from monitoring done through Environment Canada's National Air Pollution Surveillance (NAPS) program between 1982 and 2015. The program was established in 1969 to monitor and assess the quality of ambient (outdoor) air in the populated regions of Canada. There are currently 286 sites in 203 communities in every province and territory. These pollutant stations measure across these sites in part: sulphur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3) and particulate matter (PM, both PM_{10} and $PM_{2.5}$). Figure 4.1 shows the NAPS network with more stations in populated areas and less or none in underpopulated areas. These stations fluctuate in operation over the period 1982 - 2015, which means station data is not continuous for the period. This off-line and on-line nature of the sites vary for different reasons: some are retired due to population decreasing, while others are brought on-line as population increases; other factors that affect the fluctuation of stations include maintenance and accessibility issues.

The monitoring stations are clustered in Census Divisions (CD). One division of keen interest is the Toronto Census Division (CD3520) given that this area has the most population in Canada. The Toronto CD has a total of 14 stations (Figure 4.2). The Toronto census area has an approximate area of 630.2 km^2 which is home for

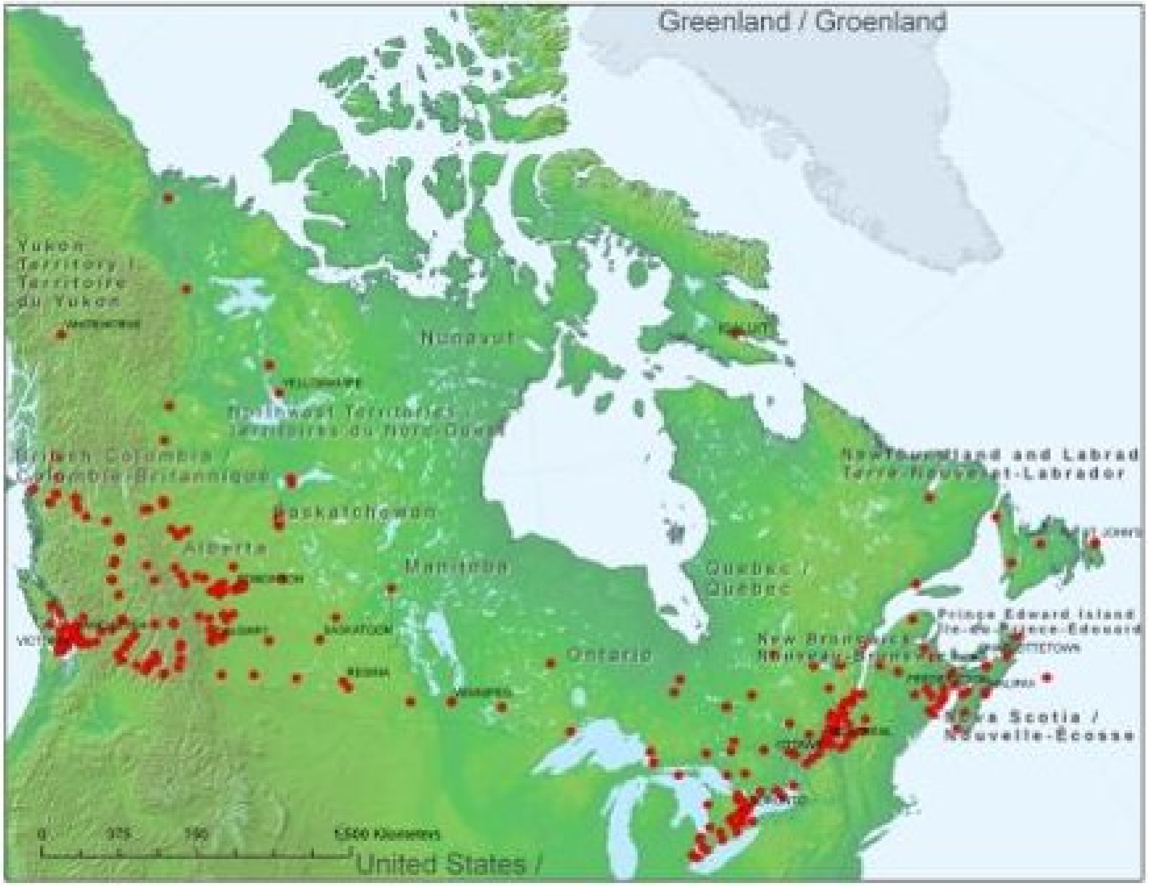

Greenland / Groenland
Yukon Territory / Territoire du Yukon
WHITEHORSE
Nunavut
IQALUIT
YELLOWKNIFE
Northwest Territories / Territoires du Nord-Ouest
British Columbia / Colombie-Britannique
Saskatchewan
Alberta
EDMONTON
Manitoba
Quebec / Québec
Newfoundland and Labrad
Terre-Neuve-et-Labrador
ST. JOHN'S
Prince Edward Island / Île-du-Prince-Édouard
CHARLOTTETOWN
New Brunswick / Nouveau-Brunswick
Nova Scotia / Nouvelle-Écosse
Ontario
VICTORIA
CALGARY
SASKATOON
REGINA
WINNIPEG
OTTAWA
MONTREAL
TORONTO
0 375 750 1,500 Kilometres
United States /

Figure 4.1: National Air Pollution Surveillance Program (NAPS) Network

approximately 2.81 million Canadians. As with other stations across the country the Toronto CD stations' active monitoring fluctuates with as few as 6 stations during the 1974 - 1984 period to as many as ten stations during the 2007 - 2016 period. This fact presents an additional problem in modelling the effects of air pollution, because not only do the sites change but also the demographics, as people move into and out of the Toronto census division.

The data set of study for this thesis was retrieved from Environment Canada's National Air Pollution Surveillance Program (NAPS). This data is presented in a hourly format for each site and each pollutant is reported independently in separate text files. The retrieval process is best described in the following steps:

1. The data was downloaded into R (statistical programming language) in its raw

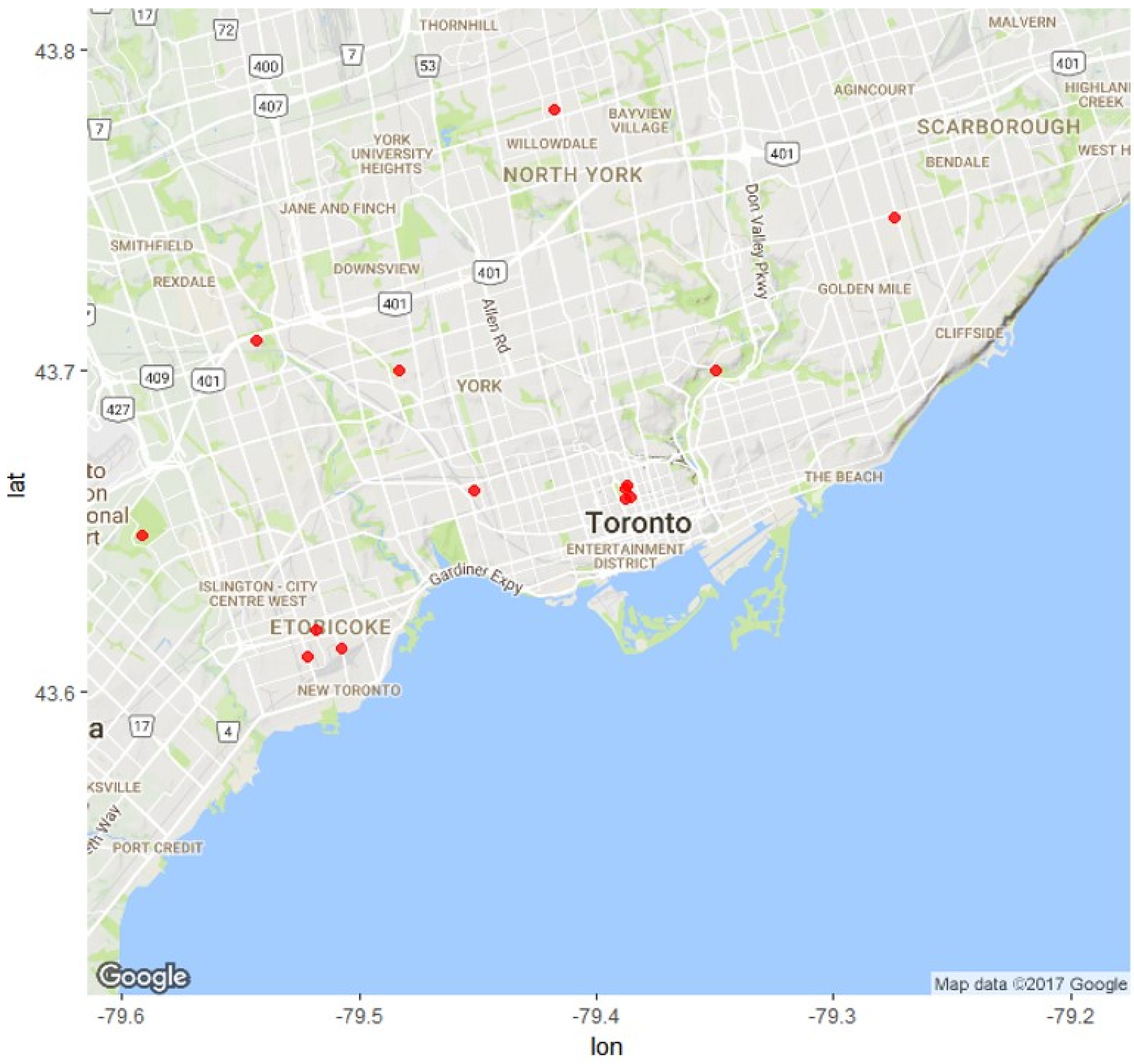


Figure 4.2: Toronto NAPS monitoring sites

form

2. The data was then pre-processed to account for missing values
3. It was then reordered using various R packages with the primary one being the *dplyr* package
4. We then used various averages (which will be discussed further below) to compress the hourly observations to daily readings.

4.2 Data Processing

In addition to the data retrieval methods employed we had to do an extensive amount of data processing to clean the raw data. This raw data was the complete Canada data set for all monitoring stations for the time period 1974 - 2016. Cleaning this included resolving issues such as missing data and corrupt data. The data was sorted by city (Toronto), by station and over the required period. The period for the data used in the study was cut to include only data for 2005 - 2016. This was done as the data was too sparse before this period that it would not provide any additional value.

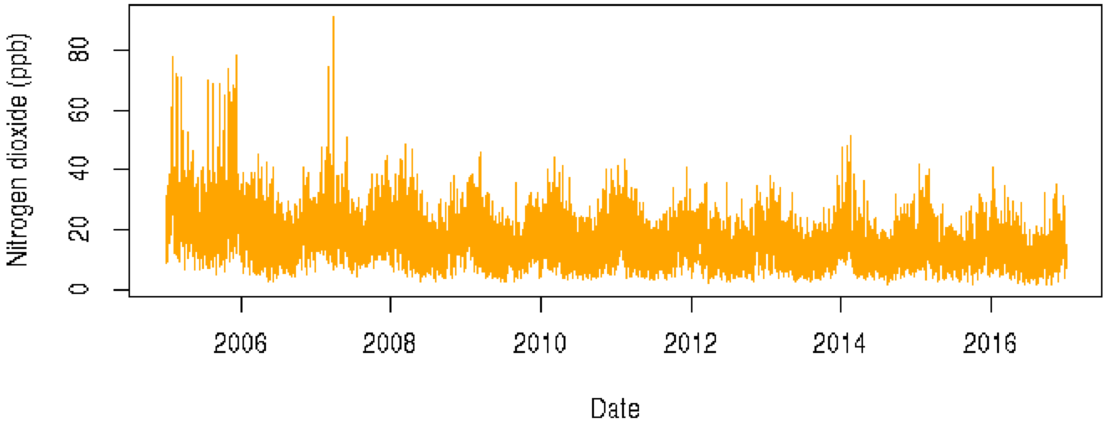


Figure 4.3: Toronto observed NO_2 time series 2005 - 2016

We present simple time series plots for the pollutants in Figures 4.3, 4.4, and 4.5 for pollutants NO_2, O_3, and $PM_{2.5}$ respectively. These time series plots are not free of measurement errors and are not the true representation of the Toronto city-level concentration levels. The following chapter will investigate the spatio-temporal factors discussed in the previous chapter to arrive at a more accurate "time series". It is also interesting to see how observed pollutant concentrations vary by stations.

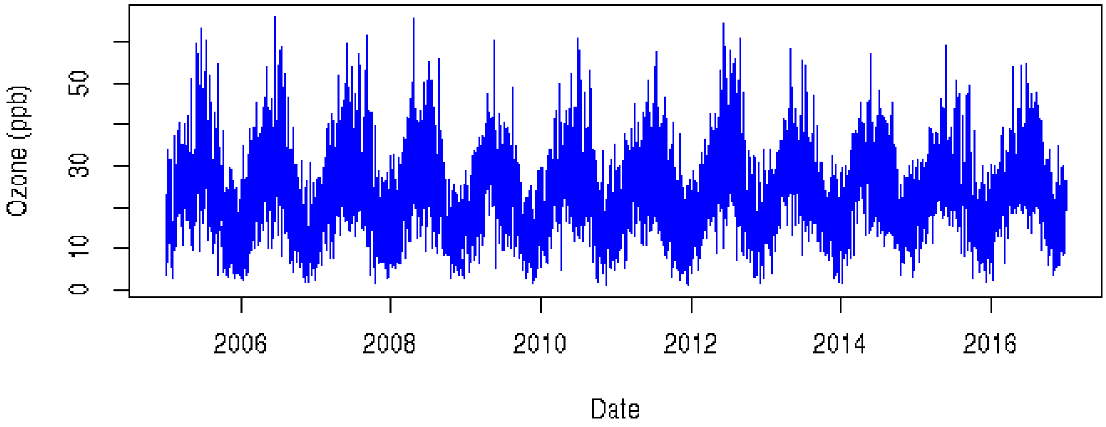


Figure 4.4: Toronto observed O_3 time series 2005 - 2016

In Figure 4.6 we present boxplots for each pollutant over the period 2005 - 2016 by stations. It can be seen that there is not a huge variance between stations. Stations that were closer to major highways however had elevated concentrations. We will examine these trends in the next chapters.

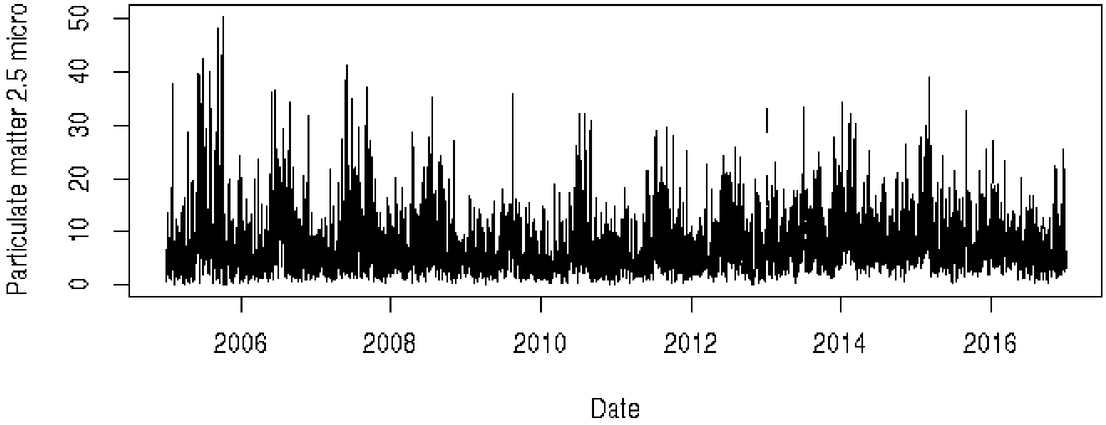


Figure 4.5: Toronto observed $PM_{2.5}$ time series 2005 - 2016

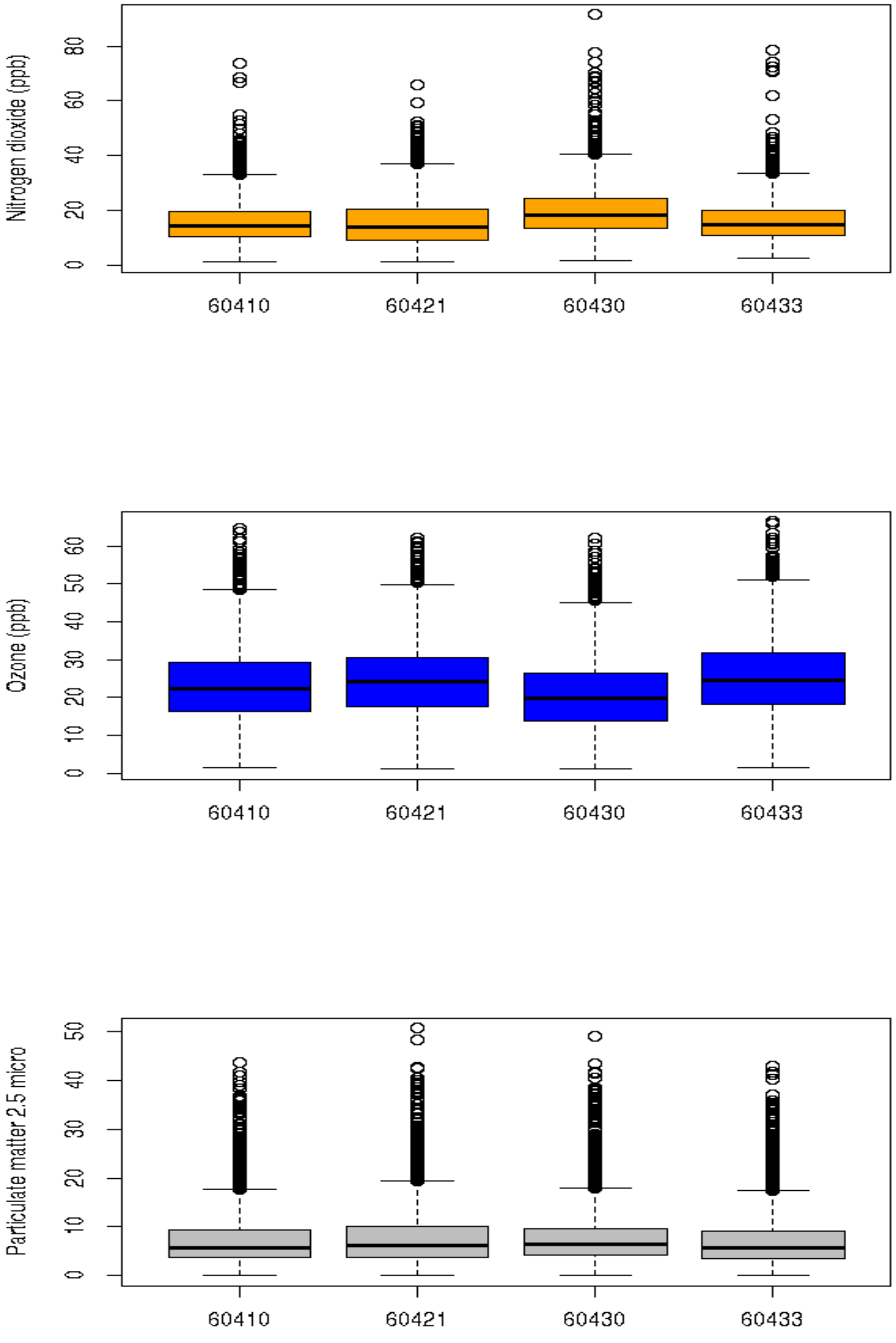


Figure 4.6: Toronto pollutant concentration by station 2005 - 2016

4.3 Metrics Comparisons

Here we will look at the effect of different averages on the daily reported metric for Ozone (O_3) and Nitrogen Dioxide (NO_2). The selected method will be used in our analysis for the Spatio-Temporal Analysis and we therefore see this as a critical step in our analysis of the overall air quality metric calculations and predictions. Before we present these methods we first look at a few plots of our data to have a feel of the representative nature.

The data collection at stations is done on a "continuous" basis and aggregated into hourly readings. For clarity, a day is considered to be the twenty four hour period between 12am and 12am. We looked at different averages to aggregate these hourly, 24 hour readings to a daily reading. The methods used are listed below, being that the method currently used by Environment Canada is the 24-hour arithmetic mean:

24-hour mean This is done by compressing the 24 hourly observations per day using the arithmetic mean to produce one daily observation.

24-hour trimmed-mean This is done by compressing the 24 hourly observations per day using the arithmetic mean then trimming a percentage of the lower and upper tails of the data to produce one daily observation.

Daylight mean The daylight mean compresses the 12 hourly observations during daylight hours (7am to 7pm) to a single daily metric utilising the arithmetic mean

Max 8-hour Using the arithmetic mean, an eight hour maximum mean is computed using eight of the 24 hourly observations per iteration and sliding throughout the 24 observations. The maximum of these computations is then used for the daily metric for the given day

Max 6-hour mean This is computed similarly to the *max 8-hour* mean above using six observations instead of eight.

24-hour median The median of the 24 hourly observations is computed as a daily metric for this method.

Figure 4.7 shows the plots of each of the mentioned averages above.

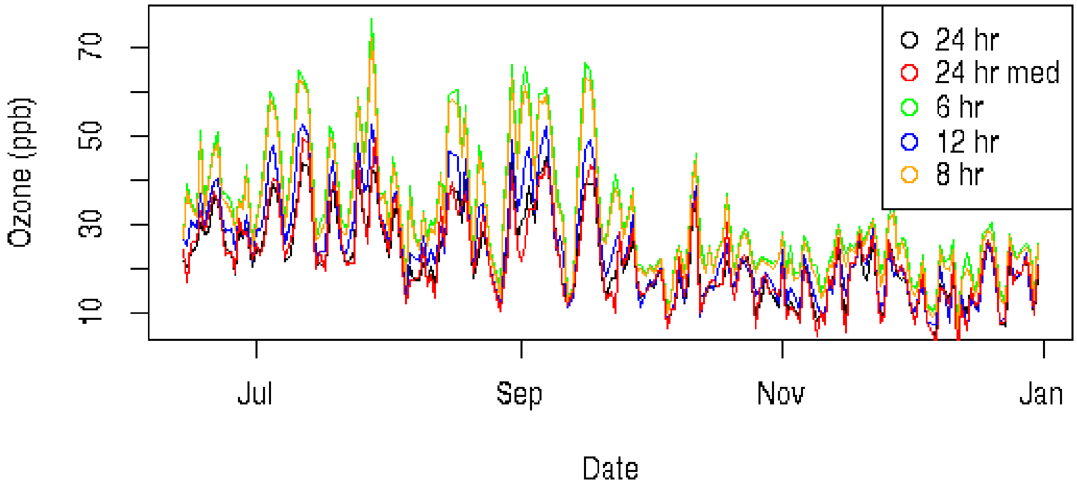


Figure 4.7: Examining Means

4.4 The Study Area: Census Tract

Our study area, Toronto, has as one of its features Census Tracts as defined for the purpose of the census. The 2006, 2011, and 2016 census tracts differ slightly. This is due to the increase in population which in turn causes an increase in housing which results in an increase in the number of census tracts. For this study, we used the 2016 Census Tract which consists of five hundred and seventy two (572) geographic areas. It is worth noting that the change in geography of these areas between 2006 to 2016 was minimal with the main change being additional areas (2006 had 531 such areas, 2011 had 544 areas).

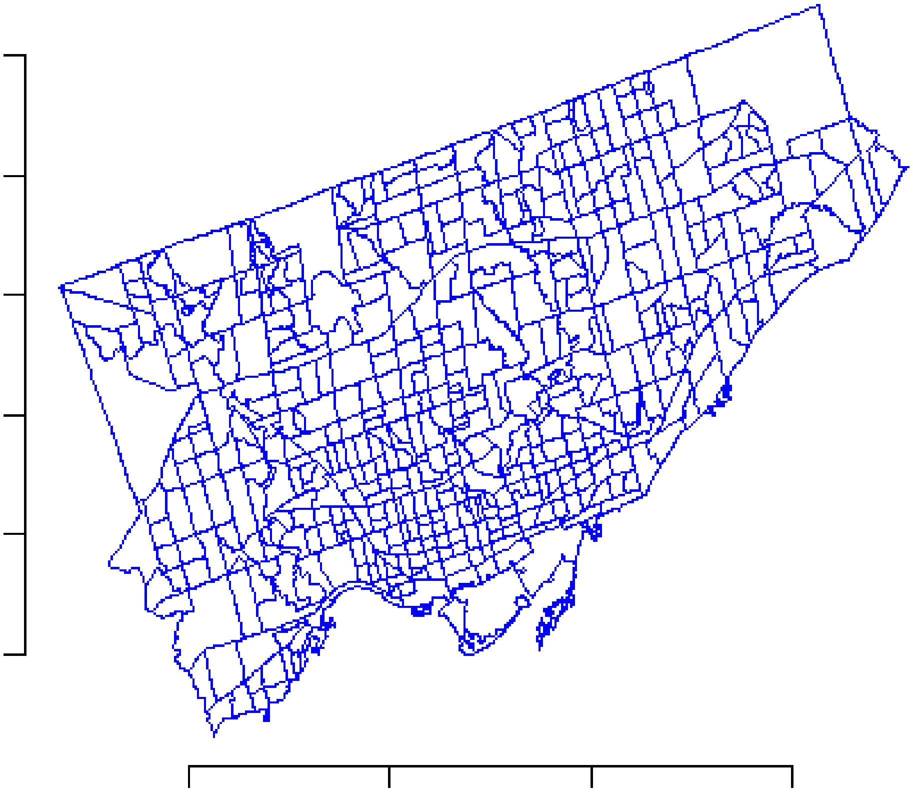
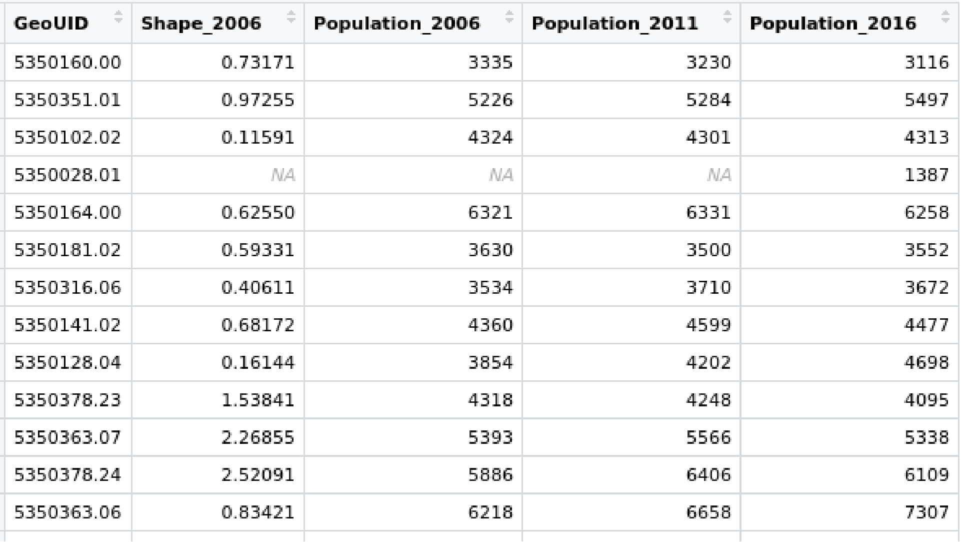


Figure 4.8: 2016 Toronto Dissemination Tracts

Figure 4.8 shows the 2016 dissemination tract used in the study. These tracts are identified by a unique geographical identifier, *GeoUID* (see Figure 4.9). Every five years there is a complete census which is available for public use from Statistics Canada [7]. We used the data for the 2006, 2011, and 2016 population census which gives population down to the census dissemination tract level. These years were used to interpolate the yearly (2006 to 2016) population for each tract assuming a linear relationship.



GeoUID	Shape_2006	Population_2006	Population_2011	Population_2016
5350160.00	0.73171	3335	3230	3116
5350351.01	0.97255	5226	5284	5497
5350102.02	0.11591	4324	4301	4313
5350028.01	NA	NA	NA	1387
5350164.00	0.62550	6321	6331	6258
5350181.02	0.59331	3630	3500	3552
5350316.06	0.40611	3534	3710	3672
5350141.02	0.68172	4360	4599	4477
5350128.04	0.16144	3854	4202	4698
5350378.23	1.53841	4318	4248	4095
5350363.07	2.26855	5393	5566	5338
5350378.24	2.52091	5886	6406	6109
5350363.06	0.83421	6218	6658	7307

Figure 4.9: Population data for 2006, 2011, and 2016 for Toronto Census Dissemination Tracts

In Chapter 3, we explained how our model predicts the posterior mean over a 500 *metre* grid. This grid's resolution is finer that the census tracts' area (meaning each census tract area will have one or multiple grid points). To project the predicted mean values to the census tract we used the overlay function (`over`) from the *sp* package in R. We then used the arithmetic mean of the posterior means that were located in each tract to achieve an average posterior mean per census dissemination tract.

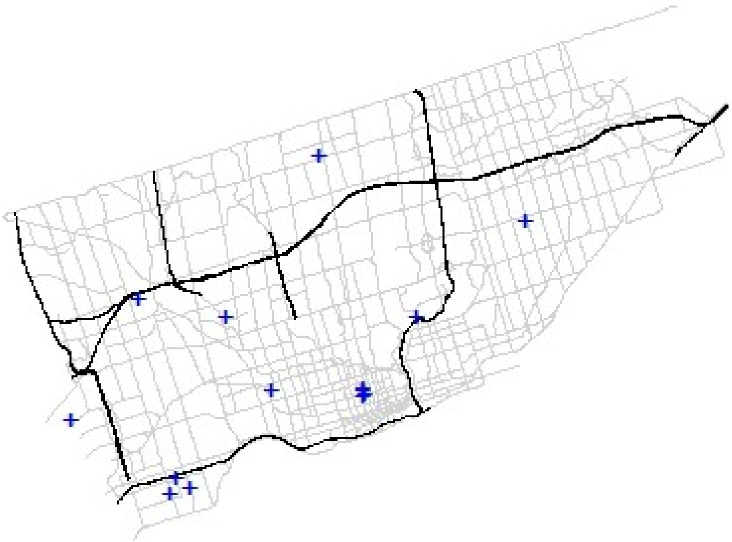


Figure 4.10: Toronto Receptors proximity to Roads. Stations are denoted by blue, while light grey denotes the local roads and black denotes the major highways

4.4.1 Stations' proximity to roads

As with many of the world's cities one of the main sources of air pollution for Toronto is traffic. We have included a plot (see Figure 4.10) of the pollutant stations and their proximity to the larger roads. It is important to note the proximity from the the main highways in particular (black). Note that Figure 4.10 shows all the stations that exist in Toronto over the last twenty years. These stations are not always active. We show this full picture for visualization purposes of all sites and their proximity to major roadways. In the next chapter we will compare how pollutant levels vary with proximity to major highways in particular.

5. *Discussion and Results*

We begin our analysis by presenting and discussing results from the model described in Chapter 3. The data used for each pollutant ($NO_2,\ O_3,\ and\ PM_{2.5}$) were prepared for the model as described in Chapter 4. Our model is implemented through R-INLA [2] which requires the following for computing the required mean field for Toronto:

1. Pollutant data which we have called

   ```
   Toronto_Data
   ```

2. Monitoring Station coordinates which we denote

   ```
   TO_coords
   ```

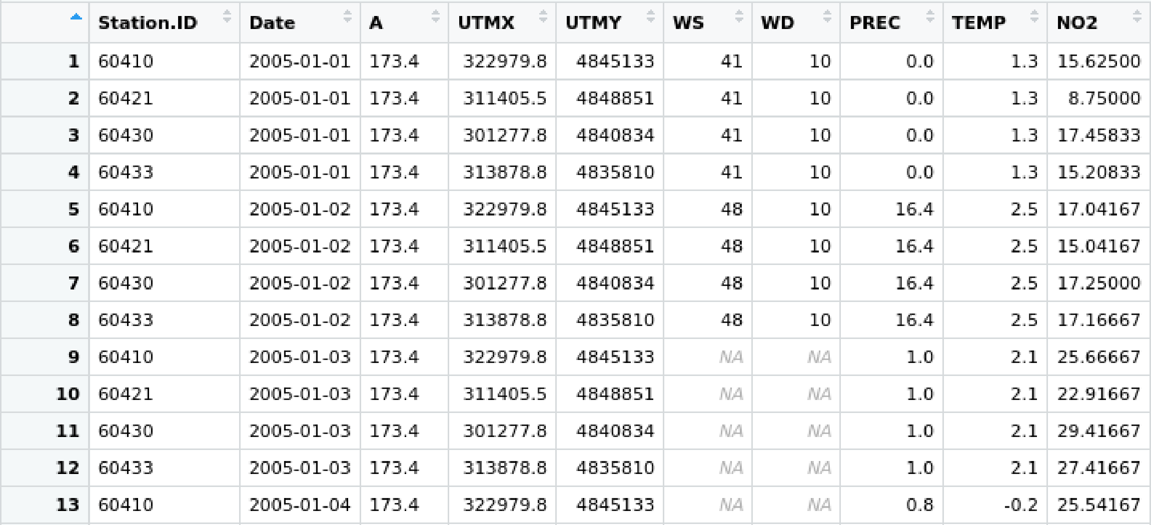
3. Toronto borders which we denote

   ```
   TO_border
   ```

4. Toronto grid discussed in Chapter 3.3.2 denoted

   ```
   TO_grid
   ```

The pollutant data frame mentioned above is stacked by days (see Figure 5.1). Predictions are computed per "stacked" day (where each day's data is presented per station, then the following day per station). As mentioned in Chapter 4, our data



	Station.ID	Date	A	UTMX	UTMY	WS	WD	PREC	TEMP	NO2
1	60410	2005-01-01	173.4	322979.8	4845133	41	10	0.0	1.3	15.62500
2	60421	2005-01-01	173.4	311405.5	4848851	41	10	0.0	1.3	8.75000
3	60430	2005-01-01	173.4	301277.8	4840834	41	10	0.0	1.3	17.45833
4	60433	2005-01-01	173.4	313878.8	4835810	41	10	0.0	1.3	15.20833
5	60410	2005-01-02	173.4	322979.8	4845133	48	10	16.4	2.5	17.04167
6	60421	2005-01-02	173.4	311405.5	4848851	48	10	16.4	2.5	15.04167
7	60430	2005-01-02	173.4	301277.8	4840834	48	10	16.4	2.5	17.25000
8	60433	2005-01-02	173.4	313878.8	4835810	48	10	16.4	2.5	17.16667
9	60410	2005-01-03	173.4	322979.8	4845133	NA	NA	1.0	2.1	25.66667
10	60421	2005-01-03	173.4	311405.5	4848851	NA	NA	1.0	2.1	22.91667
11	60430	2005-01-03	173.4	301277.8	4840834	NA	NA	1.0	2.1	29.41667
12	60433	2005-01-03	173.4	313878.8	4835810	NA	NA	1.0	2.1	27.41667
13	60410	2005-01-04	173.4	322979.8	4845133	NA	NA	0.8	-0.2	25.54167

Figure 5.1: Pollutant (NO_2) data frame used as input for INLA computations

set consists of data for each pollutant for the period January 1, 2006 to December 31, 2016. We also presented the pollutant stations for Toronto in Chapter 4, with some of these stations going online (active) or offline (inactive) due to various reasons. These reasons include geographical relocation of stations, changes in population, and other factors. We have employed four stations for our computations as these stations were online for the entire study period, and furthermore they represented Toronto geographically.

Stations are spread as North Toronto (North York area, north of highway 401), East Toronto (Scarborough area), South Toronto (downtown area), and West Toronto (highway 401 north of Etobicoke and west of Pearson International Airport). Note the weather related covariates in the data frame `Toronto_Data,` were chosen from the Toronto Pearson weather station as the station was active for the study period. Missing weather data was replaced with a moving arithmetic mean. These covariates along with the geostatistical covariates for each pollutant station is also presented in Figure 5.1. In addition, we have assumed that the average weather conditions per day were the same throughout Toronto (as measured at the Toronto Pearson weather station). It is also important to note that the respective pollutant measurements are

Symbol	**Definition**
A	Altitude (meters)
UTMX and UTMY	Universal Transverse Mercator (UTM) X and Y coordinates see Chapter 4.4.1
WS	Wind Speed measured in kilometres per hour (km/h)
WD	Wind Direction measured in tens of degrees (10's degrees, where 10 10's deg is 100 degrees
PREC	Precipitation measured in millimetres (mm)
TEMP	Temperature, collected as degrees Celsius and converted to Kelvins (K) for computation

Table 5.1: Covariates used in INLA

log transformed as required by our model (see Equation (3.1)).

The grid explained in Section 3.3.2 is plotted in Figure 5.2. This grid consists of 5332 points, and our interest lies in predicting the posterior mean for each of these points. The mesh used in the Toronto study (see Figure 3.6) was created from $G = 306$ vertices. This mesh was then used to create the Matèrn SPDE object and the projector matrix **A** (from Section 3.2) using the following code:

```
Toronto_spde <- inla.spde2.matern(mesh = Toronto_mesh, alpha=2)
A_est <- inla.spde.make.A(mesh = Toronto_mesh,
                          loc = coords.allyear,
                          group = Toronto_Data$time,
                          n.group = n_days)
stack_est <- inla.stack(data = list(logNO2 = Toronto_Data$logNO2),
                        A = list(A_est, 1),
                        effects = list(c(s_index, list(Intercept = 1)),
                                       list(Toronto_Data[,3:9])),
                                       tag = "est")
```

Note the projector matrix **A**'s (object `A_est`) number of columns is given by the number of mesh vertices times number of time points (days). In running the INLA

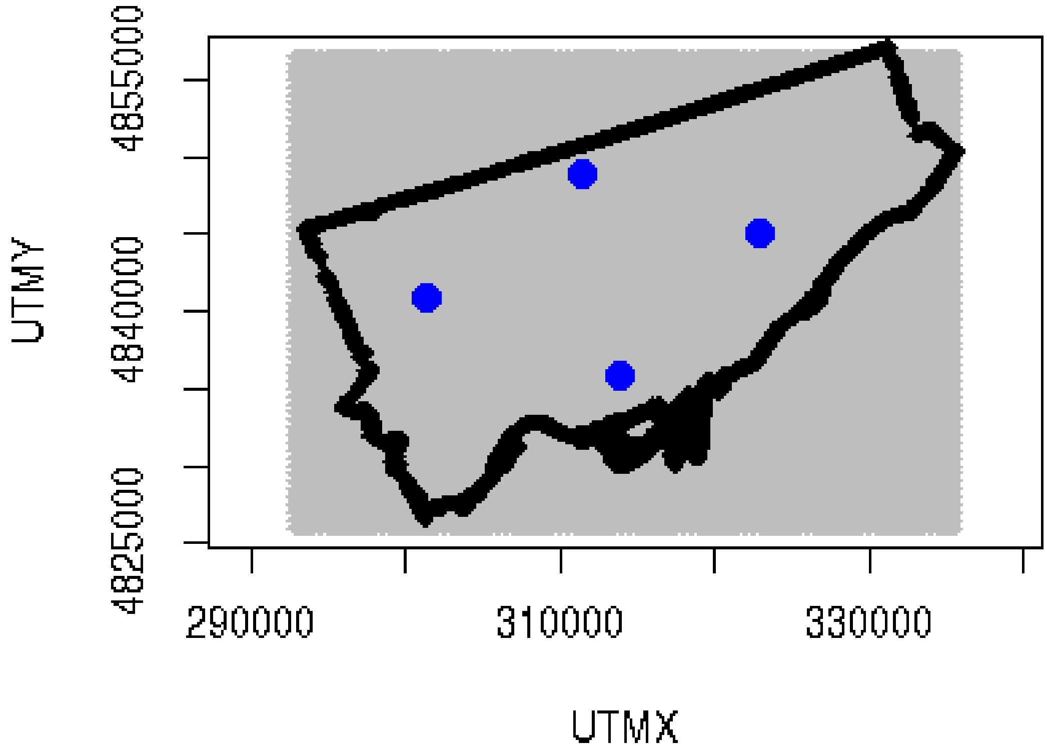


Figure 5.2: Toronto grid of 5332 points (grey area of plot), blue dots denote the monitoring stations

trials for the chosen mesh (306 vertices) it was noticed that the program would break for time points beyond forty points (days) as a result of the limited computing power. The data was computed by months which reduced the time points to at most 31 days, which resulted in a 306 x 31 columns by 5332 number of grid points rows projector matrix. The `group` argument represent the time grouping (4 measurements for each day, one from each of the stations of study) and the `n.group` arguments represent the number of groups (28, 29, 30 or 31 days depending on the month). The spatial field object being computed (with the object number of time points replicated by the number of vertices) by the `inla.stack` function was used to create the stack object for the estimation of the spatio-temporal model. The function takes the logarithm of the data and the covariates for the monitoring stations. The similar object for the prediction of the particular day (`i_day`) is then created again by the use of the `inla.spde.make.A` function:

```
A_pred <- inla.spde.make.A(mesh = Toronto_mesh,
                           loc = as.matrix(TO_grid),
                           group = i_day,
                           n.group = n_days)
```

The stack object for the prediction is created as follows. Note that the standardised covariates `cov_matrix_std` are employed for the prediction stack.

Our code thus far has been for the general case for each pollutant, and we now use the specific case for pollutant NO_2 for the remainder of this section with the process from this point on.

```
stack_pred <- inla.stack(data = list(logNO2 = NA),
                         A = list(A_pred, 1),
                         effects = list(c(s_index, list(Intercept = 1)),
                         list(cov_matrix_std)), tag="pred")
```

The stack for both the estimate and the prediction was then combined to a full stack. The output for the `inla` function then provided the estimate directly of the linear predictor at the grid level for the day required for prediction. The formula that was used for our prediction included an explicit intercept for the covariates. The `group` and `control.group` arguments that are passed to the function are used to specify that the spatial locations are linked by the SPDE model object (`spde`) at each time point. The process then transforms across time by an AR(1) process (see Equation (3.3)).

```
formula <- logNO2 ~ -1 + Intercept + UTMX + UTMY + WS + A + WD + PREC +
  TEMP + f(spatial.field, model = Toronto_spde,
           group = spatial.field.group,
           control.group = list(model = "ar1"))
```

To obtain our output, the final step in the process is then computed by using the following code. This is the step that is computationally expensive and the cause for reducing the density of the mesh (thus reducing the vertices to achieve computation without breaking):

```
output <- inla(formula, data = inla.stack.data(stack, spde = Toronto_spde),
               family = "gaussian", verbose = T,
               control.predictor = list(A = inla.stack.A(stack),
                                        compute = TRUE))
```

From our output we retrieved the posterior summary statistics for the fixed effects β, $1/\sigma_e^2$ and the AR(1) coefficient a. The posterior estimates for the spatial range r and variance σ^2 are also retrieved. We then extracted the posterior marginals of the linear predictor (available for all of the Toronto grid locations) to obtain the smooth prediction (without measurement error) for i_day and then compute the posterior mean of the exponential distribution to achieve the natural NO_2 prediction. This process was replicated for O_3 and $PM_{2.5}$.

5.1 INLA metrics and mean fields comparison for pollutants

We begin our comparison by looking at the posterior mean field prediction produced from our results for a specific day July 10, 2010 for our three pollutants. We first

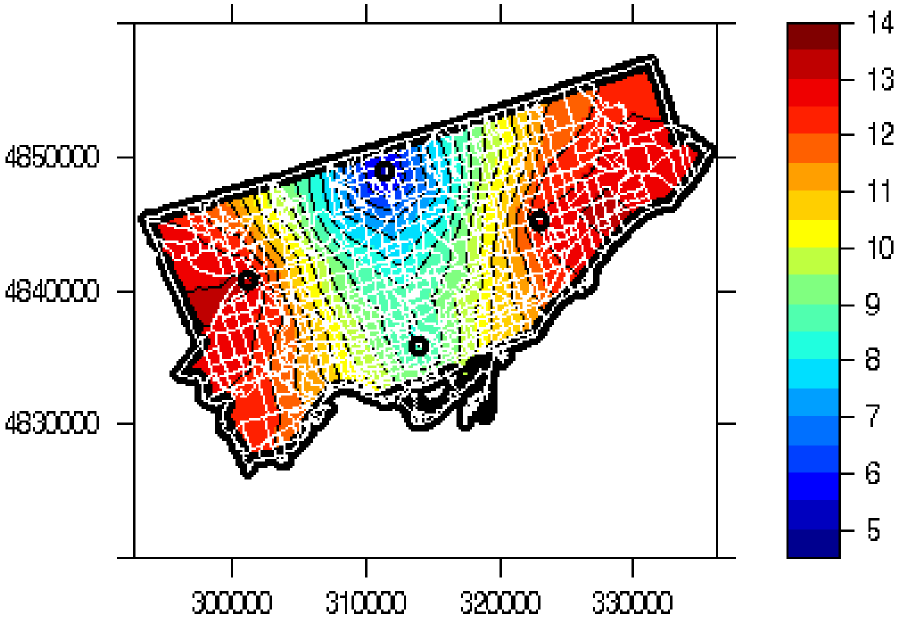


Figure 5.3: Toronto posterior mean of Nitrogen Dioxide (NO_2) for July 10, 2010. Figure plotted using Universal Transverse Mercator (UTM) coordinate system

present the posterior mean field for NO_2. Figure 5.3 shows that over our region, predicted values for NO_2 range from 4.5 *ppb* to 14 *ppb*. Pollutant levels are lowest in the central northern Toronto area and relatively average (from our range) in the central southern Toronto area. Levels are closer to the high values in our range in the eastern and western Toronto areas. Nitrogen dioxide is part of a group of gaseous air pollutants produced as a result of road traffic and other fossil fuel combustion processes as discussed in Chapter 1. For this given day we notice that this reasoning would be flawed if we were to argue in a vacuum, since our lowest predictions for the day is just north of the 401 highway (where traffic is usually high).

There are other factors which affect our predictions (covariates). Our posterior mean predictions are likely explained by these variations. We mentioned proximity to major highways, but there are a variety of other sources within the metro Toronto area that will affect pollutant measurements and hence predictions. For our project, we are mainly looking at the main contributors (e.g. large highways). Figure 5.4 shows the predicted posterior mean for Ozone with range of 20 *ppb* to 31 *ppb*, with higher predictions for central north and central south Toronto, and lower reading for Toronto east and west. Note again that lower and higher readings are relative to the given range. We will look at changes over the yearly range later in this chapter.

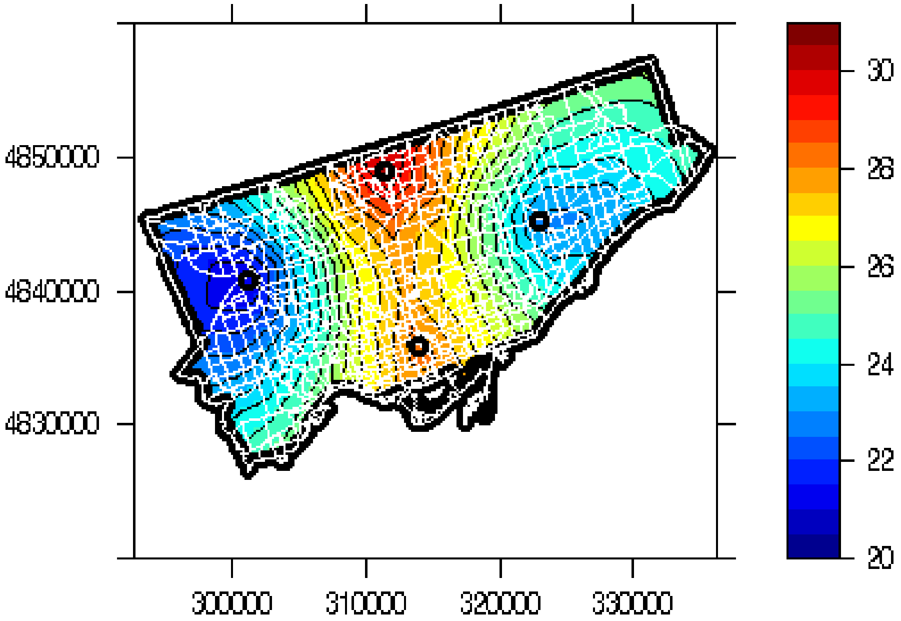


Figure 5.4: Posterior mean of Ozone (O_3) for July 10, 2010. Figure plotted using Universal Transverse Mercator (UTM) coordinate system

Figure 5.5 show the posterior mean for particulate matter with a range of 4 μg/m^3 to 9 μg/m^3. Here predictions are lower in the central and northern Toronto area with slightly higher readings in the east and west Toronto area. The highest posterior means were recorded in the far east of Toronto (close to the border). Another in-

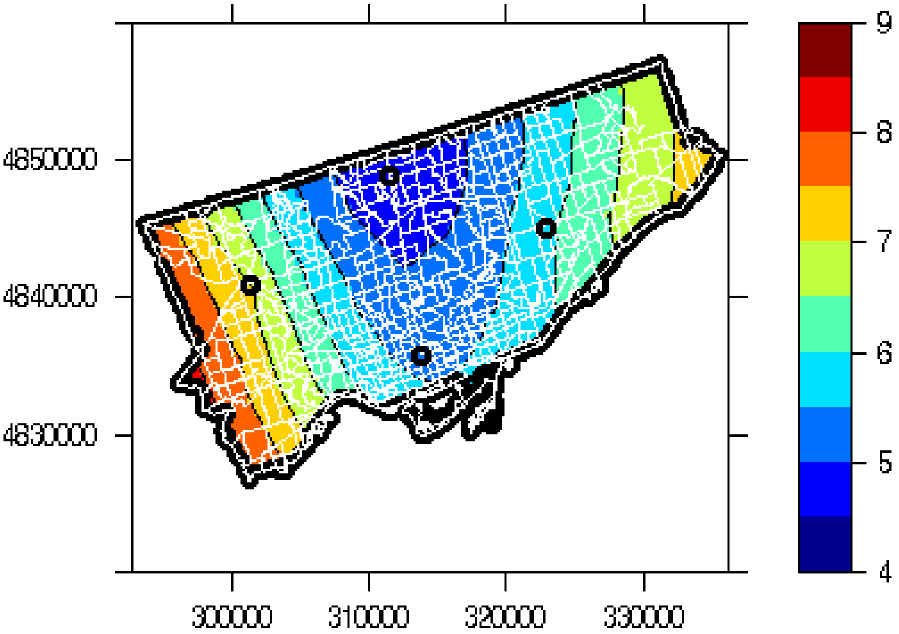


Figure 5.5: Posterior mean of Particulate Matter 2.5 (PM$_{2.5}$) for July 10, 2010. Figure plotted using Universal Transverse Mercator (UTM) coordinate system

teresting plot is the probability of exceeding a given threshold. Table 5.2 gives the annual maximum acceptable concentrations for each pollutant for Canada. For July 10, 2010 we see that the maximum predicted posterior mean for nitrogen dioxide and particulate matter (2.5 microns) are 14 *ppb* and 9 μg/m^3 respectively, since these are well below the limit we need not examine the plot of posterior probability exceeding acceptable levels. Ozone, however, is well above the annual limit of 15 *ppb*, we will relax this limit to 20 *ppb* for the purpose of visualising the probability. Figure 5.6 shows the posterior probability of exceeding the maximum concentrations for July 10, 2010. Note that there are areas where it is certain (Probability = 1) that the concentration limits will be exceeded. It must be noted that without relaxing the limit, all of the region would have exceeded the limit for the chosen day.

Pollutant	**Exposure Period**	**Concentration**
NO_2	Yearly	50
O_3	Yearly	15
$PM_{2.5}$	Yearly	70

Table 5.2: National ambient air quality objectives for pollutants (maximum acceptable concentrations)

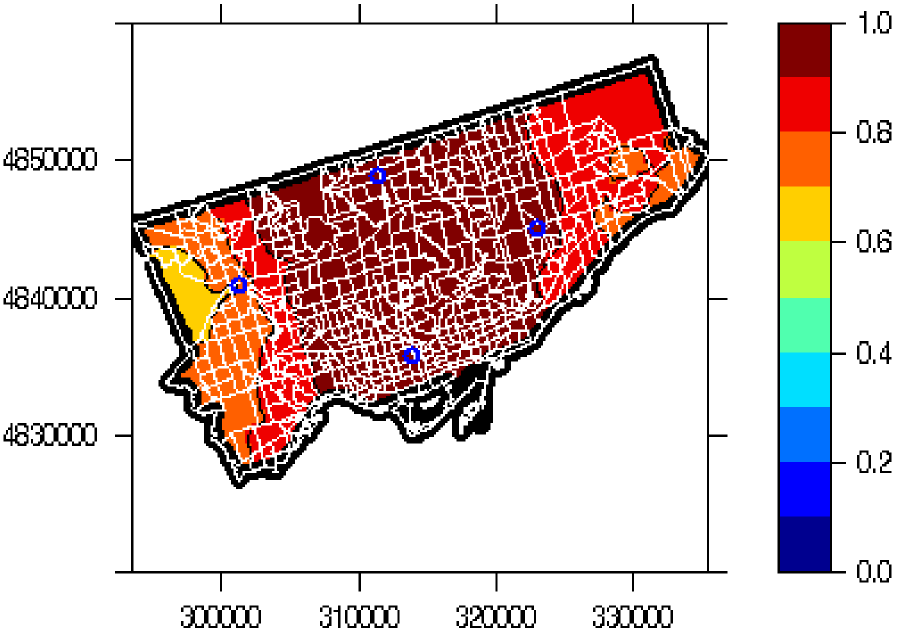


Figure 5.6: Probability of O_3 Posterior mean field exceeding allowable limit for July 10, 2010. Figure plotted using Universal Transverse Mercator (UTM) coordinate system

5.2 Seasonal Trends

The seasons of a year bring changes in the weather and hence environmental effects that result from these changes. We present the trends due to the effects of these seasons. For clarity, the seasons are defined as Winter (December, January, February), Spring (March, April, May), Summer (June, July, August), Fall (September, October, November) for our area of study, Toronto. Figure 5.7 shows the posterior mean by the seasons for nitrogen dioxide. It can be seen that the posterior mean is highest in the colder seasons and lower in the warmer seasons. We could argue that this is due to higher NO_2 being produced from sources such as: heating requirements in the winter, and vehicles (people tend to drive instead of commuting). One could argue that there are also activities in the hotter months that produce high NO_2. We could not argue the case of traffic effect by season as only the average annual daily traffic (AADT) data is publicly available [24].

There is however another phenomenon that causes lower levels for NO_2 in warmer months. Looking at Figure 5.8 we can see that the opposite is true for O_3 levels - they are higher in warmer months. While NO_2 is directly emitted from various sources, the formation of ground level O_3 requires four key ingredients: volatile organic compounds (VOC), nitrogen oxides (NO_x), heat and sunlight. We know that NO_2 is directly emitted, and VOC is also directly emitted as a result of human activity. Heat and sunlight depends on weather which changes with seasons. Outdoor air quality can therefore be expected to worsen with high temperature, and especially in hot summer days without clouds, where there is an affluence of sunlight. This is due to the fact that NO_x and VOC provide extra oxygen atoms that combine with atmospheric oxygen (O_2) to form ozone (O_3). We use the following simple chemical equation to explain how O_3 can be formed from NO_2:

$$NO_2 + Sunlight = NO + O,$$

$$O + O2 = O3$$

it can be seen that this chemical reaction "takes" an oxygen atom from NO_2. This "available" oxygen atom then combines with O_2 forming O_3, generally resulting in lower levels of NO_2 and higher levels of O_3.

Figure 5.9 shows the seasonal changes in $PM_{2.5}$. Particulate matter can be emitted or formed. As with O_3 formation becomes faster with heat. PM is considered primary when it is emitted directly from combustion, and secondary when it is produced from the chemical reaction of other air pollutants. Concentration of secondary PM increases during heat waves, since the reactions that lead to its formation are accelerated as heat acts as a catalyst (similar formation of O_3 explained earlier). NO_x, ammonia (NH_3), sulphur oxides (SO_x), and VOC have been identified as reagents that contribute to the formation of secondary particulate matter (explaining the higher levels of $PM_{2.5}$). When high temperatures coincide with periods of high atmospheric pressure, ozone and particulate matter can reach dangerously high levels. While high temperature leads to faster formation of ozone and fine particulate matter, high pressure makes it difficult for natural air currents to dissipate these pollutants. Rainfall is generally beneficial for air quality, since particulate matter in the air adheres to water droplets and falls to the ground. We will look at the risks associated with pollutant levels in the next chapter.

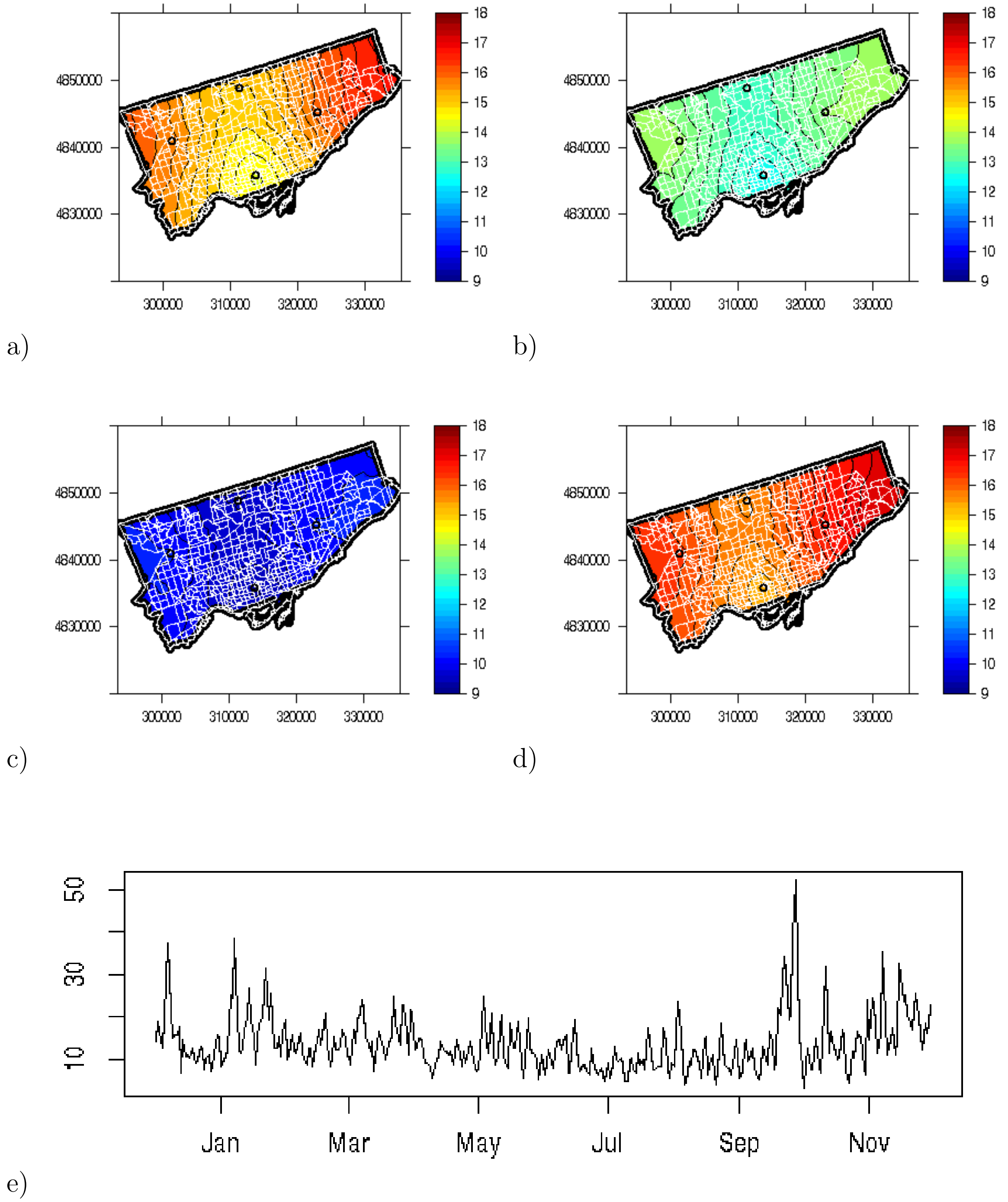


Figure 5.7: Plots a) to d) are the posterior mean field for NO_2 concentration (in *ppb*) by seasons: Winter, Spring, Summer, and Fall respectively for Toronto plotted using the UTM scale. Plot e) is the time series plot for the period Dec 2015 to Nov 2016 of daily NO_2 predicted mean field concentrations (in *ppb*) for Toronto

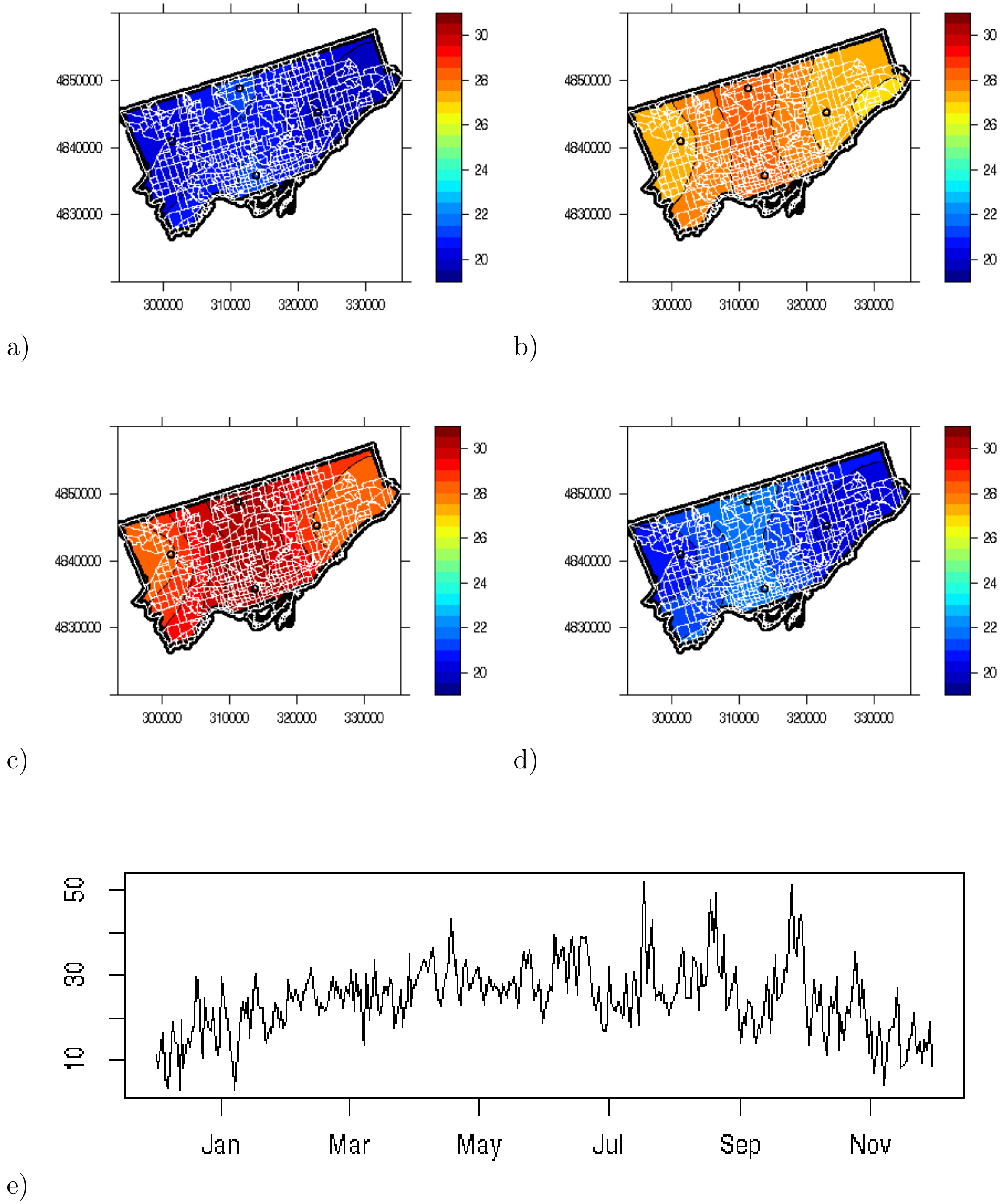


Figure 5.8: Plots a) to d) are the posterior mean field for O_3 concentration (in *ppb*) by seasons: Winter, Spring, Summer, and Fall respectively for Toronto plotted using the UTM scale. Plot e) is the time series plot for the period Dec 2015 to Nov 2016 of daily O_3 predicted mean field concentrations (in *ppb*) for Toronto

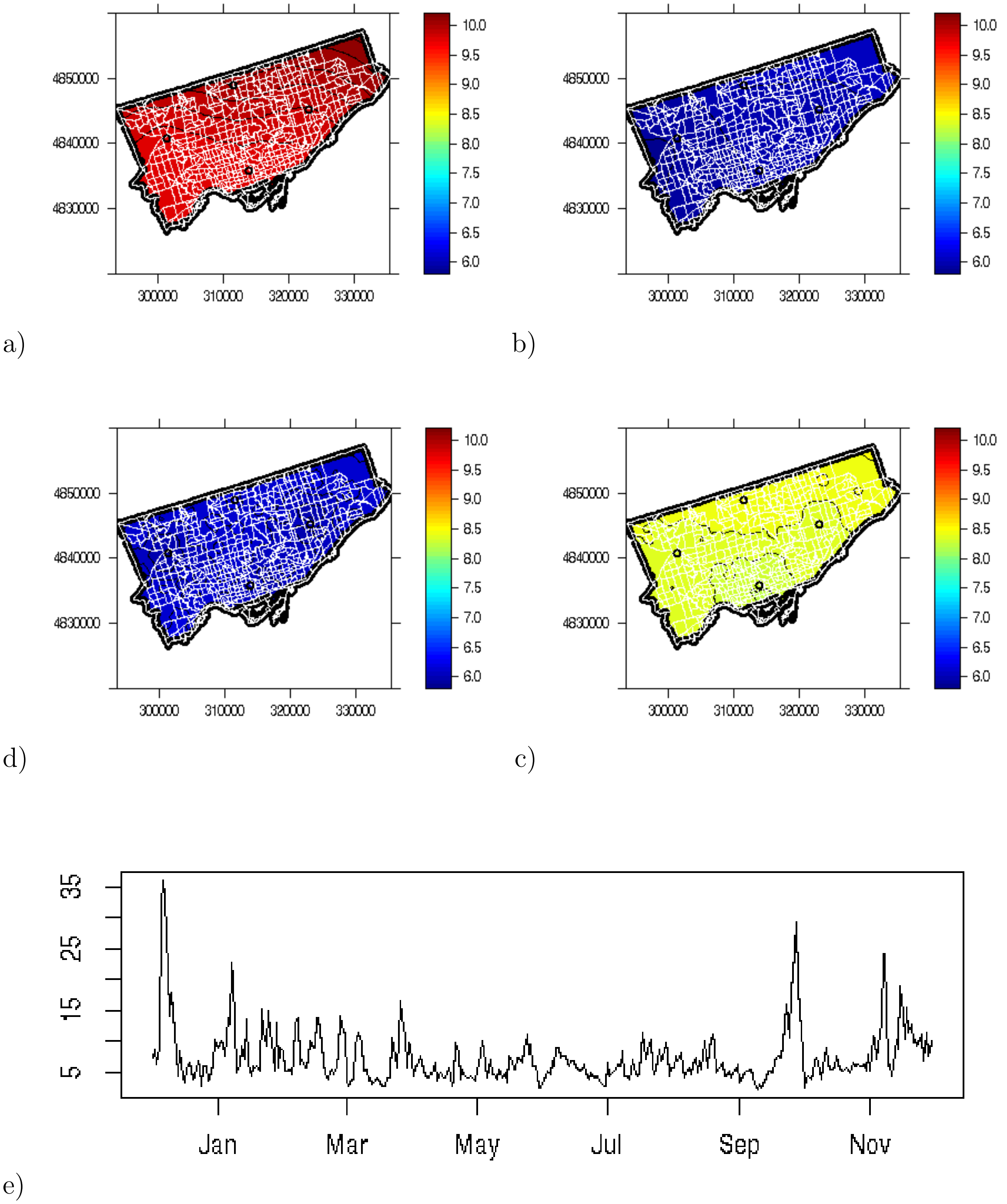


Figure 5.9: Plots a) to d) are the posterior mean field for $PM_{2.5}$ concentration (in $\mu g/m^3$) by seasons: Winter, Spring, Summer, and Fall respectively for Toronto plotted using the UTM scale. Plot e) is the time series plot for the period Dec 2015 to Nov 2016 of daily $PM_{2.5}$ predicted mean field concentrations (in $\mu g/m^3$) for Toronto

5.3 Spatio-temporal trends

We will first look at the trend of each pollutant. Changes in the natural logarithm are (almost) equal to percentage changes in the original series, it follows that the slope of a trend line fitted to log transformed data is equal to the average percentage growth in the original series. For this reason we will use the log transformed posterior mean field data to investigate the trend for each pollutant of the study period (January 2006 to December 2016). We fitted a General Additive Model (GAM) [14] to this log transformed data in order to achieve the trend.

Figure 5.10 shows the log transformed posterior mean field average for NO_2. Here we see that there is a slight downward trend in NO_2's percent concentration. We discussed earlier that there is a link between NO_2 emissions and O_3 formation, that is O_3 is formed from NO_2. From Figure 5.11 we can see that the O_3 levels are rising (at a slightly lower rate that the decreasing rate of NO_2). The "faster" rate at which NO_2 is decreasing when compared to O_3 can be attributed to other factors such as stricter regulations for emissions for the province of Ontario over the years.

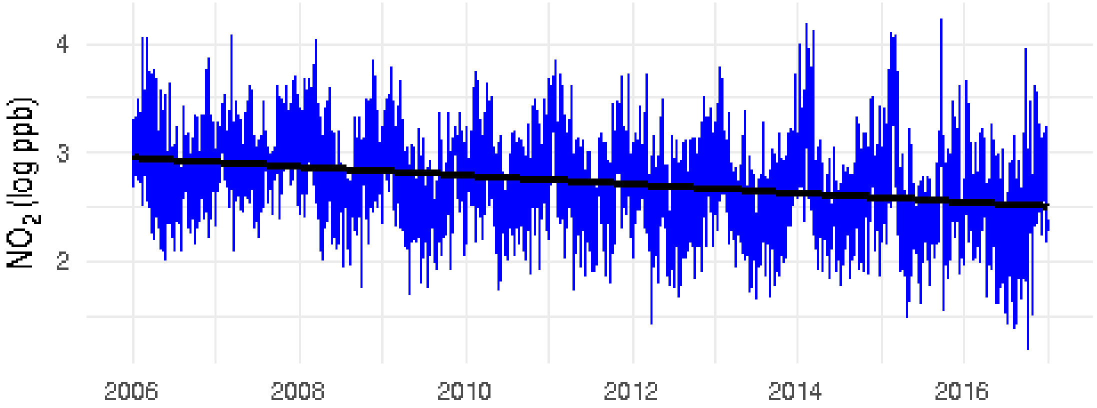


Figure 5.10: Trend for NO_2 over the period Jan 2006 to Dec 2016.

Figure 5.11 shows the log transformed posterior mean field average for O_3. From this plot we can see that there is a steady increase in the percent concentration (approximately 0.5%). This trend is likely to continue in the coming years unless actions are taken to reduce emissions within the city. The effects of global warming will add to this steady rise, as rising temperatures will also cause rising levels of O_3 (as heat is a catalyst in converting NO_2 and other emitted compounds to O_3). Fighting global warming will therefore aid in reducing O_3 levels for the future.

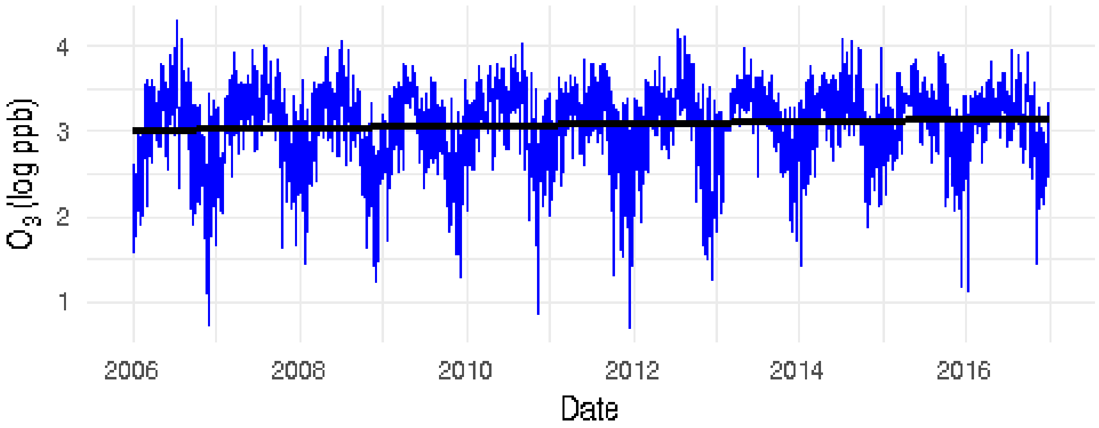


Figure 5.11: Trend for O_3 over the period Jan 2006 to Dec 2016.

We can see from Figure 5.12 that there is a slight upward trend in $PM_{2.5}$ concentration levels. As noted earlier in this chapter, particulate matter is produced in two ways: direct emissions and formation from other emitted substances. This noticeable upward trend can be attributed to this fact. While there is a steady increase in emitted pollutants, we will continue to see this rising trend continue. While the province has put various regulations in place that restrict emission levels, there is still more work needed to reduce particulate matter in the city of Toronto. We will discuss the risk implications of pollutant levels in the following chapter. We will now look at the differences in three time series plots.

(i) *Station Mean* - the daily averaged pollutant concentration observations across stations over study period,

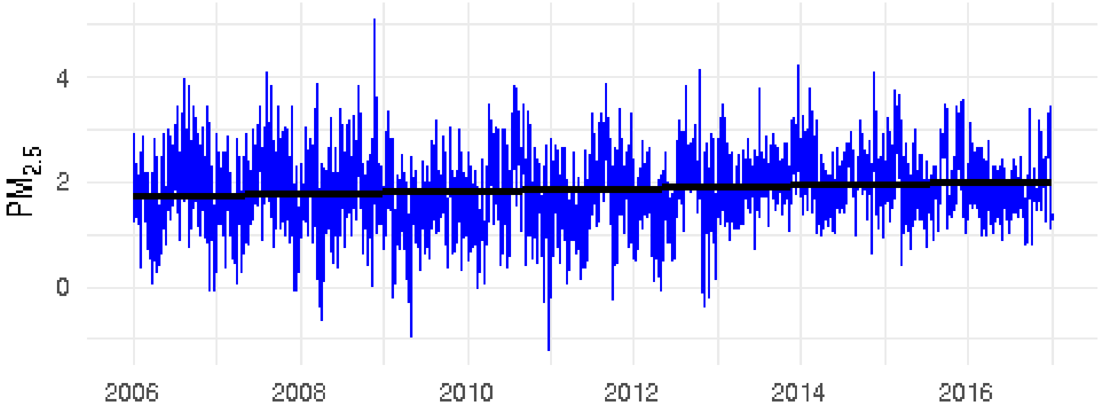


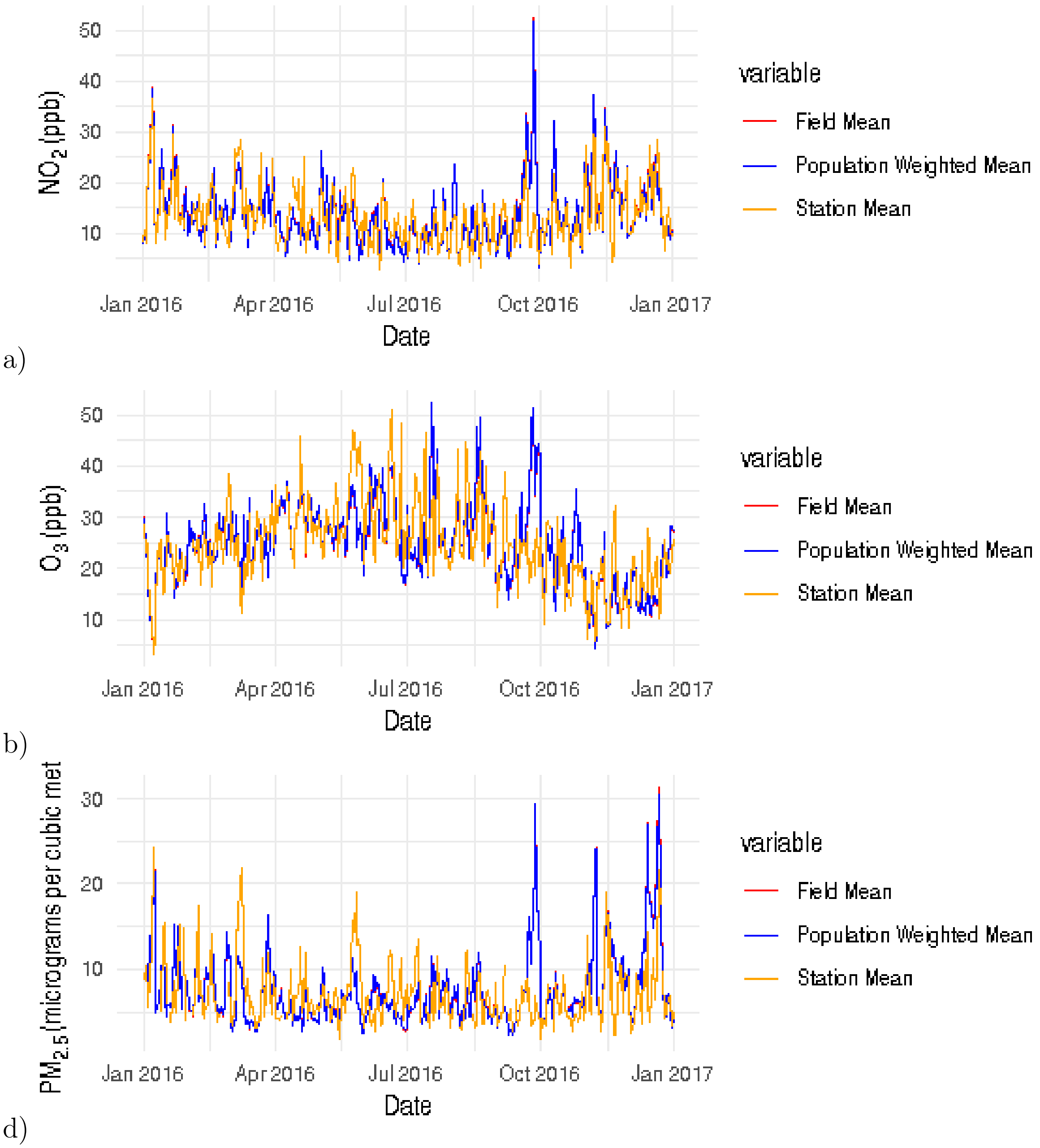
Figure 5.12: Trend for $PM_{2.5}$ over the period Jan 2006 to Dec 2016.

(ii) *Mean Field* - the daily averaged posterior mean field for Toronto (from spatio-temporal INLA results), and

(iii) *Population Weighted Mean* - the posterior mean fields above weighted by population (from Toronto Tracts) then averaged for a daily results.

For these time series we will be looking at one year's data (2016) to see the daily differences. Figure 5.13 shows the three stated time series for the pollutants. We can see that the differences between the *Mean Field* and the *Population Weighted Mean* are not huge. However, when we compare these to the *Station Mean* we can see that there is a larger noticeable difference; we can see that there were days where the *Station Mean* presented an under representation of the concentration for pollutant, while there were days that it gave an over representation.

As noted in the previous chapter pollutant levels may be affected by their proximity to highways. We will now examine the effects of a major highway (Highway 401) on the pollutant levels of the posterior mean compared to a location farther from said highway. In particular, we examine the posterior mean close to the Highway 401 and Don Valley Parkway intersection and another posterior mean in South West Toronto (Etobicoke) area. Figure 5.14 shows the comparison of the above mentioned areas for



a)

b)

d)

Figure 5.13: Plots a) Comparison of daily concentration levels of *Station Mean*, *Mean Field*, and *Population Weighted Mean* for each pollutant a) NO_2, b) O_3, and c) $PM_{2.5}$.

each pollutant (for the year 2016). The red line in the plot represents the highway location mentioned above. The blue line represents the location in the Etobicoke area. When we look at the plot for NO_2 we can see that there is a clear difference at various dates. On average all the pollutants show higher concentrations for the highway location. For NO_2 the maximum difference in concentration exceeded 12 ppb in the summer (June 2016) and again in the Fall (October 2016). O_3 for the highway location concentration difference exceeded 10 ppb at its maximum in the spring (April 2016). For $PM_{2.5}$ the highway location exceeded 5 $\mu g/m^3$ at its maximum concentration difference (January and August 2016).

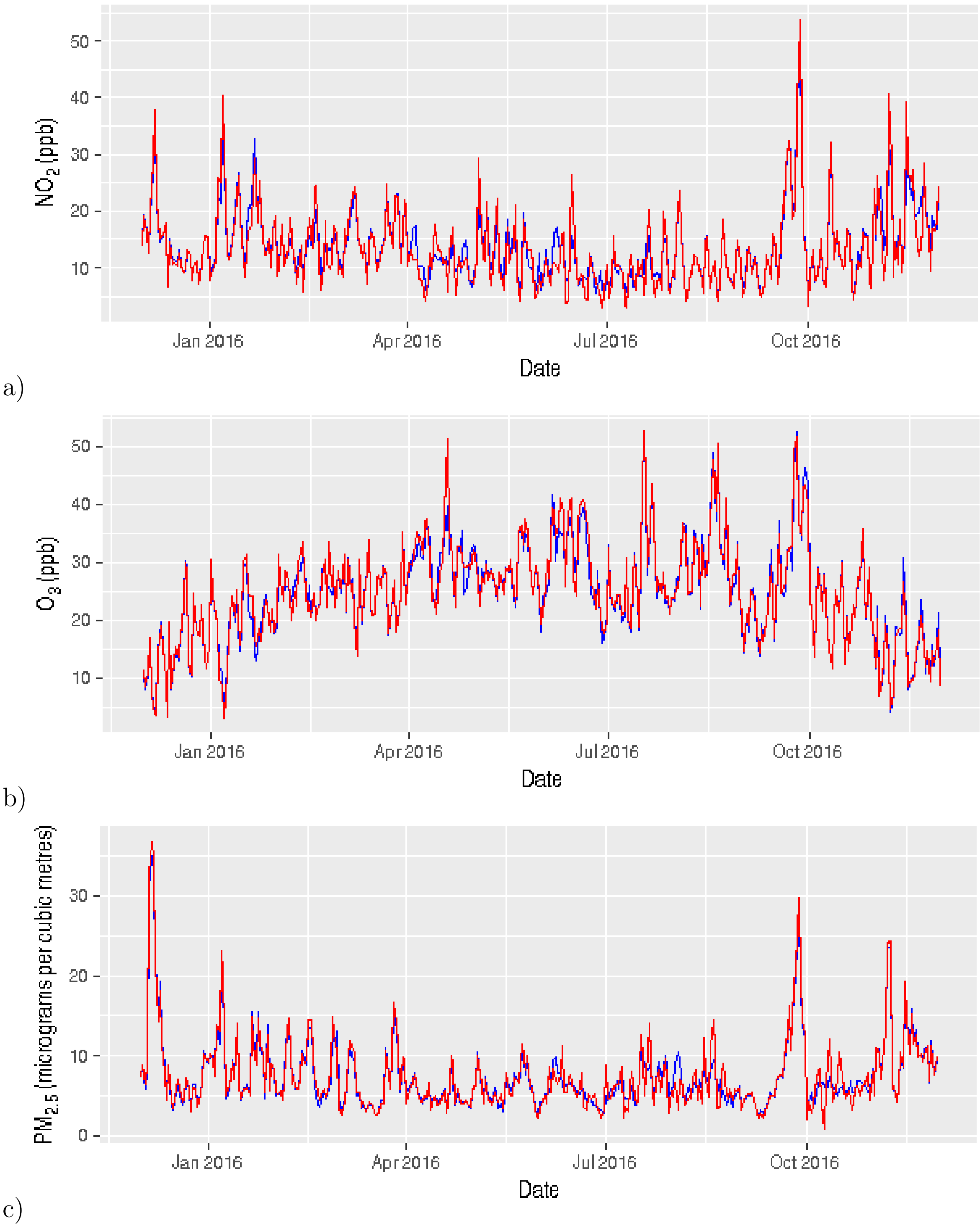


Figure 5.14: Plots a) Comparison of daily concentration levels of *Population Weighted Mean* for each pollutant a) NO_2, b) O_3, and c) $PM_{2.5}$.

6. *Risk Analysis*

Various studies have shown the adverse health effects that short-term exposure to air pollutants present. Specifically, these health effects include increased risk of premature mortality. In this chapter, we set out to examine the risk estimates between the simplistic population exposure metric and our new metric presented in the previous chapter, using previous work.

We set out by using a model for the estimation of population health risk as presented in the work by Burr, Takahara and Shin [3]. This model is built around a Poisson Generalised Linear Model (GLM) or GAM. The log transformed daily mean mortality counts is modelled as an additive combination of a smooth function of daily mean temperatures, a smooth function of time, a categorical variable for the day-of-week effect, and a parametric air pollution covariate (NO_2, O_3, or $PM_{2.5}$). We will adjust the baseline version of the model employed in [3] as required. We present the baseline, for Y_t, a daily mortality count series indexed by day, the following can be written:

$$\log(\mu_t) = \beta_0 + x_t\beta_1 + \mathrm{DOW}_t\beta_2 + \mathrm{S}(\mathrm{Temp}, df = 3)_t + \mathrm{S}(\mathrm{Temp}, df = 6/\mathrm{yr})_t \qquad (6.1)$$

where $\mu_t = E[Y_t]$ is the time-varying daily mean mortality, x_t is the pollutant of interest, also measured by a daily metric, such as sample mean, DOW_t is an indicator function for the day-of-week, both S functions are natural cubic regression splines with the indicated df, computed on calendar time and daily mean temperature, both on the same day as the response. Seasonality presents a strong variability for mortality in our study area Toronto. The smooth function presented in Equation 6.1 accounts for unmeasured risk factors such as the long timescales structures that may be present. For example, influenza epidemic is highest in mid-winter for our study area and demographic shifts vary on a multi-year or decadal time scales.

We will compare the conventional population health risk explained above (we will refer to this as *population health risk (average)*, estimated with the naive average pollution metric) with the risk for our weighted-by-population mean pollution exposure metric (we will refer to this as *population health risk (population weighted)*). For both risk measures we take into account:

(i) the mortality caused by all non-accidental deaths for both genders and all ages birth to death, and

(ii) the mortality caused by all cardio-pulmonary deaths for both genders, also for all ages birth to death.

6.1 Mortality (all non-accidental deaths)

The response that was used for this "non-accidental death risk" include all deaths for the population of our study area that were non-accidental (all deaths not caused by accidents), both genders (male and female), and for all ages (births to deaths) of the population. The data set that was used was complete in that a complete interpolation was done with bad data points flagged and removed. Figures 6.4, 6.5, and 6.6 show risk plots for NO_2, O_3, and $PM_{2.5}$ respectively. The blue solid and dash curve in our plot represents *population health risk (average)*, and its confidence interval (with confidence level 95%) respectively. The black curve and grey fill represents *population health risk (population weighted)*, and its confidence interval (with confidence level

95%) respectively. Note we used degree of freedom (df) = 26 (30 days) for these risk calculations. In Figure 6.1 *population health risk (average)* is within the 95%

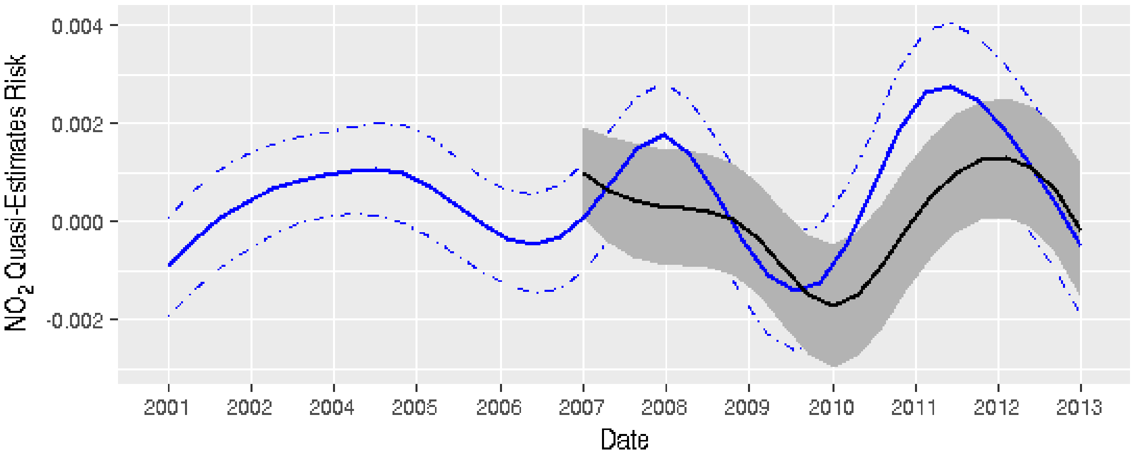


Figure 6.1: NO_2 *population health risk (average)* versus *population health risk (population weighted)*, blue and black curves respectively.

confidence interval for *population health risk (population weighted)* for most years. If we were to interpret this in a traditional sense, there is not a statistical difference between the two curves (since their confidence intervals overlap). This indicates that there is some form of relationship between *population health risk (average)* and *population health risk (population weighted)* - as expected, as the models are very similar, differing only in the exposure metric used.

In Figure 6.2, *population health risk (average)* for O_3 is only within the 95% confidence interval for years 2008 – 2009, 2011 and part of 2013. It appears that for this pollutant that the two risk curves track, indicating that the population-weighted exposure metric is similar enough to the default naive averaging approach as to provide similar interpretations. The differences are still large enough to warrant further examination, however: it's not as if the risk curves are exactly tracking, despite the somewhat subtle changes.

In Figure 6.3 *population health risk (average)* for $PM_{2.5}$ is within the 95% confidence interval of *population health risk (population weighted)* for the majority of the

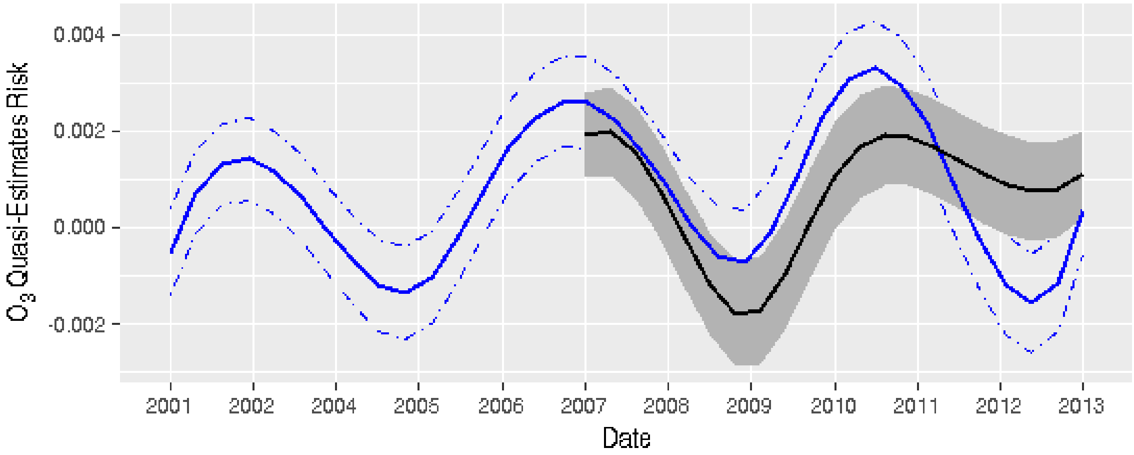


Figure 6.2: O_3 *population health risk (average)* versus *population health risk (population weighted)*, blue and black curves respectively.

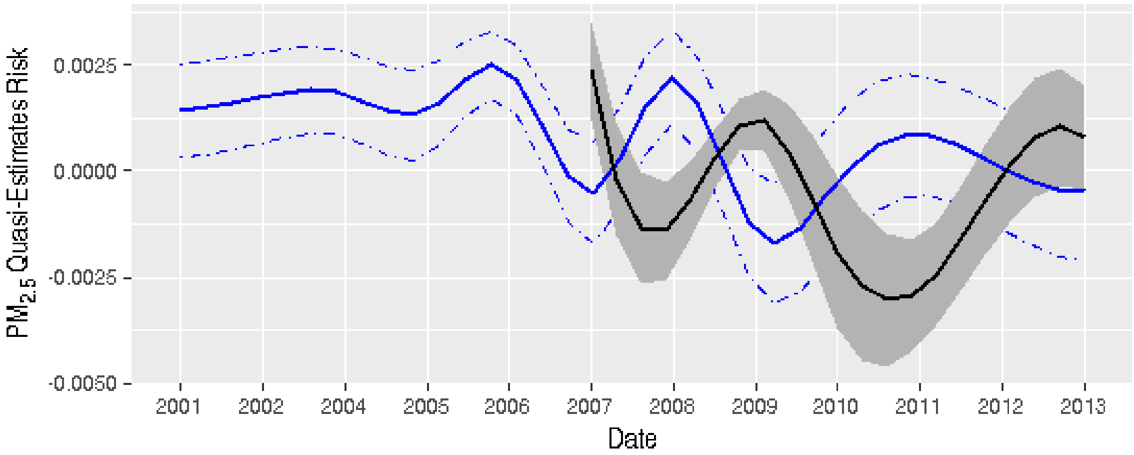


Figure 6.3: $PM_{2.5}$ *population health risk (average)* versus *population health risk (population weighted)*, blue and black curves respectively.

time period 2007 – 2013. Again if we were to explain this in the traditional statistical sense like NO_2 there is not a statistical difference between both curves (as there confidence intervals overlap, for the most part). Further, the seasonality of the estimated risks appears to be opposing - in 2009 the blue curve is at a minimum, while the black is at a maximum, while in mid 2010-11, the opposite effect occurs, and then in 2012-13, again reversal. This is very curious, and definitely a cause for future study.

6.2 Mortality (cardio-pulmonary deaths)

In this section, the response that was used for this "cardio-pulmonary death risk" include all deaths for the population of our study area that were caused by cardio-pulmonary related deaths, both genders (male and female), and for all ages (births to deaths) of the population.

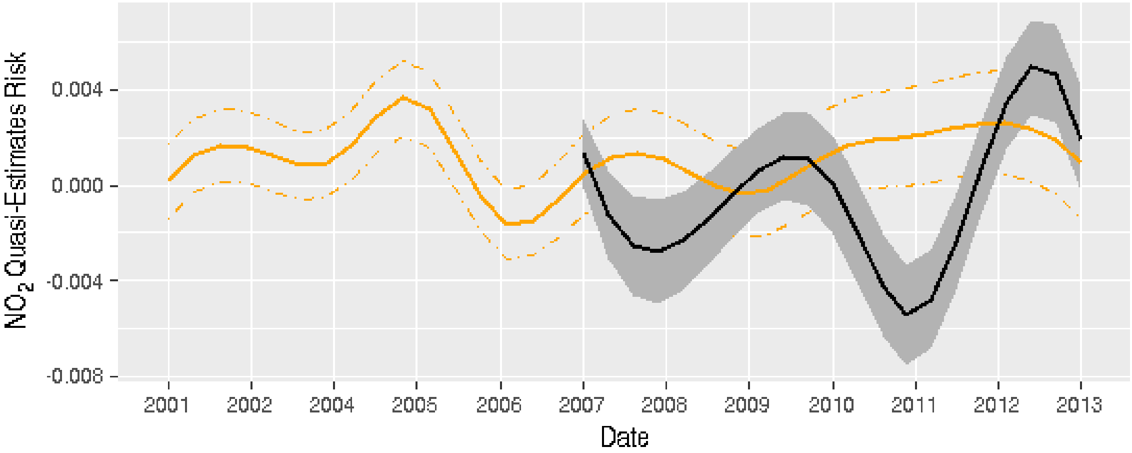


Figure 6.4: *NO_2 population health risk (average)* versus *population health risk (population weighted)*, orange and black curves respectively (cardio-pulmonary related deaths).

The data set that was used was complete in that a complete interpolation was done with bad data points flagged and removed, then interpolated (only for the pollutants). Figures 6.4, 6.5, and 6.6 shows risk plots for NO_2, O_3, and $PM_{2.5}$ respectively. Here

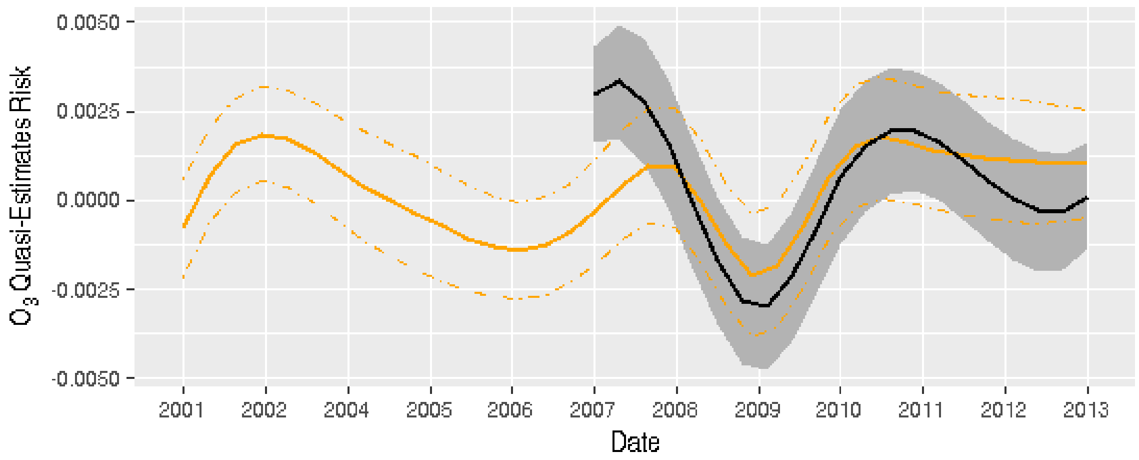


Figure 6.5: O_3 *population health risk (average)* versus *population health risk (population weighted)*, orange and black curves respectively (cardio-pulmonary related deaths).

we use the yearly seasonality January to December (Jan – Dec). The orange solid dash curve in our plot represents the *population health risk (average)* and its confidence interval (with confidence level 95%) respectively. The black line, and the grey fill represents the *population health risk (population weighted)* and its confidence interval (with confidence level 95%) respectively.

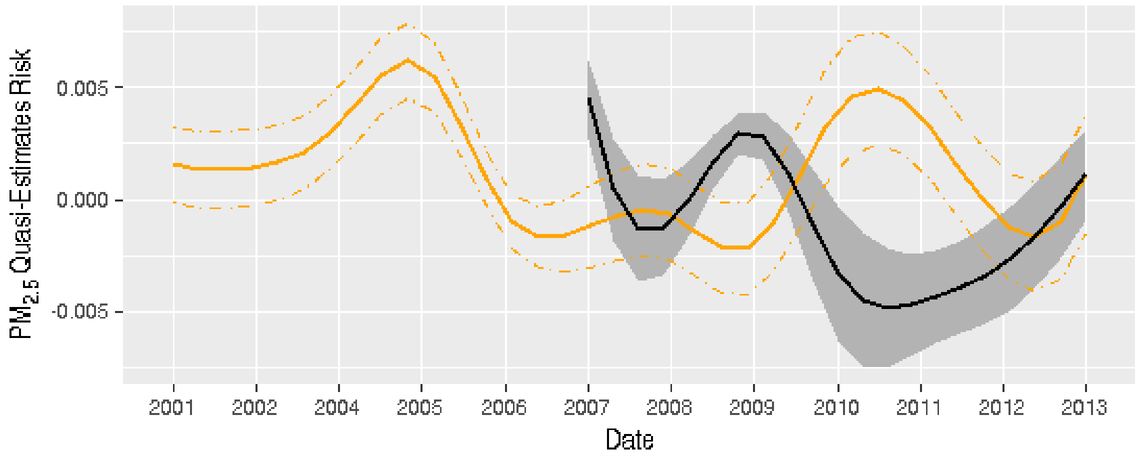


Figure 6.6: $PM_{2.5}$ *population health risk (average)* versus *population health risk (population weighted)*, orange and black curves respectively (cardio-pulmonary related deaths).

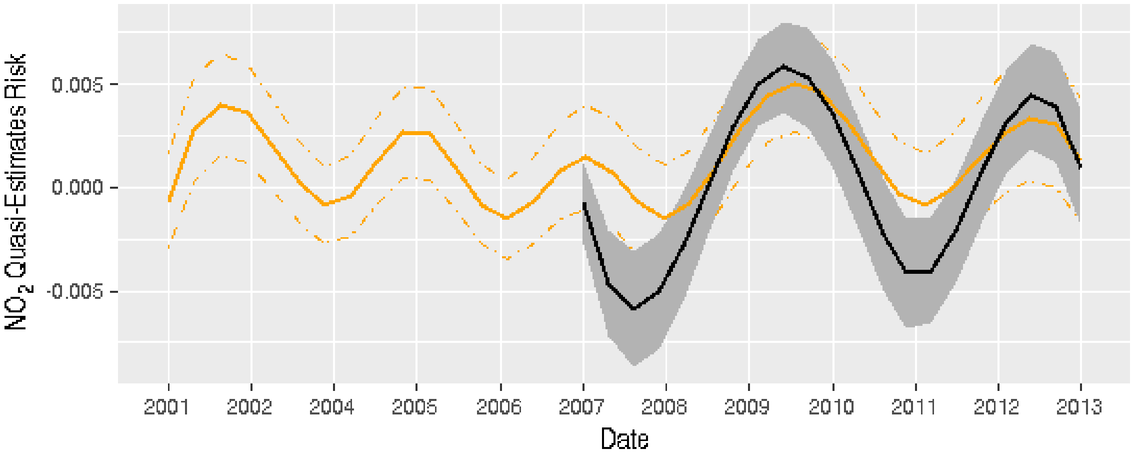


Figure 6.7: *NO_2 population health risk (average)* versus *population health risk (population weighted)*, orange and black curves respectively (cardio-pulmonary related deaths). Oct – Mar seasonality.

In Figures 6.4, 6.5, and 6.6 we can see that there is again differences in structure between the two exposure series, with significant “seasonality” being exhibited in the population-weighted exposure metrics which are not present in the naive average. We also note that there is again some minor anti-phase behaviour, with peaks in the orange curve where troughs exist in the black curve (especially Figure 6.6).

There are a number of years for which the *population health risk (average)* is not within the 95% confidence interval of *population health risk (population weighted)* with the exception of O_3 (see Figure 6.5). This indicates that the changes we have made to the pollutant through the efforts of Chapters 4 and 5 have seemingly strong influences on the risk relationship between pollution and health, something which previously to our knowledge not been noted in the published literature.

We also examined the seasonality, October to March (Oct – Mar), for the risk models. This was done because we know that from numerous previous studies cardio-pulmonary related illnesses are more frequent in the colder months. We present these plots in Figures 6.7, 6.8, and 6.9 for the pollutants NO_2, O_3, and $PM_{2.5}$ respectively. From these plots we can see, that using the colder half of the year for seasonality,

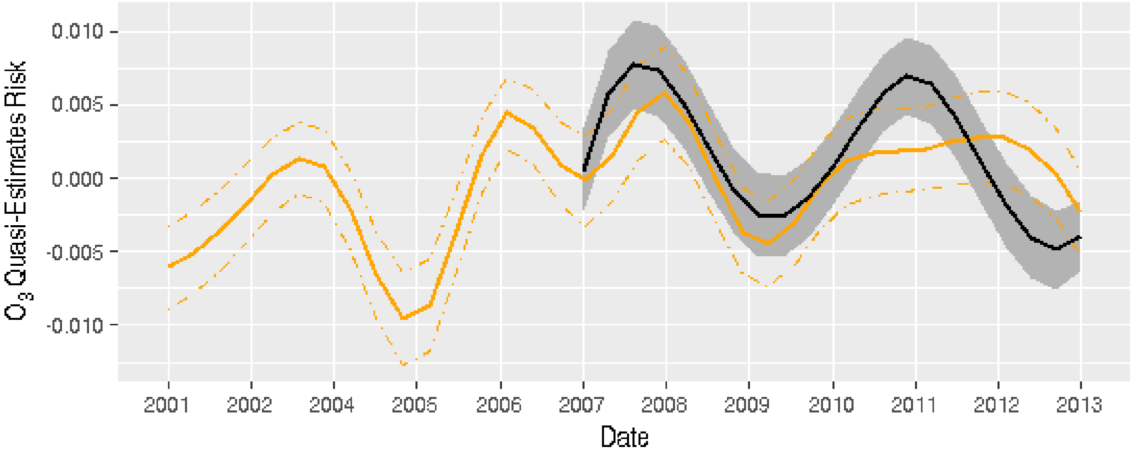


Figure 6.8: O_3 *population health risk (average)* versus *population health risk (population weighted)*, orange and black curves respectively (cardio-pulmonary related deaths). Oct – Mar seasonality.

in a statistical sense the relationship between the risks are stronger (as the curves confidence interval overlapped more than the prior results). In particular, we can see that for NO_2 the yellow and black curve where almost identical. There is also significant increase in the overlap of the confidence intervals for O_3 that is the case when we use the Jan – Dec seasonality. $PM_{2.5}$ however still presented a clear opposing effect in seasonality and again this calls for future study.

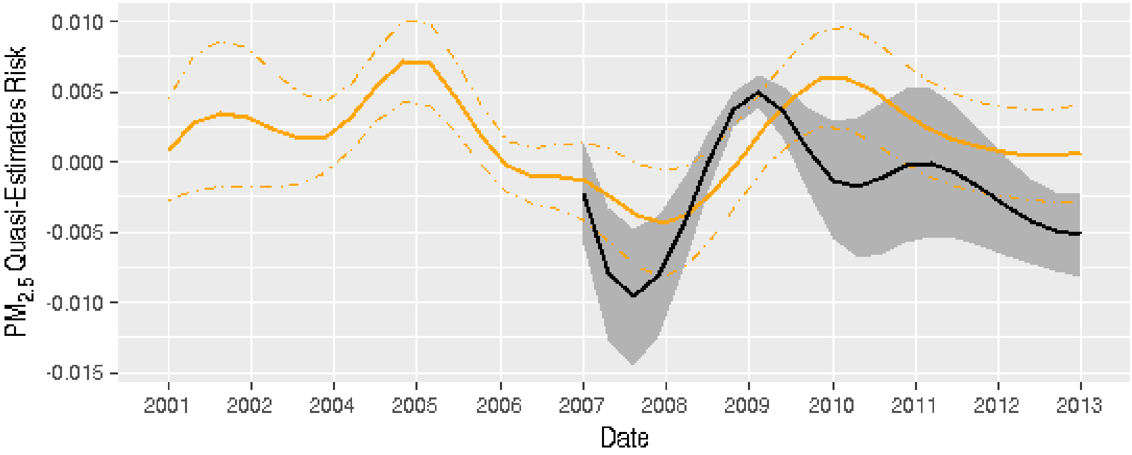


Figure 6.9: $PM_{2.5}$ *population health risk (average)* versus *population health risk (population weighted)* (orange and black curves respectively), cardio-pulmonary related deaths Oct – Mar seasonality.

Overall, the differences in these risk curves are larger than they may appear at first glance. In particular, the fact that the risk curves are different at all is curious: the models are the same, the predictors are the same, and the only difference is the re-calculation and re-weighting of the pollution exposure. However, this seemingly minor change has extremely large effects on the final risk, indicating that the relationships being measured are quite sensitive to this input. More work needs to be done on the stability of the population-weighted estimated pollution exposure in order to determine if the changes being observed are due to specific influences.

Due to scope, we were unable to expand this project beyond Toronto to the rest of Canada, but we feel that this is a fascinating question to try to answer, and hope it can be a topic for future work.

7. *Conclusions and Future Work*

The representative nature of the air population metric averages on the Canadian (Toronto) data set studied in this thesis required careful thought as these reported metrics estimate the true health effects associated with air-quality. A careful determination of the spatial and temporal correlations was implemented employing the INLA spatio-temporal model. From our results, there is a clear disparity between the station-level naive average of pollutants and the spatio-temporal posterior mean field.

Given this disparity that was discussed in Chapter 5, we conclude that the station-level average does not provide a true population-level exposure metric for air pollution. We feel that our INLA spatio-temporal model allowed us compute a more accurate mean for each pollutant, especially in the integration of the local population density and estimation of the field strength across the city. In addition, our computation of a mean field gave the opportunity to predict concentration levels for locations that did not have observations. This was important in modelling the population health risk for the city of Toronto as it provided complete accurate pollutant concentrations in our calculation of the weighted population mean. Areas that had higher population density would not have been properly accounted for with a simple city level average.

In our discussion we saw that there were instances where there the allowable limit for pollutant concentrations were exceeded. These events, though not frequent, showed that air pollution for the city is a real concern. This concern gives rise to health related risk which we investigated in Chapter 6. We found that pollutant concentration levels had an impact on population health. Our investigation of population health risk revealed that of the three pollutants studied NO_2 had the strongest impact.

Expanding this work to Canada wide data set would be desirable as there are areas in Canada where air pollution monitoring is limited. Spatio-temporal analysis would fill in the gaps and give a complete picture of pollutant modelling across the country. We saw that while our model gave accurate predictions between station locations, these predictions are less accurate as we moved farther from these stations. Including observations for stations that are outside of the study area and also stations with sparse data within our study area would aid in improving the accuracy. As an example, including observations for Oshawa would give rise to better prediction for eastern Toronto. In addition there are some additional covariates that would improve our model. As was discussed in Chapter 5 there was not a good continuity in available meteorological data that we used for the study. For future works we could employ methods through the INLA model which could improve the continuity and spatial accuracy of meteorological data (also using stations outside of our study area).

Using available satellite data for future work would also improve our model, this would provide us with better observations to increase prediction. Li *et al.* [19], developed a method for predicting high resolution spatial-temporal air pollutant maps from various data sources. These data sources were varied (they utilised satellite and mobile sensing data which we did not employ in our study), and they had good volume (data was representative for each variety). Going forward we could build on our work by including available data. We can also look at obtaining daily traffic data (not publicly available) at locations along major roadways to include as covariates in our model.

Bibliography

[1] R. W. Allen, C. Carlsten, B. Karlen, S. Leckie, S. V. Eeden, S. Vedal, I. Wong, and M. Brauer. Air filter intervention study of endothelial function among healthy adults in a woodsmoke-impacted community. *American Journal of Respiratory and Critical Care Medicine*, 2011.

[2] M. Blangiardo and M. Cameletti. *Spatial and Spatio-temporal Bayesian Models with R - INLA*. Wiley, United Kingdom, 2015.

[3] W. S. Burr, G Takahara, and H. H. Shin. Bias correction in estimation of public health risk attributable to short-term air pollution exposure. *Environmetrics*, 7 April 2015.

[4] Health Canada. Residential indoor air quality guideline ozone. 2010.

[5] Health Canada. Guidance for fine particulate matter ($pm_{2.5}$) in residential indoor air. 2012.

[6] Health Canada. Residential indoor air quality guideline: Nitrogen dioxide. 2015.

[7] Statistics Canada. Census program, census datasets. `https://www12.statcan.gc.ca/datasets/index-eng.cfm`. Accessed on 2018-01-11.

[8] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, Stanford, California, 2011.

[9] N. A. C. Cressie. *Statistics for spatial data*. Wiley Interscience, Wiley Series in Probability and Statistics, 1993.

[10] R. J. Delfino, P. J. E. Quintana, J. Floro, V. M. Gastanaga, B. S. Samimi,

M. T. Kleinman, L. S. Liu, C. Bufalino, C. F. Wu, and C. E. McLaren. Association of fev1 in asthmatic children with personal and microenvironmental exposure to airborne particulate matter. *Environmental Health Perspectives*, pages 112(8):932–941, 2004.

[11] F. Dominici, J. M. Samet, and S. L. Zeger. Combining evidence on air pollution and daily mortality from the 20 largest us cities: A hierarchical modelling strategy. *Journal of the Royal Statistical Society*, 163 No. 3:263–302, 2000.

[12] Environment and Climate Change Canada. Canadian ambient air quality standards. `http://www.ec.gc.ca/default.asp?lang=En&n=56D4043B-1&news=A4B2C28A-2DFB-4BF4-8777-ADF29B4360BD`, 2013. Accessed on 2018-03-12.

[13] Environment and Climate Change Canada. Environmental sustainability indicators: Air quality. 2016.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, Stanford, California, 2008. Second Edition.

[15] S. B. Henderson, B. Beckerman, M Jerrett, and M Brauer. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental Science & Technology*, pages 41(7), 2422–2428, 2007.

[16] C. E. Housecroft and E. C. Constable. *Chemistry: An Introduction to Organic, Inorganic and Physical Chemistry*. McGraw Hill, Department of Chemistry University of Basel, Switzerland, 2006.

[17] J. Q. Koenig, K. Jansen, T. F. Mar, T. Lumley, J. Kaufman, C. A. Trenga, J. Sullivan, L. S. Liu, G. G. Shapiro, and T. V. Larson. Measurement of offline exhaled nitric oxide in a study of community exposure to air pollution. *Environmental Health Perspectives*, pages 111(13):1625–1629, 2003.

[18] J. Q. Koenig, T. F. Mar, R. W. Allen, K. Jansen, T. Lumley, J. H. Sullivan, C. .

Trenga, T. V. Larson, and L. S. Liu. Pulmonary effects of indoor and outdoor-generated particles in children with asthma. *Environmental Health Perspectives*, pages 113(4):499–503, 2005.

[19] Y. Li, Y Zhu, W. Yin, Y. Liu, G Shi, and Z. Han. Prediction of high resolution spatial-temporal air pollutant map from big data sources. *BigComm, The First International Conference on Big Data Computing and Communication*, pages 9196, 273–282, 2015.

[20] G. R. Liu and S. S. Quek. *The Finite Element Method, A Practical Course*. Butterworth Heinemann, Department of Mechanical Engineering, National University of Singapore, 2003.

[21] L. Liu, T. Ruddy, M. Dalipaj, R. Poon, M. Szyszkowicz, H. You, R.E. Dales, and A.J Wheeler. Effects of indoor, outdoor, and personal exposure to particulate air pollution on cardiovascular physiology and systemic mediators in seniors. *Journal of Occupational and Environmental Medicine*, pages 51(9):1088–1098, 2009.

[22] D. Medkovà. *The Laplace Equation: Boundary Value Problems on Bounded and Unbounded Lipschitz Domains*. Springer, Institute of Mathematics of the Czech Academy of Sciences, 2018.

[23] City of Toronto. City of toronto maps. `http://map.toronto.ca/maps/map.jsp?app=TorontoMaps_v2`, 2018. Accessed on 2018-02-12.

[24] Highway Standards Branch Ontario Ministry of Transportation. Provincial highway traffic volumes, 1988-2016. 1988-2016.

[25] Y. Ramos, St-Onge. B., J. P. Blanchet, and Smargiassi A. Spatio-temporal models to estimate daily concentrations of fine particulate matter in montreal: Kriging with external drift and inverse distance-weighted approaches. *Journal of Exposure Science & Environmental Epidemiology*, pages 26(4), 405–414, 2015.

[26] Y Ramos, W Réquia, B St-Onge, J.-P. Blanchet, Y Kestens, and Smargiassi.

Spatial modeling of daily concentrations of ground-level ozone in montreal, canada: A comparison of geostatistical approaches. *Environmental Research*, page 166(10) 1016, 2018.

[27] F.E. Relton. *Applied Bessel Functions.* Blackie, London, 1946.

[28] H. Rue and H. Tjelmeland. *Fitting Gaussian Markov random fields to Gaussian fields.* Scandinavian Journal of Statistics, United Kingdom, 2002. 29(1): pages 31-29.

[29] K. Sabaliauskas, C.-H. Jeong, X Yao, C. Reali, T Sun, and G. J. Evans. Development of a land-use regression model for ultrafine particles in toronto, canada. *Atmospheric Environment*, pages 110, 84–92, 2015.

[30] J. G. Smith. *Principles of General, Organic, and Biological Chemistry.* McGraw Hill, University of Hawaii at Manoa, 2012. pages 170 - 181.

[31] C.A. Trenga, J. H. Sullivan, J. S. Schildcrout, K. P. Shepherd, G. G. Shapiro, L. S. Liu, J. D. Kaufman, and J. Q. Koenig. Effect of particulate air pollution on lung function in adult and paediatric subjects in a seattle panel study. *Chest*, pages 129(6):1614–1622, 2006.

[32] S Weichenthal, K. V. Ryswyk, A Goldstein, M Shekarrizfard, and M. Hatzopoulou. Characterizing the spatial distribution of ambient ultrafine particles in toronto, canada: A land use regression model. *Environmental Pollution*, pages 208(A), 241–248, 2016.

A. *R Code used for data development and model computation*

Check this appendix

```
Conv_longlat_XY <- function(data){
  library(sp)
  colnames(xy) <- c(’lon’, ’lat’)
  coordinates(xy) <- ~ lon + lat
  proj4string(xy) <- CRS("+proj=longlat +datum=WGS84")
  p <- spTransform(xy, CRS("+proj=tmerc +lat_0=0 +lon_0=-79.5
                           +k=0.9999 +x_0=304800 +y_0=0
                           +datum=NAD27 +units=m +no_defs
                           +ellps=clrk66 +nadgrids=@conus,
                           @alaska,@ntv2_0.gsb,@ntv1_can.dat"))
  tran_cood <- coordinates(p)
  colnames(tran_cood) <- c(’x’, ’y’)
  XY_Coords <- cbind(tran_cood[,1], tran_cood[,2])
  XY_Coords
}
```

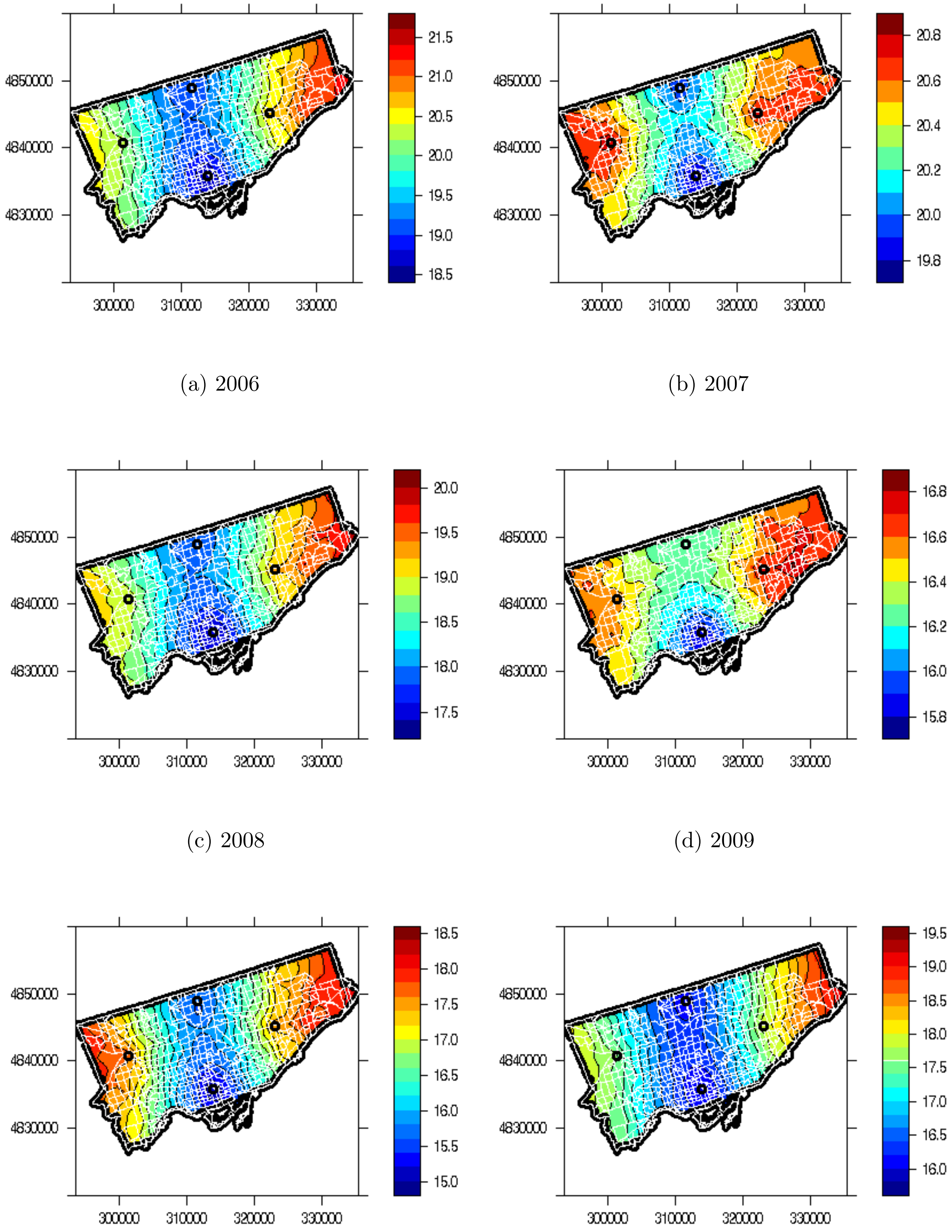
Figure A.1: Function used to convert longitude, latitude to UTMX, UTMX

```
TO_Mean <- function(x1, x2, days, data){
  # Call data to be used
  TO_Data <- data[x1:x2,]
  Fix_TO_Data <- TO_Data[6:7]
  Fix_TO_Data[] <- lapply(Fix_TO_Data,
                          function(x) ifelse(is.na(x),
                                             mean(x,
                                                  na.rm = TRUE), x))
  TO_Data$WS <- Fix_TO_Data$WS
  TO_Data$WD <- Fix_TO_Data$WD
  # Data for calc
  Toronto_Data <- TO_Data
  # Prepare Borders for Toronto
  TO_border <- TO_borderN

  TO_coords <- CD_Toronto_X[c(3, 8, 12, 13), 1:3]
  colnames(TO_coords) <- c("Station.ID", "UTMX", "UTMY")

  # Prepare Data for Pollutant
  n_stations <- length(TO_coords[,1]) #number of stations (data sparced)
  n_data <- length(Toronto_Data[,1]) #number of space-time data
  n_days <- n_data/n_stations #number of time points

  rownames(TO_coords) <- 1:n_stations
  TO_coords$Station.ID <- 1:n_stations
  Toronto_Data$Station.ID <- rep(1:n_stations, n_days)

  Toronto_Data$time <- rep(1:n_days, each = n_stations)
  coords.allyear <- as.matrix(TO_coords[Toronto_Data$Station.ID,
                                        c("UTMX","UTMY")])

  # Standardize covariates and calculate log pollutant
  Toronto_Data$logNO2 <- log(Toronto_Data$NO2)
  mean_covariates <- apply(Toronto_Data[, 3:9], 2, mean, na.rm = TRUE)
  sd_covariates <- apply(Toronto_Data[, 3:9], 2, sd, na.rm = TRUE)
  Toronto_Data[, 3:9] <- scale(Toronto_Data[, 3:9],center = mean_covariates,
                               scale = sd_covariates)
  # Altitude standardized by factor of 1/100 as study are is flat
  Toronto_Data$A <- Toronto_Data$A/100
```

Figure A.2: Mean Field Function

Function used for computing Posterior mean field

B. *Posterior mean field plots for Toronto*



(a) 2006

(b) 2007

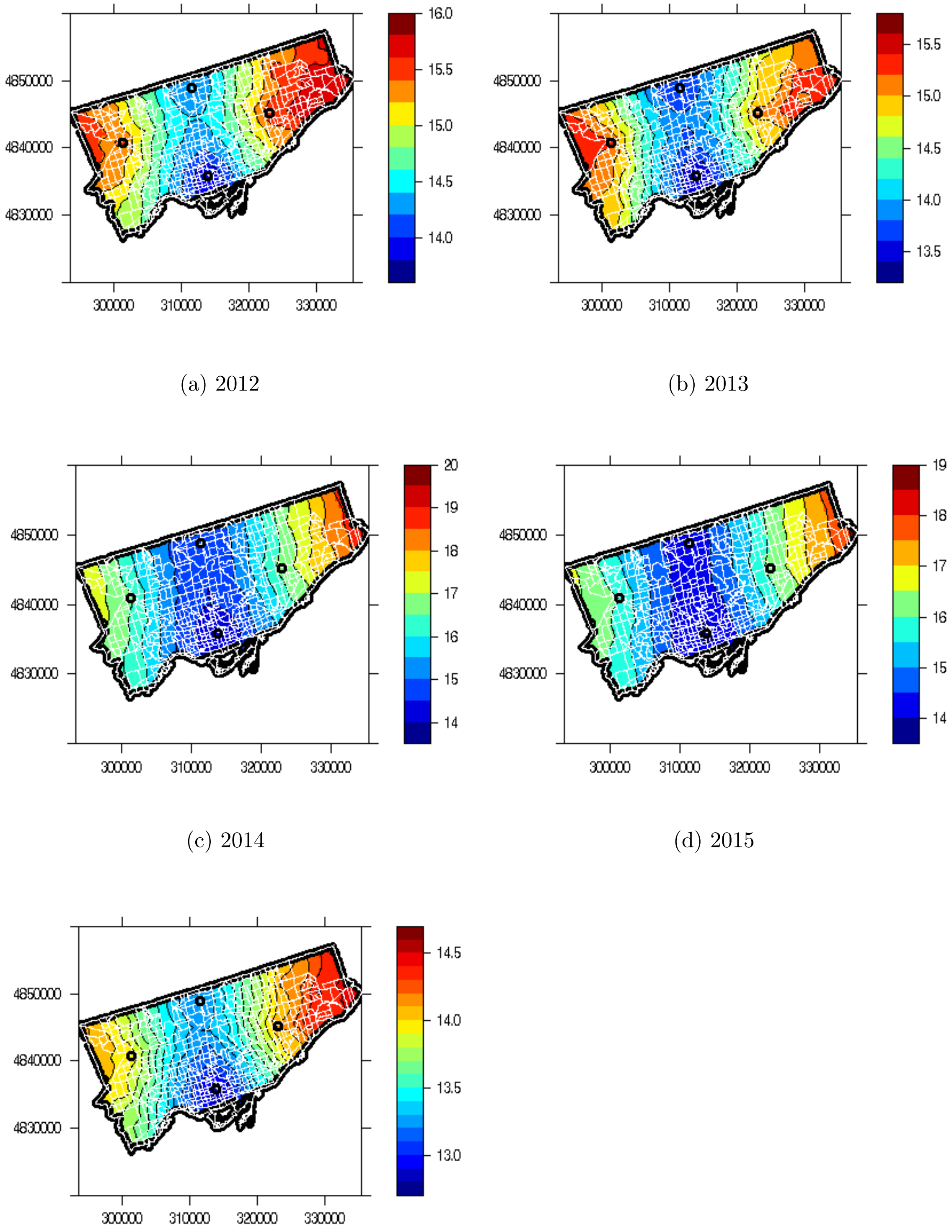
(c) 2008

(d) 2009

(e) 2010

(f) 2011

Figure B.1: NO_2 Spatial mean field (2006 - 2011)



(a) 2012

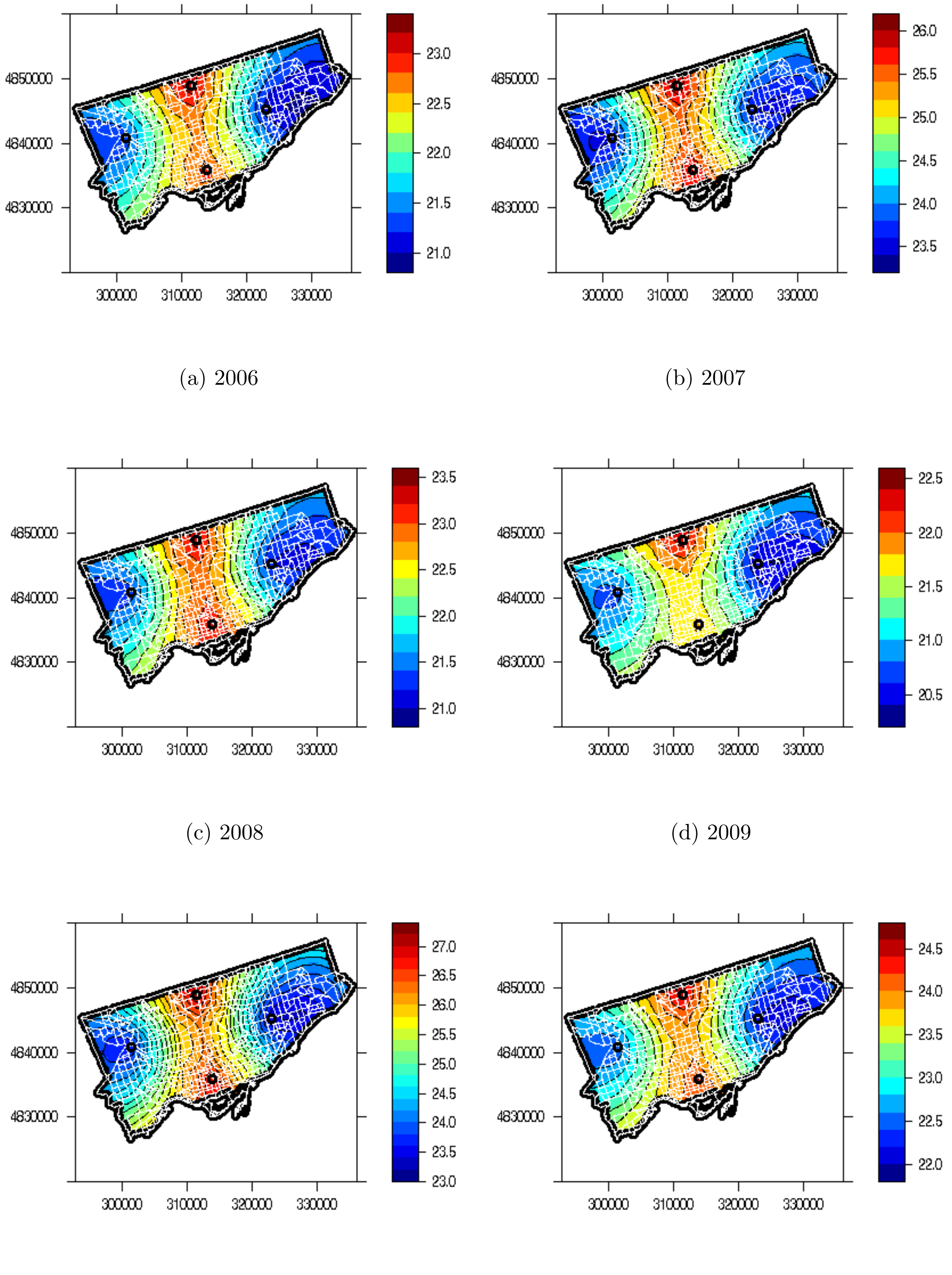
(b) 2013

(c) 2014

(d) 2015

(e) 2016

Figure B.2: NO_2 Spatial mean field (2012 - 2016)



(a) 2006

(b) 2007

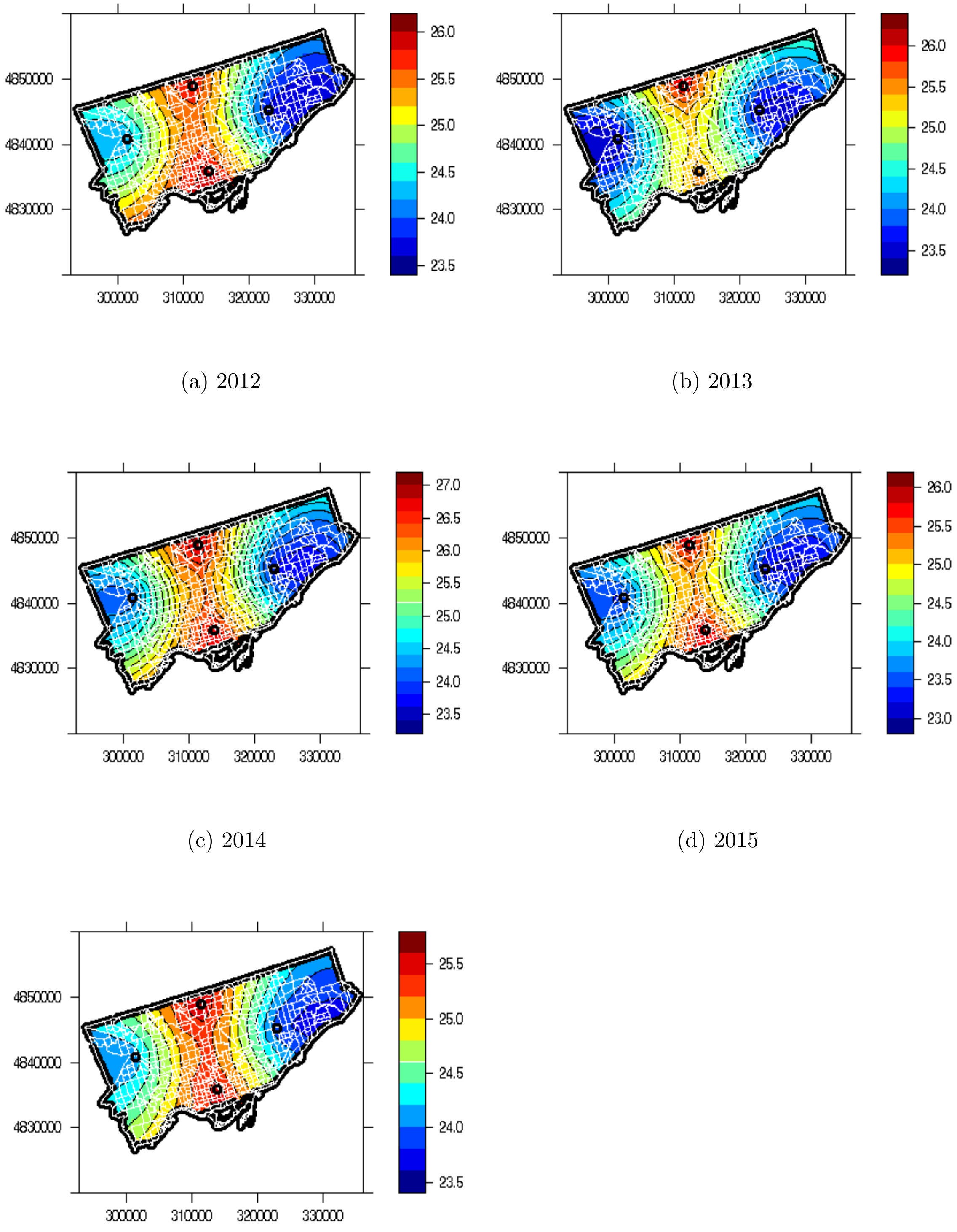
(c) 2008

(d) 2009

(e) 2010

(f) 2011

Figure B.3: O_3 Spatial mean field (2006 - 2011)



(a) 2012

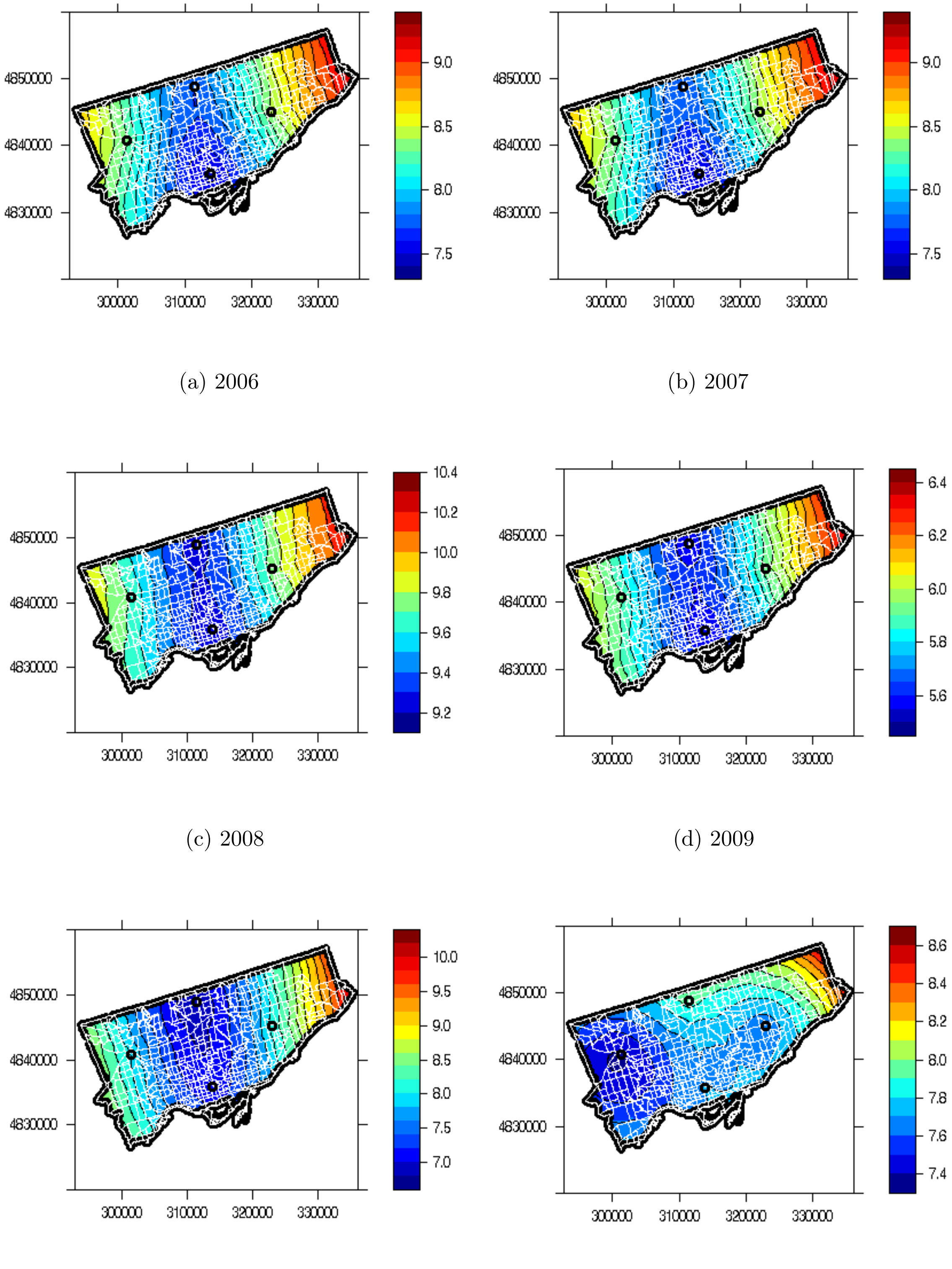
(b) 2013

(c) 2014

(d) 2015

(e) 2016

Figure B.4: O_3 Spatial mean field (2012 - 2016)



(a) 2006

(b) 2007

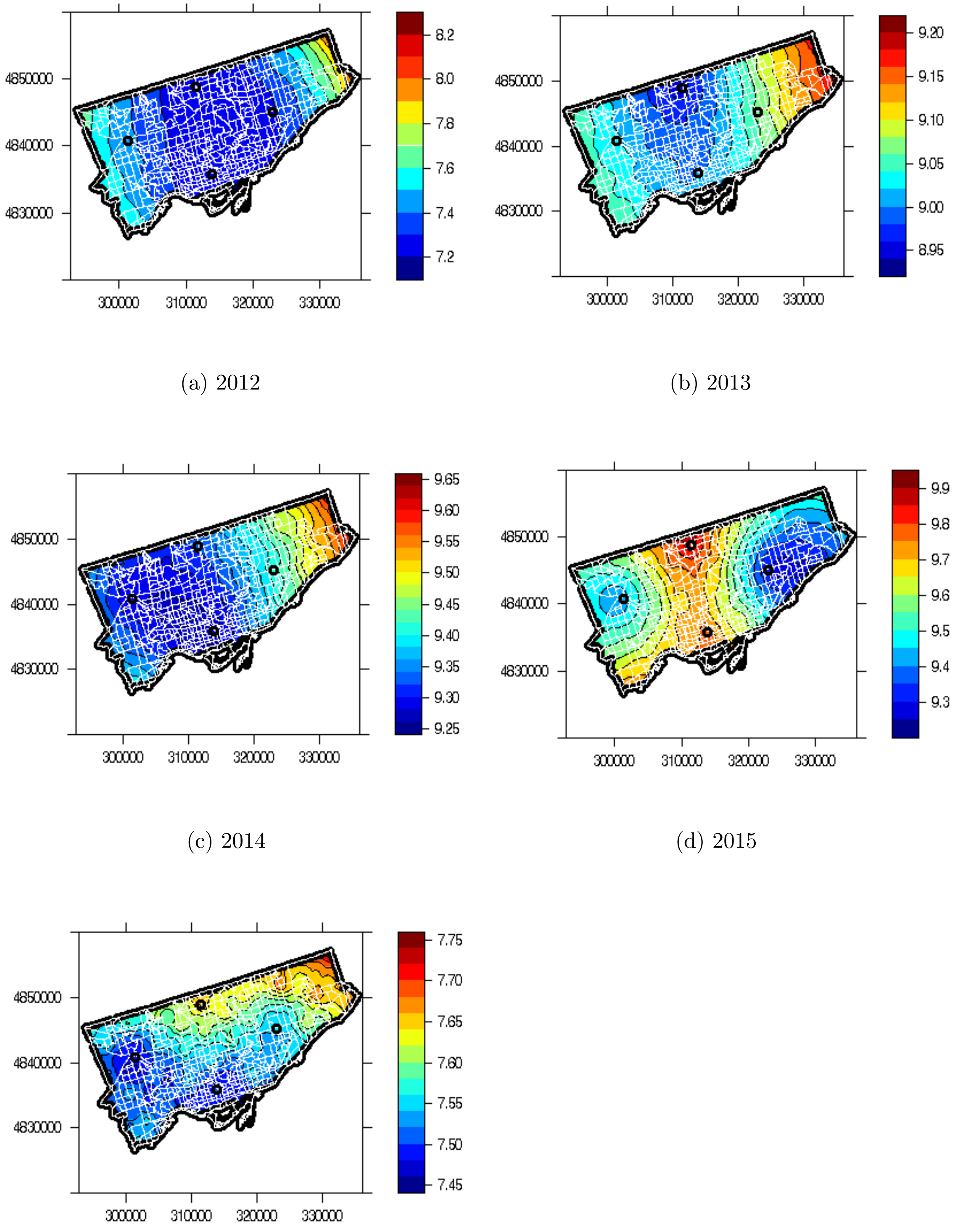
(c) 2008

(d) 2009

(e) 2010

(f) 2011

Figure B.5: $PM_{2.5}$ Spatial mean field (2006 - 2011)



(a) 2012

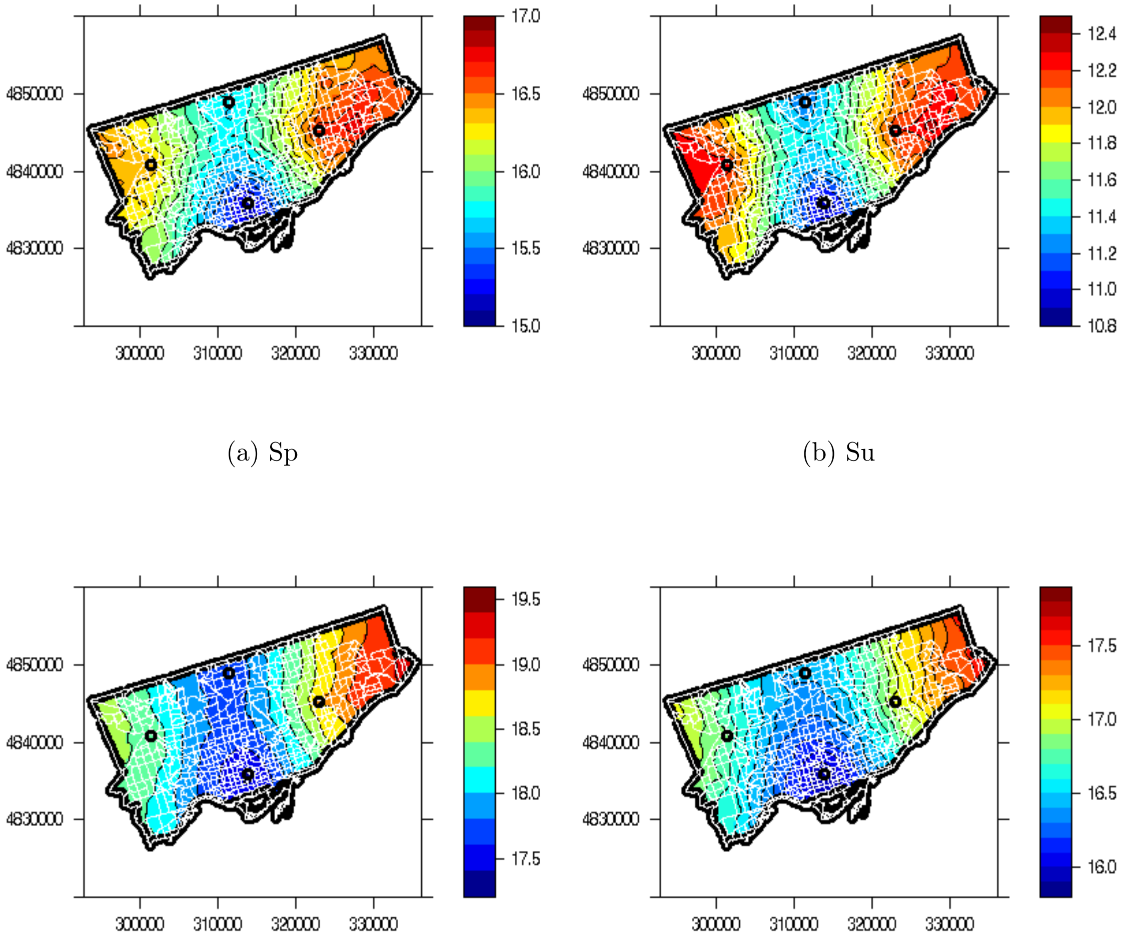
(b) 2013

(c) 2014

(d) 2015

(e) 2016

Figure B.6: $PM_{2.5}$ Spatial mean field (2012 - 2016)



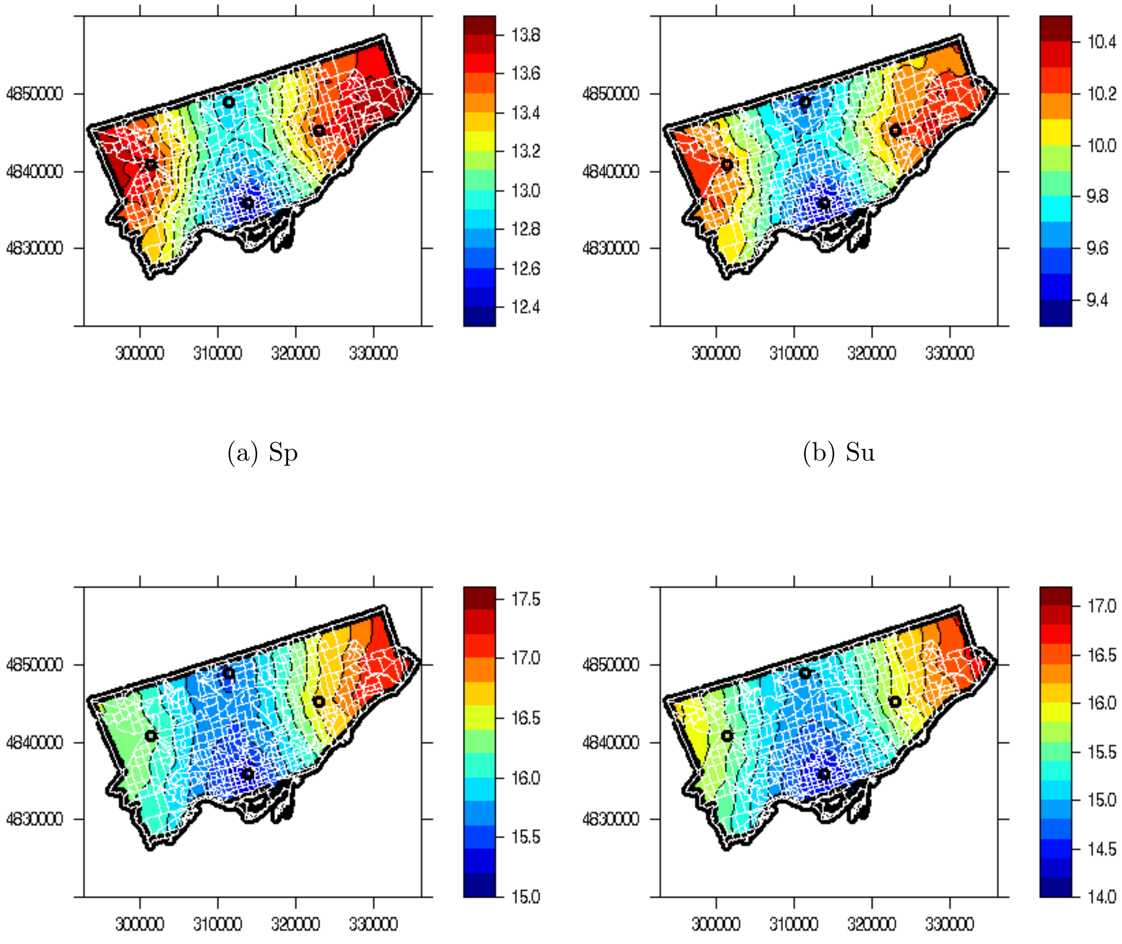
(a) Sp

(b) Su

(c) Fl

(d) Wr

Figure B.7: NO_2 Spatial mean field by seasons 2006



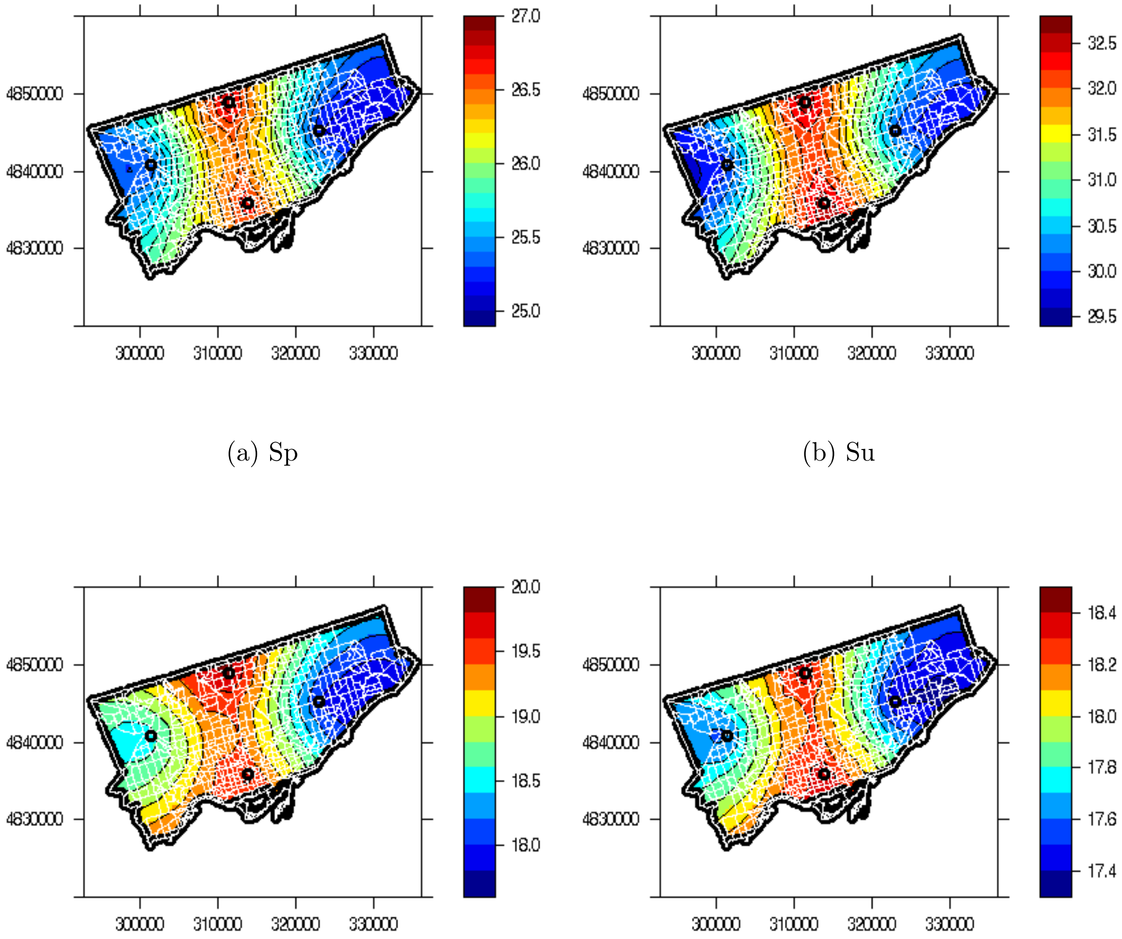
(a) Sp

(b) Su

(c) Fl

(d) Wr

Figure B.8: NO_2 Spatial mean field by seasons 2016



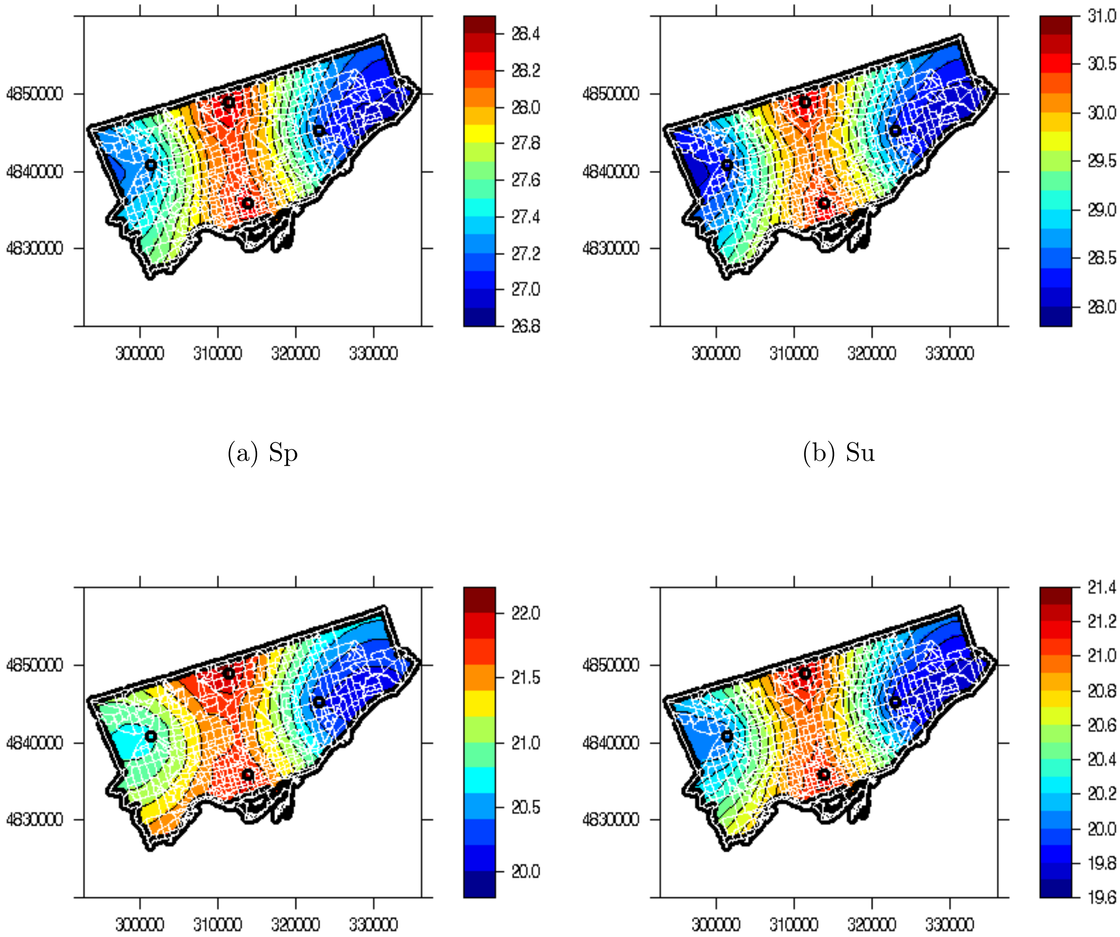
(a) Sp

(b) Su

(c) Fl

(d) Wr

Figure B.9: O_3 Spatial mean field by seasons 2006



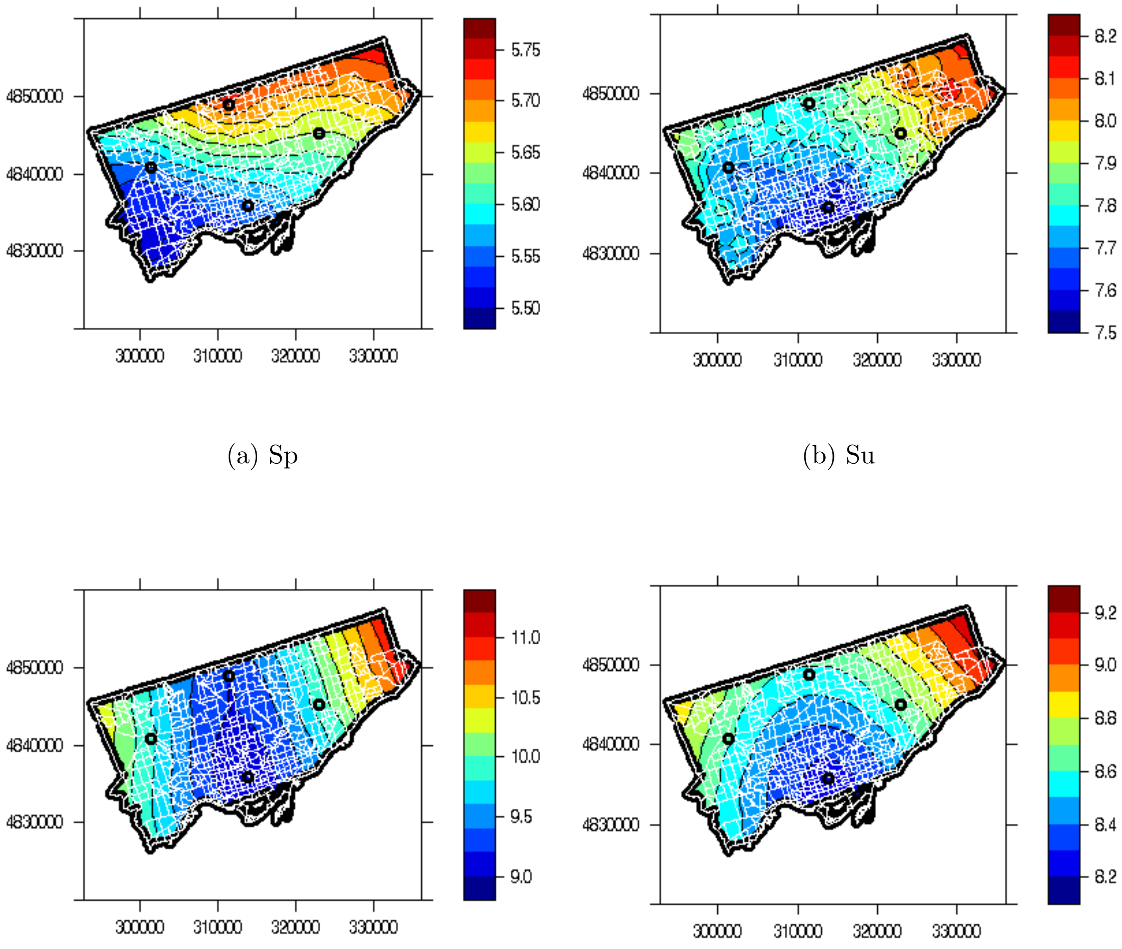
(a) Sp

(b) Su

(c) Fl

(d) Wr

Figure B.10: O_3 Spatial mean field by seasons 2016



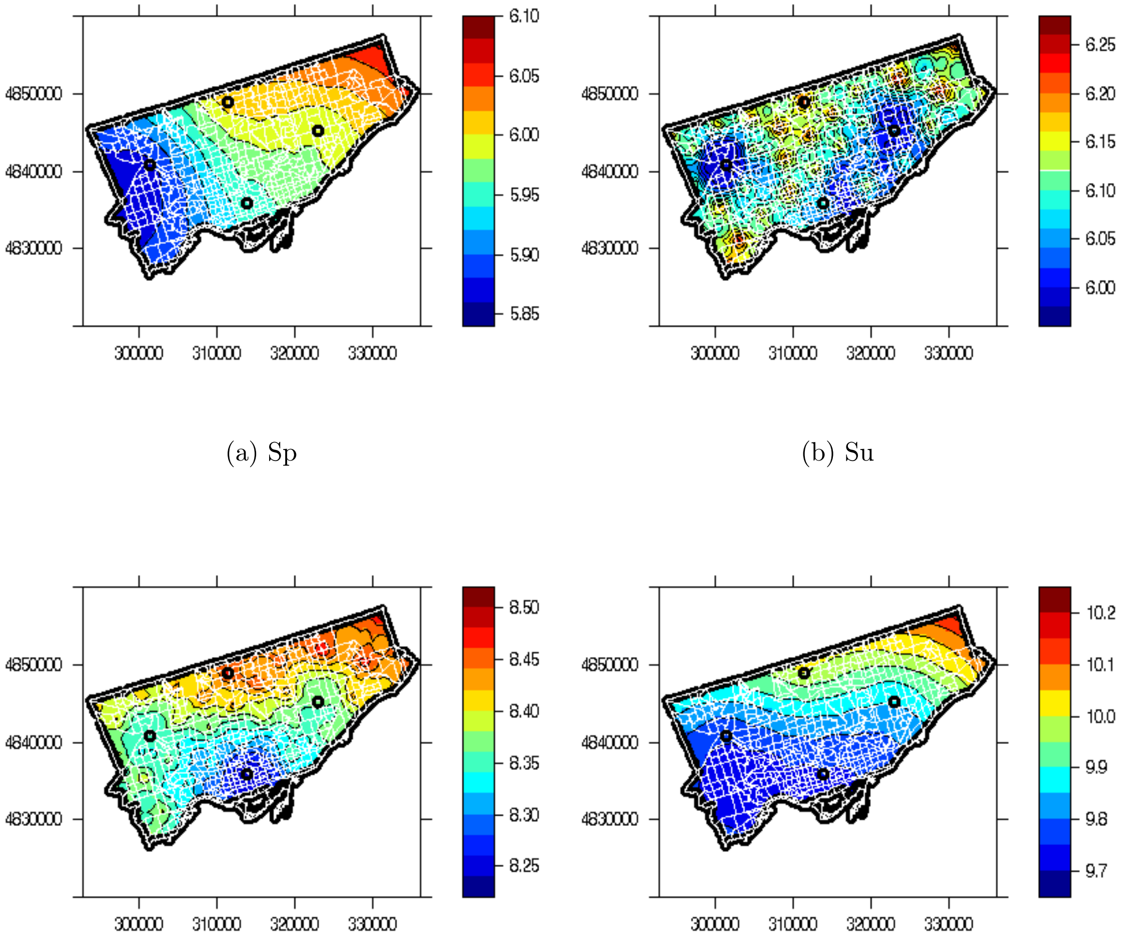
(a) Sp

(b) Su

(c) Fl

(d) Wr

Figure B.11: $PM_{2.5}$ Spatial mean field by seasons 2006



(a) Sp

(b) Su

(c) Fl

(d) Wr

Figure B.12: $PM_{2.5}$ Spatial mean field by seasons 2016