LAND-USE REGRESSION AND SPATIO-TEMPORAL HIERARCHICAL MODELS FOR ENVIRONMENTAL PROCESSES

By Sara Zapata-Marin

Quantitative Life Sciences Program

McGill University, Montreal

April, 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

Contents

List of Figures and Tables

Chapter 3

Chapter 4

Chapter 5

Appendix A

Appendix B

Appendix C

Abstract

Land-use regression is a popular method used to describe the spatial variability of different environmental processes using local variables. However, there are situations in which there might be some complex spatio-temporal structure left after accounting for land-use variables.

In this work, three different Bayesian hierarchical models are proposed to model the spatial and spatio-temporal dispersion of air pollutants and aeroallergens within cities. Bayesian inference can easily accommodate complex interactions while naturally accounting for uncertainties in the estimation of unknowns in the model when performing predictions.

In the first study, a spatial hierarchical model is used to analyze the concentration of volatile organic compounds (VOCs) in Montreal, Canada. The data consists of concentration measurements of five VOCs measured over two-week periods for three monitoring campaigns between 2005 and 2006 over 130 locations in the city. The five VOCs of interest are: benzene, decane, ethylbenzene, hexane, and trimethylbenzene. Four different models are fitted to each of the five VOCs. These models extend land-use regression by accounting for any spatial structure left after including the covariates while also capturing the across campaign variation through an indicator variable or campaign-specific coefficients. Predicted surfaces are obtained for each campaign. For all VOCs higher levels are found during the December campaign, and the predicted areas with the highest levels correspond to multiple sections of major highways.

For the second and third studies, we have available data on the daily and weekly measurements of pollen concentration in Toronto, Canada collected in 2018. The measurements consist of tree, weed, grass, and total pollen concentration at 18 monitoring sites and were obtained daily for eleven of these sites and weekly for the other seven sites.

In the second study, the weekly concentration of each of the four pollen types is modeled. Instead of considering the temporal window that only has positive values, that is, removing the zeros, a hurdle model is proposed to account for the high number of measurements equal to zero. This structure allows for the estimation of the probability of the pollen concentration being equal to zero at any given week, which provides further information on temporal windows with positive concentrations of the different types of pollen. Additionally, a dynamic linear model is used to capture the weekly trend of pollen concentration in the city.

In the third study, the daily concentration of total pollen is modeled. Rather than aggregating the data to the weekly scale, a temporal misalignment model is proposed to account for the difference in scale and to take advantage of the daily measurements. Using the properties of dynamic linear models and the multivariate normal distribution, a spatio-temporal model to account for temporal misalignment is proposed. This model allows to estimate the fine-scale measurements at locations where only coarse-scale observations were available. Additionally, the model is fitted to artificial data with different temporal structures, including trend and seasonality.

The predicted surfaces obtained in these three studies will help inform future health-related studies. Furthermore, the methods proposed here are flexible, easily adaptable, and can improve our understanding of similar environmental processes. All codes are publicly available such that the implementation of the proposed approach in similar situations is easily achieved.

Abrégé

La modélisation de l'occupation de l'espace par la régression est une méthode très répandue et utilisée pour décrire la variabilité spatiale de différents processus environnementaux à l'aide de variables locales. Cependant, une structure spatio-temporelle complexe subsiste après l'ajustement aux variables d'occupation de l'espace.

Dans cette thèse, trois modèles hiérarchiques bayésiens sont proposés pour modéliser la dispersion spatiale et spatio-temporelle des polluants et allergènes aériens au sein de villes. L'inférence bayésienne peut facilement tenir compte des interactions complexes tout en tenant compte naturellement, lors de prédictions, du niveau d'incertitude lié à l'estimation d'inconnues dans un modèle.

Dans le premier projet, un modèle spatial hiérarchique est utilisé pour analyser la concentration de composés organiques volatils (COVs) à Montréal, au Canada. Les données sont constituées de mesures de concentrations de cinq COVs relevées au cours de périodes de deux semaines lors de trois campagnes de surveillance sur 130 sites urbains entre 2005 et 2006. Les cinq COVs d'intérêts sont le benzène, le décane, l'éthylbenzène, l'hexane et le triméthylbenzène. Quatre modèles différents sont considérés pour les cinq COVs. Ces modèles développent la modélisation de l'occupation de l'espace par la prise en compte de toute structure spatiale qui subsiste après l'inclusion de variables explicatives; et par leur considération de la variation inter-campagne au moyen d'une variable indicatrice ou de coefficients propres à chaque campagne de surveillance. Pour tous les COVs, les niveaux les plus élevés sont relevés au cours d'une campagne en décembre et les zones prédites avec les plus hauts niveaux correspondent à plusieurs sections de principales autoroutes.

Pour les deuxième et troisième projets, nous utilisons des données quotidiennes et hebdomadaires liées à la concentration de pollen mesurée à Toronto, au Canada en 2018. Les mesures concernent la concentration de pollen d'arbre, d'herbes, de gazon et total sur 18 sites de surveillance, qui ont été relevées quotidiennement sur onze de ces sites et hebdomadairement sur les sept autres.

Dans le deuxième projet, la concentration hebdomadaire de chacun des quatre types de pollen est modélisée. Au lieu de ne considérer que la période où les valeurs sont positives, c'est-à-dire d'enlever les zéros, un modèle hurdle est proposé afin de considérer les nombreuses mesures égales à zéro. Cela permet l'estimation de la probabilité que la concentration de pollen soit nulle, ce qui donne davantage d'information sur les périodes avec des concentrations positives des différents types de pollen. De plus, un modèle linéaire dynamique est utilisé pour représenter la tendance hebdomadaire de la concentration de pollen en ville.

Dans le troisième projet, la concentration quotidienne du pollen total est modélisée. Plutôt que d'agréger les données à une échelle hebdomadaire, un modèle de désalignement temporel est proposé afin de considérer les différences d'échelle et afin de profiter des mesures quotidiennes. En utilisant les propriétés des modèles linéaires dynamiques et de la distribution normale multivariée, un modèle spatio-temporel qui tient compte du désalignement temporel, est proposé. Ce modèle estime des mesures à une échelle précise à des localisations où seules des mesures d'échelle plus grossières étaient relevées. De plus, ce modèle est ajusté à des données artificielles présentant différentes structures temporelles telles qu'une tendance générale et une saisonnalité.

Les surfaces prédites dans ces projets aideront à façonner de futures études en santé. D'autre part, les méthodes proposées ici sont flexibles, facilement adaptables et peuvent aider à améliorer notre compréhension de processus environnementaux similaires. Tous les codes sont disponibles publiquement afin que l'implémentation de l'approche proposée dans des situations similaires puisse facilement être implémentée.

Acknowledgements

I would like to thank my supervisor Dr. Alexandra M. Schmidt for her support, guidance, and patience during this process. The dedication and passion that you put into your work and each one of your projects are contagious. With you, I not only learned about spatial statistics, but I became a better researcher and even a better person. I still have so much to learn, but it has been an absolute pleasure to work with you throughout these years.

I thank Dr. Erica Moodie for her support and great advice as a committee member during these five years.

I am also grateful to Dr. Eric Lavigne, and Dr. Scott Weichenthal for providing me with the data, and for allowing me to work with them on the pollen project. I have always appreciated your advice, support, and encouragement.

I would like to thank Dr. Mark Goldberg for providing me with the VOCs data, and for your infinite patience and support. I have learned so much from working with you on this project, and I will always be grateful for having had the opportunity to work with such an amazing and comprehensive researcher.

I am grateful to Daniel S.W. Katz, Tim Takaro, and Jeffrey Brook for their insightful advice while working on the pollen project. Also, I would like to thank Liu Sun for her assistance with the GIS data for this project.

I would also like to thank Dan Crouse, Vikki Hod, France Labrèche, and Marie-Élise Parent for their useful insights on the VOCs project. I am also grateful to Marianne Hatzopoulou and Jad Zalzal for extracting the variables for this project.

I wish also to thank Selin Jessa, Alex Diaz-Papkovich, Jeffrey Hyacinthe, Matt D'Iorio, and Yixiao Zeng, for making me laugh and going through part of this process with me. I would also like to thank Victoire Michal for helping me translate the abstract and, together with Dirk Douwes-Schultz, for your support and help. I am also grateful for the backstage support from Marco A. Rodríguez.

I want to thank my program director, Dr. Celia Greenwood, for her guidance and help throughout these five years. Even with a busy schedule and your many academic responsibilities, you have always been supportive and you have always helped each of the students in the program to succeed. I also want to thank my amigo and program coordinator Alexander DeGuise. Thank you for all your help navigating all the administrative steps and for making my Ph.D. experience much more fun. I am sure that the success of the QLS program is thanks to the hard work and dedication of both of you.

I am so grateful for my family and friends, without them this would have never been possible. To Myriah Haggard and Mariana Carmona, thank you for always being there for the good times and also the bad times, this wouldn't have been possible without your unconditional support. I will always be in debt to you.

To my mom and dad, none of this would have been possible without your help. Thank you for always encouraging me to do what I love and for teaching me to do my best even through the tough times. To my brother, Santiago, thank you for always being there for me, I have always learned so much from you.

Finally, I would like to acknowledge my sponsors CONACYT, COMECYT, AMEXCID, and FRQNT.

Contribution to original knowledge

The work presented in this thesis proposes possible extensions to land-use regression using spatial and spatio-temporal Bayesian hierarchical models to understand the distribution of environmental processes and to make these types of models more accessible to the environmental health audience.

This approach provides additional information on the two processes under study, namely, the spatial distribution of volatile organic compounds and pollen within urban settings. For example, Bayesian hierarchical models in the study of air pollutants can help describe the spatial distribution and the variation across campaigns for different types of air pollutants. When analyzing the pollen concentration, by considering a time-varying mean component, it is possible to identify the period of the pollen season when higher concentrations of certain types of pollen are present. Finally, accounting for the temporal misalignment in the pollen concentration data enables estimating the measurements for all sites at the finer temporal scales and shows how, depending on the temporal scale considered, different associations between pollen concentration and environmental variables can be found.

The methods shown here are flexible and easy to implement for any other air pollutant or aeroallergen within any city. Finally, the results obtained by this work can provide reliable exposure estimates to better inform future environmental health studies.

Contribution of Authors

The present thesis consists of three manuscripts covering topics of spatial and spatio-temporal hierarchical modeling in the study of the distribution of volatile organic compounds in Montreal, QC, Canada (Chapter 3), the weekly distribution of grass, tree, weed, and total pollen in Toronto, ON, Canada (Chapter 4), and the daily distribution of temporally misaligned measurements of total pollen in Toronto, ON, Canada (Chapter 5).

The authors of each manuscript and their contributions vary from one manuscript to another; therefore, the contributions of each co-author are listed at the beginning of each chapter.

1 Introduction

Environmental processes such as air and water pollution, and climate change, can negatively impact individual and population health. It has been estimated that 24% of the global deaths in 2016 were attributable to modifiable environmental risks [1]. While some of these events may have natural causes, human damage to the environment has also caused negative consequences at different scales.

1.1 Air pollutants: Volatile Organic Compounds

Air pollution is the largest environmental risk for human health [2], accounting for an estimated 4.2 million deaths worldwide every year [3]. Exposure to air pollutants can result in various negative health effects such as respiratory illnesses, cardiovascular diseases, allergic reactions, central nervous system dysfunctions, and cancer [4]. These negative health effects depend on the amount and duration of exposure, as well as the type of pollutant, sources, and accumulation over time [5].

Recently, the World Health Organization (WHO) has updated the global quality guidelines due to evidence of adverse health effects at lower concentrations than previously expected [6]. Additionally, it has been estimated that more than 99% of the global population lives in a place where the air pollution exceeds the WHO guideline limits, with higher exposures in mid-and low-income countries [3].

Pollutants that negatively impact air quality include particulate matter ($PM_{2.5}$, PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), and different types of volatile organic compounds (VOCs). Relevant VOCs classes include alkanes such as ethane and

propane; aromatic compounds such as benzene, ethylbenzene, styrene, toluene, and trimethylbenzenes; and oxygenated VOCs such as formaldehyde, oxygenated aromatics, and acetone [7].

To model and control the negative impact of air pollutants on human health, it is necessary to understand their sources, chemical transformations, spatial distribution, and relationship with other environmental factors [8]. Exposure assessment methods focus on the amount, duration, and intensity of exposure to a substance as well as the routes and pathways of exposure [9]. These methods allow epidemiological studies to estimate the concentration of substances to which human populations might be exposed. Land-use regression methods are the most common exposure assessments used to study airborne pollutants such as VOCs [10].

1.2 Aeroallergens: Pollen

Allergic illnesses associated with exposure to aeroallergens include atopic dermatitis, asthma, and allergic rhinitis [11]. The development and severity of these allergic reactions depend on the host and the environmental factors to which the host is exposed, such as specific allergens or other environmental conditions [12].

Environmental exposure to air pollutants and aeroallergens such as pollen, has been suspected to impact the prevalence of allergenic conditions [13]. Additionally, the interaction between these two can also increase asthma symptoms' severity and influence the sensitivity to aeroallergens [11, 14, 15]. Furthermore, living in urban areas is also a known risk factor for developing respiratory allergies due to exposure to aeroallergens [16-18].

Environmental factors such as wind, humidity, precipitation, urbanization, dust, and CO_2 levels can affect when, how and how many pollen grains are released into the environment, as well as their spatial distribution within a region [19]. Climate change can also increase the production and abundance of aeroallergens in many ways. For example, an increase in global temperatures might affect the timing and length of growing seasons, plant distribution, and aeroallergens' dispersion patterns [20, 21] which can lead to an increase in sensitization rates and the severity of allergic reactions [14, 20].

Understanding the spatial variation of air pollutants and aeroallergens can help identify hot spots, show changes in spatio-temporal patterns of pollution, demonstrate compliance with local regulations, identify priorities for targeted environmental action, inform epidemiological studies and provide management recommendations in the case of allergenic pollen-producing plants [1, 22, 23].

To further reduce the exposure to air pollutants and aeroallergens, making fundamental changes to today's practices and implementing evidence-based policies is crucial [24]. This work proposes some methods to predict environmental processes of interest at different temporal scales, therefore providing reliable exposure estimates for future health-related studies.

1.3 Overview of the chapters

This thesis is organized as follows. Chapter 2 provides a literature review on aeroallergens and air pollutants modelling, and Bayesian hierarchical models and their application in environmental processes.

Chapter 3 analyzes the spatio-temporal distribution of VOCs in Montreal, Quebec, Canada. The main objective of this study is to predict the concentration at unobserved locations of five VOCs: benzene, n-decane, ethylbenzene, hexane, and 1, 2, 4-trimethylbenzene.

Variations of a general spatio-temporal model are fitted to each VOC to find the model that better describes the data in each case. This model extends the usual land-use regression by incorporating parameters to capture the variation across space and across monitoring campaigns. The predicted surfaces obtained in this study will be used in the future to investigate possible associations between exposure to VOCs and the relative risk of breast and prostate cancer.

At the time of writing this thesis, this work has been submitted to the *Environmental Epidemiology* journal.

In Chapter 4, the spatio-temporal distribution of pollen concentration is modeled. The data consist of weekly measurements of grass, weed, and tree pollen collected at 17 sites across Toronto, Canada.

The main objectives for this study are,

a) to estimate the concentration of different types of pollen across the sampled season;
b) to identify local predictors that might be associated with the intra-urban variation of the different types of pollen;
c) to capture the temporal trend across weeks over the studied region;
d) to account for the high number of zeros in order to identify the temporal windows with concentrations higher than zero of each pollen type for the sampled period.

In this study, a hurdle model is implemented to capture the probability of measurements being equal to zero at any given week. This provides information about the presence of each pollen

type across the season. Additionally, a time-varying mean component is used to capture the overall mean concentration of each pollen type across the city.

Chapter 4 has been published in the *Environmental Research Journal*.

Chapter 5 proposes a spatio-temporal model to accommodate for the temporal misalignment across the observed sites. Here, temporal misalignment is described as having measurements taken at different temporal scales across different spatial locations, e.g., some sites provide weekly measurements while others provide daily ones. In this chapter, the temporal misalignment is analyzed using dynamic linear models.

The application of this model consists of analyzing the total pollen concentration in Toronto when, for some of the locations, measurements are taken daily, whereas, for others, measurements are obtained as weekly averages.

The main objectives of this study are:

a) to estimate the daily measurements at sites where only weekly measurements are available;
b) to assess how the temporal scale may impact the estimated associations with the local predictors.

Chapter 5 will be submitted to the *Environmetrics* journal.

The work presented in this thesis aims to expand the methods used in environmental sciences for the study of spatio-temporal processes, particularly regarding the study of aeroallergens and air pollutants dispersion. Here, we propose to use state-of-the-art hierarchical Bayesian models that provide the flexibility to obtain additional information from the data.

The results presented in this thesis will help inform future studies in environmental epidemiology to link environmental exposure to pollen and VOCs with their impact on human health.

References

1. Prüss-Üstün A, Wolf J, Corvalán C, Bos R, Neira M. Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks: World Health Organization; 2016.

2. World Health Organization. Ambient air pollution: a global assessment of exposure and burden of disease. Geneva: World Health Organization; 2016.

3. World Health Organization. Air Pollution. Available from https://www.who.int/health-topics/air-pollution

4. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health. 2020;8.

5. Almetwally AA, Bin-Jumah M, Allam AA. Ambient air pollution and its influence on human health and welfare: an overview. Environmental Science and Pollution Research. 2020;27(20):24815-30.

6. World Health O. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: World Health Organization; 2021 2021.

7. Reimann S, Lewis AC. Anthropogenic VOCs. Volatile Organic Compounds in the Atmosphere2007. p. 33-81.

8. Koppmann R. Volatile organic compounds in the atmosphere: John Wiley & Sons; 2008.

9. Muralikrishna IV, Manickam V. Chapter Eight - Environmental Risk Assessment. In: Muralikrishna IV, Manickam V, editors. Environmental Management: Butterworth-Heinemann; 2017. p. 135-52.

10. Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment. 2008;42(33):7561-78.

11. Reid CE, Gamble JL. Aeroallergens, allergic disease, and climate change: impacts and adaptation. EcoHealth. 2009;6:458-70.

12. Peden D, Reed CE. Environmental and occupational allergies. Journal of Allergy and Clinical Immunology. 2010;125(2, Supplement 2):S150-S60.

13. Rider CF, Yamamoto M, Günther OP, Hirota JA, Singh A, Tebbutt SJ, et al. Controlled diesel exhaust and allergen coexposure modulates microRNA and gene expression in humans: effects on inflammatory lung markers. Journal of Allergy and Clinical Immunology. 2016;138(6):1690-700.

14. Singer BD, Ziska LH, Frenz DA, Gebhard DE, Straka JG. Increasing Amb a 1 content in common ragweed (Ambrosia artemisiifolia) pollen as a function of rising atmospheric CO2 concentration. Functional Plant Biology. 2005;32(7):667-70.

15. Zhang Q, Qiu Z, Chung KF, Huang S-K. Link between environmental air pollution and allergic asthma: East meets West. Journal of thoracic disease. 2015;7(1):14.

16. D'Amato G, Bergmann KC, Cecchi L, Annesi-Maesano I, Sanduzzi A, Liccardi G, et al. Climate change and air pollution. Allergo Journal International. 2014;23(1):17-23.

17. D'Amato G, Cecchi L. Effects of climate change on environmental factors in respiratory allergic diseases. Clinical & Experimental Allergy. 2008;38(8):1264-74.

18. Sierra-Heredia C, North M, Brook J, Daly C, Ellis AK, Henderson D, et al. Aeroallergens in Canada: Distribution, Public Health Impacts, and Opportunities for Prevention. Int J Environ Res Public Health. 2018;15(8).

19. Park HJ, Lee J-H, Park KH, Kim KR, Han MJ, Choe H, et al. A Six-Year Study on the Changes in Airborne Pollen Counts and Skin Positivity Rates in Korea: 2008–2013. Yonsei Med J. 2016;57(3):714-20.

20. Ariano R, Canonica GW, Passalacqua G. Possible role of climate changes in variations in pollen seasons and allergic sensitizations during 27 years. Annals of Allergy, Asthma & Immunology. 2010;104(3):215-22.

21. Ziello C, Sparks TH, Estrella N, Belmonte J, Bergmann KC, Bucher E, et al. Changes to airborne pollen counts across Europe. PloS one. 2012;7(4):e34076.

22. Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, et al. Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science. 1997;11(7):699-718.

23. Katz DSW, Carey TS. Heterogeneity in ragweed pollen exposure is determined by plant composition at small spatial scales. Sci Total Environ. 2014;485-486:435-40.

24. Amann M, Kiesewetter G, Schöpp W, Klimont Z, Winiwarter W, Cofala J, et al. Reducing global air pollution: the scope for further policy interventions. Philosophical Transactions of the Royal Society A. 2020;378(2183):20190331.

2 Literature review

2.1 Environmental processes

The distribution of air pollutants and aeroallergens can be provided as estimates of exposure for epidemiological studies, which in turn can help define demographic groups at risk and identify health and safety concerns [22]. This also implies that a failure in properly modelling spatial distributions can lead to misclassifications in exposure to air pollutants in epidemiological studies [22]. However, the difficulties in modelling some of these environmental processes lie in the complex interactions across multiple spatial and temporal scales.

When monitoring air quality in urban settings, data are collected at a given set of locations, which form the monitoring network, over an area of interest. Frequently, the main goal in these studies is to predict the variable of interest at a set of unobserved sites given the measurements obtained at the monitoring network locations. Furthermore, it is often the case that these measurements are available at different temporal or spatial scales.

Instead of simplifying the problem by making assumptions and manipulating the data before fitting a model, one can decompose the problem into different spatial or temporal levels. Bayesian hierarchical modelling provides a flexible framework that accounts for the uncertainty at the different stages of the modelling process.

2.2 Land-use regression

Land-use regression (LUR) methods are popular in environmental studies, given the data availability and how relatively easy it is to adapt to local circumstances. LUR models combine data collected from a network of monitoring stations over the area of interest with predictor

variables obtained through geographic information systems (GIS) to develop stochastic models that can help understand the spatial distribution of the variable of interest [10, 22]. Additionally, it is also possible to predict the variable of interest at a set of unobserved locations in the same study area.

First called regression mapping, LUR models were used to map the distribution of NO_2 in three different cities [22]. This first study showed the importance of including GIS-derived variables to account for the local factors that would drive the spatial variation of urban air pollution.

Since then, some analyses on particle dispersion using LUR models have shown how the data from a single monitoring site is not representative of the levels of air pollutants or aeroallergens within a city, given the spatial variation associated with their dispersion [25].

In the study of airborne pollutants, these methods have been extensively used to understand the spatial variability of nitrogen oxides and particulate matter [26-28], but also BTEX (benzene, toluene, ethylbenzene and xylenes) [29, 30], ozone [31-33] and other VOCs [34]. Additionally, LUR methods have also been used to estimate the concentration of airborne pollen in various environments and across different cities [35, 36].

However, there are some limitations to LUR methods, for example, it is case and area-specific, and the number of locations can affect the relationships found with local factors [22, 37]. Additionally, one of the main obstacles in LUR models is the lack of guidelines in the monitoring process, e.g., there is no consensus on the appropriate number of sites per unit area, measurement periods, or the number of monitoring campaigns [37]. In this thesis, latent Bayesian structures are included to account for some of these potential issues.

2.3 Bayesian Inference

Bayesian inference focuses on quantifying uncertainty through probabilities. This probabilistic framework allows for a more natural or common-sense interpretation of statistical conclusions. For example, a Bayesian interval (credible interval) of an unknown quantity is interpreted in terms of the probability of the interval containing the "true value" as opposed to the frequentist interval (confidence interval), which represents the number of times the interval contains the "true value" after a sequence of repeated inferences under similar conditions. More importantly, Bayesian inference allows the incorporation of prior knowledge such as expert opinions or previous studies and theory in the inferential procedure [38].

The first step in Bayesian inference is to define a joint probability model for the observed and unobserved quantities that is consistent with previous knowledge [38]. Unobserved quantities can refer to potential observable quantities such as future observations or quantities that might not be directly observable, like parameters [38]. Here, y denotes the observed data or outcome, $\tilde{y}$ the potentially observable data, and $\boldsymbol{\Omega}$ the parameters or unobservable quantities. In many studies, y can be collected as a set of n observations such that $y = (y_1, \dots, y_n)'$, where each of the y_i's, for $i = 1, \dots, n$, is considered as a realization from a random variable whose probability distribution is described by $p(y|\Omega)$. Furthermore, the outcome variables are considered as random to acknowledge the fact that the observed data could have been different depending on the sampling process and the population variations [38].

Then, it is possible to make probability statements on the unobserved quantities given the observed quantities by defining the joint distribution of $\boldsymbol{\Omega}$ and y such that,

$$p(\Omega, y) = p(y|\Omega)p(\Omega),$$

where $p(y|\Omega)$ is the data distribution or likelihood, and $p(\Omega)$ is the prior distribution expressing uncertainty on the parameters. The choice of prior distributions usually depends on the information available. If there is no previous reliable information on the process, it is common practice to use non-informative priors to indicate prior ignorance or vague prior information [38].

Next, to estimate the unobserved quantities of interest, it is necessary to compute their probability distribution conditional on the available information. Bayes' rule allows us to model the conditional distribution of the parameters given the model and the data after assigning a prior distribution to the parameter vector. Conditioning on the observed value of y and following Bayes' theorem, the posterior distribution is defined as,

$$p(\Omega|y) = \frac{p(\Omega, y)}{p(y)} = \frac{p(\Omega)p(y|\Omega)}{p(y)},$$

where the marginal distribution of y is defined by,

$$p(y) = \int p(y|\Omega)p(\Omega)d\Omega,$$

an integral over all possible values of $\boldsymbol{\Omega}$. This marginal distribution is essentially a normalizing constant, so the unnormalized posterior density can be written as,

$$p(\Omega|y) \propto p(\Omega)p(y|\Omega).$$

Every unique combination of data, likelihood and prior distribution leads to a unique posterior distribution [39].

Similarly, after observing y, it is also possible to predict the potentially observable values $\tilde{y}$. Then, the posterior predictive distribution of $\tilde{y}$ is defined as

$$p(\tilde{y}|y) = \int p(\tilde{y}|\Omega)p(\Omega|y)d\Omega,$$

that is, the posterior predictive distribution of $\tilde{y}$ is obtained after integrating out the parameter vector, and this integration is performed with respect to the posterior distribution of Ω. In other words, from a Bayesian point of view, predictions of interest naturally account for the uncertainty in the estimation of the parameters in the model.

2.4 Hierarchical spatio-temporal models

Bayesian hierarchical models provide the flexibility needed to study the complex interactions in environmental processes. Additionally, these hierarchical structures naturally account for the uncertainties in the data, the model parameters, the process, and possible future scenarios [40].

Following convention [41], the hierarchical structure comprises three different stages:

1) Data model: the distribution of the data given the process and parameters;
2) Process model: the distribution of the process given the parameters;
3) Parameters model: the distribution of the parameters.

These three stages can include additional sub-stages that would better describe the process.

Hierarchical models can be implemented through classical or Bayesian methods, but as the model gets more complex, the Bayesian framework allows for more flexibility and easier implementation [42].

Depending on the definition of the model and the parameters, sometimes the resulting posterior distribution does not have a closed form. An alternative is to use computational methods such as Markov Chain Monte Carlo (MCMC) to obtain samples from the resulting posterior distribution [43]. MCMC is a technique in which, instead of computing or directly approximating the

posterior distribution, samples from this distribution are drawn [43]. This is achieved by constructing a Markov chain whose stationary distribution is the posterior distribution. MCMC includes algorithms such as Metropolis-Hastings and Gibbs sampler that, together with the development of more computational power, are responsible for the resurgence of the Bayesian framework in the last few years [38]. These methods can be implemented through packages such as Nimble [44] and Stan [45] that aim to facilitate the implementation of Bayesian methods.

2.4.1 Spatio-temporal models

Let $Y_t(s)$ be a measurement of some spatio-temporal process at time *t* and location *s*, such that $t = 1, \dots, T$, and $s \in D$, where *D* is a continuous spatial domain. Assume that $Y_t(s)$ follows a spatio-temporal model of the form,

$$Y_t(s) = \mu_t(s) + e_t(s),$$

where $\mu_t(s)$ is the mean structure defined as $\mu_t(s) = \boldsymbol{x}_t(s)^T \boldsymbol{\beta}_t(s)$, such that a vector of covariates $\boldsymbol{x}_t(s)$ is available and the coefficients $\boldsymbol{\beta}_t(s)$ can be constant across time or space or both. The residual $e_t(s)$ can be rewritten as the sum of two independent processes: a white noise $\epsilon_t(s)$ and a spatio-temporal process $\omega_t(s)$. Finally, there are many ways in which $\omega_t(s)$ can be modeled that can yield to different relationships between space and time [46].

2.4.2 Dynamic Linear models

Dynamic linear models (DLMs) are probabilistic models that describe a set of observable measurements of a dynamic system as a function of a non-observable state process affected by random dynamics [47, 48]. Their hierarchical structure has made DLMs a popular tool in many

fields such as biology, economics, engineering, neuroscience, and environmental sciences. Although these models can be implemented using a classical frequentist approach, the Bayesian framework is a more natural approach that allows updating the parameter estimates as new information becomes available [47].

The three stages for a dynamic linear model in a hierarchical form are described as [48],

1) Data model: the distribution of the data given the state-space and the parameters;
2) State-space dynamics: the distribution of the state-space given the parameters;
3) Parameters model: Prior distributions.

Multivariate dynamic linear models extend the univariate models and can accommodate temporal observations made across space. More specifically, let $\boldsymbol{Y}_t = (Y_t(s_1), \dots, Y_t(s_n))'$ be a set of observations at discrete time t such that $t = 1, \dots,$ at locations $s_i \in D$. A multivariate DLM is described as a set of two equations, an *observation equation*,

$$\boldsymbol{Y}_t = \boldsymbol{F'}_t\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{V_t}),$$

and a *system* or *state equation*

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim N(0, \boldsymbol{W}_t),$$

with initial state,

$$\boldsymbol{\theta}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0),$$

where the matrix $\boldsymbol{F}_t$ is a $p \times n$ dimensional known matrix of covariates, $\boldsymbol{G}_t$ is a $p \times p$ matrix called evolution matrix, and $\boldsymbol{W}_t$ is a $p \times p$ covariance matrix. Both $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\omega}_t$ are independent, mutually independent, and independent of $\boldsymbol{\theta}_0$. Dynamic linear models are defined by the

quadruple $\{\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{V}_t, \boldsymbol{W}_t\}$. Different combinations can lead to different classes of DLMs, such as a local level, local trend model, or trigonometric seasonal models [49].

Sequential Bayesian inference in DLMs has three crucial steps: filtering, state prediction, and smoothing [47]. A popular algorithm for performing filtering is the Kalman filter to update the latent states.

In Chapters 4 and 5, a combination of spatio-temporal models and a dynamic linear structure is used to analyze the spatio-temporal dispersion of weed, grass, tree, and total pollen in Toronto, Canada.

2.4.3 Hurdle model

Due to the nature of the environmental processes and the sampling methods, there is often an excess of measurements equal to zero, commonly indicating the absence of the process of interest. A common practice in environmental health is to average the observations across the points in time when observations are greater than zero to reduce the dimensionality and complexity of the model. However, the zeros can provide valuable information about the process.

In statistics, zero-inflated and hurdle models are used to handle an excess of zero measurements in the data [50, 51]. Zero-inflated models assume a mix of zero-generating processes by considering an additional probability mass at zero outcomes. Hurdle models assume a mixture of zero and non-zero outcomes described as,

$$p(y|\boldsymbol{\Omega}) = \begin{cases} p(y = 0 \mid \boldsymbol{\Omega}) = \gamma \\ p(y \neq 0 \mid \boldsymbol{\Omega}) = p_{y\neq 0}(y \mid \boldsymbol{\Omega}), \end{cases}$$

where $\boldsymbol{\Omega}$ is a parameter vector, and $p_{y\neq 0}(y\,|\boldsymbol{\Omega})$ is a probability distribution for which zero is not a possible value, e.g., a Poisson distribution truncated at zero or a lognormal distribution.

In Chapter 4, a hurdle model combined with a DLM structure is proposed to analyze the different types of pollen on a weekly scale.

2.5 Modelling of Volatile Organic Compounds

The development of models to study intraurban exposure to air pollutants has been identified as a priority area in public health research [37, 52, 53]. Environmental exposure studies have primarily focused on criteria air pollutants such as particle matter, ozone, carbon monoxide, and nitrogen dioxide, which are regulated in many countries with established air quality standards [37, 53, 54]. However, despite their known adverse health effects, VOCs have been less studied in the health risk literature. This might be due to their low boiling point, chemistry, half-life, interactions with other pollutants, toxicity, and volatility, which makes it challenging to capture small-scale variations in urban areas [53, 54]. Furthermore, the existing studies on VOCs have primarily focused on the aromatic alkylbenzene group, specifically on benzene, toluene, ethylbenzene, and xylene, also known as BTEX [37]. Some studies have highlighted the importance of studying VOC groups other than aromatic alkylbenzenes, given their potential adverse health effects [37, 55].

Samples of air pollutants can be collected using passive samplers, the preferred technique used in situations that do not require a high temporal resolution, allowing for monitoring atmospheric concentrations averaged over extended periods [7, 37]. Moreover, passive samplers have proven to be a successful and reliable alternative to other, more expensive sampling techniques [56].

These samplers are then deployed throughout the study area to capture spatial variation in the concentration of the pollutants, thus, forming a monitoring network. Different monitoring campaigns can be conducted by collecting samples for short periods at different times of the year.

Some types of models for assessing small-scale variations of air pollutants include proximity-based assessments, statistical interpolation (e.g., kriging, inverse distance weighting), dispersion models, land-use regression, integrated emission-meteorological models, and hybrid models (a combination of personal or regional monitoring with other air pollution exposure method) [52]. However, due to their performance and relatively low cost, LUR models have been the preferred method for exposure assessment in urban areas [10, 37].

When data is available from multiple monitoring campaigns for the whole monitoring network, some studies compute the average across campaigns and perform the exposure analysis [27, 57-59]. However, information on the seasonal variability and spatial distribution across campaigns might be lost in the analysis. Incorporating knowledge of the factors related to spatial variation can potentially make models more readily transferable to other areas [10]. Furthermore, including spatial and temporal components in these models might be of interest to provide more detailed information for exposure assessment studies [10].

Even though to obtain these long-term exposure estimates (e.g., annual) from a monitoring network, the ideal would be to sample at all sites throughout the study period, depending on the number of sites, this might increase the cost of epidemiological studies [60]. Alternatively, one can use reference sites which are a small set of sites with available measurements throughout the whole study period [37, 60]. Using data from the monitoring network and the reference sites, the long-term mean can be adjusted, for example, with multiplicative or additive imputation methods

[60]. Nevertheless, extreme weather conditions, operating costs, and equipment limitations might hinder the possibility of having a set of reference sites. Moreover, there is no consensus in the literature concerning the number of monitoring locations, reference sites, sampling periods, number of campaigns, or the measurement frequency [37].

Most LUR models for VOCs focus on BTEX, whose main sources are vehicle exhaust and industrial activities. Given their nature, the most important variables identified in LUR models for these pollutants included traffic surrogates, population density, geographical variables, and industrial and commercial land use [30, 37, 58, 59, 61].

There is no consensus on the best model evaluation method for LUR models, but some common methods include cross-validation and bootstrapping [37]. Furthermore, the standard in the literature is to compute R^2 for model comparison and model selection. However, R^2 is a measure that compares the fitted model with the simplest model, which considers a constant mean. Moreover, as you increase model complexity, R^2 increases, and you do not have a component that penalizes overly complex models.

2.6 Modelling of Pollen

In Canada, there are five floristic regions where different aeroallergens-producing plants can grow [62]. The aeroallergens more commonly associated with allergies in Canada include trees, grasses, and ragweed pollen [63-65]. In Northern climates, tree pollen begin to pollinate in early spring to late summer, grass pollination occurs in mid-spring to mid-summer, and weeds release pollen in mid-summer and early fall [66, 67]. However, plant groups do not strictly follow the timing of pollination [67].

In large cities, pollen monitoring is usually conducted at one location positioned at rooftop-level [68]. However, this technique fails to capture the intra-urban spatial variability and does not reflect exposures at breathing levels needed for health-related studies [25, 35, 69, 70]. Although some studies have found similarities in the pollen spectrum and seasonal course for the whole city, there are inconsistencies in the spatial distribution across different monitoring stations [68]. Therefore, some studies have highlighted the need to use multiple monitors placed at breathing level heights instead of-roof level, especially for exposure assessment studies [35, 71, 72].

At the urban scale, spatial variability may depend on taxa and pollen concentration [25]. For example, days with lower pollen concentration might tend to have less spatial variability [25]. Additionally, spatial variation in plant composition and abundance, phenology, pollen traits, management of urban vegetation, and meteorological conditions might influence the spatiotemporal patterns of pollen dispersion, release, and concentration across a city [25, 68]. Furthermore, physical structures like buildings and trees can affect wind patterns and, therefore, pollen dispersal [25].

To capture different aeroallergens outdoors at multiple locations, airborne particles can be collected passively, using gravity, or actively using methods such as impaction, impingement, and other methods that can provide volumetric samples [73]. The choice of samplers can affect the time scale and measurement unit [71].

The most common sampling method in allergies-related studies is impaction samplers, specifically rotorod samplers [73]. These samplers use a motor that rotates a sampling head with adhesive-coated rods in order to capture airborne particles [73]. The collected samples are later analyzed using microscopic examination and identification based on their morphological aspects [74]. This process requires trained and experienced personnel to identify the spores and pollen

types commonly found in the samples [74]. Therefore, the process takes considerable time, and real-time data is not available [18, 74]. Furthermore, low sampling periods and dense spatial coverage are challenging to obtain [18].

To capture the spatial or spatiotemporal variation of pollen within a city, the sampling process can occur sequentially at different locations, where multiple sites are sampled at different times [35, 75]. Some studies have also integrated the data observed over long sampling periods [23, 36, 68]. However, this does not show the short-term variability needed for health-related studies [25].

Similar to monitoring air pollutants, samplers can also be deployed over the area of interest, forming a monitoring network. Then, using the data collected at these monitoring sites, it is possible to model and forecast aeroallergens concentration to estimate environmental exposure at locations not included in the sampling sites. Some methods such as habitat distribution models, land cover, circular statistics, and land-use regression can be used to predict airborne pollen concentrations [35, 36, 76, 77].

Multiple studies analyzing the intra-urban variation of pollen concentration of different pollen types have found a high spatial variability across the city [23, 35, 36, 72, 77, 78]. Furthermore, some studies found higher concentrations of pollen [72, 78] and higher variability across monitoring sites [72] near ground level. Some factors that have been found to contribute to airborne pollen include local plant populations [23, 79], the presence of parks and gardens [79], and urbanity level [80, 81].

Given the additional difficulties in pollen sampling, LUR methods are not as widely used in pollen studies as they are for air pollutants. Nevertheless, some studies have successfully used

LUR methods to provide insight into the relationship between environmental factors and pollen concentration in human-modified environments at a fine spatial scale [35, 36].

Combining maps of urban plants with pollen production estimates, phenology, and atmospheric models might improve the prediction and forecasting of pollen concentration within a city [25]. Additionally, similar to air pollutants, incorporating temporal components to these models might also be of interest for exposure assessment studies.

References

7. Reimann S, Lewis AC. Anthropogenic VOCs. Volatile Organic Compounds in the Atmosphere2007. p. 33-81.

10. Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment. 2008;42(33):7561-78.

18. Sierra-Heredia C, North M, Brook J, Daly C, Ellis AK, Henderson D, et al. Aeroallergens in Canada: Distribution, Public Health Impacts, and Opportunities for Prevention. Int J Environ Res Public Health. 2018;15(8).

22. Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, et al. Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science. 1997;11(7):699-718.

23. Katz DSW, Carey TS. Heterogeneity in ragweed pollen exposure is determined by plant composition at small spatial scales. Sci Total Environ. 2014;485-486:435-40.

25. Katz DSW, Batterman SA. Urban-scale variation in pollen concentrations: a single station is insufficient to characterize daily exposure. Aerobiologia. 2020.

26. Oiamo TH, Johnson M, Tang K, Luginaah IN. Assessing traffic and industrial contributions to ambient nitrogen dioxide and volatile organic compounds in a low pollution urban environment. Science of The Total Environment. 2015;529:149-57.

27. Crouse DL, Goldberg MS, Ross NA. A prediction-based approach to modelling temporal and spatial variability of traffic-related air pollution in Montreal, Canada. Atmospheric Environment. 2009;43(32):5075-84.

28. Henderson SB, Beckerman B, Jerrett M, Brauer M. Application of Land Use Regression to Estimate Long-Term Concentrations of Traffic-Related Nitrogen Oxides and Fine Particulate Matter. Environmental Science & Technology. 2007;41(7):2422-8.

29. Mohammadi A, Ghassoun Y, Löwner M-O, Behmanesh M, Faraji M, Nemati S, et al. Spatial analysis and risk assessment of urban BTEX compounds in Urmia, Iran. Chemosphere. 2020;246:125769.

30. Atari DO, Luginaah IN. Assessing the distribution of volatile organic compounds using land use regression in Sarnia," Chemical Valley", Ontario, Canada. Environmental Health. 2009;8(1):1-14.

31. Wang J, Cohan DS, Xu H. Spatiotemporal ozone pollution LUR models: Suitable statistical algorithms and time scales for a megacity scale. Atmospheric Environment. 2020;237:117671.

32. Alvarez-Mendoza CI, Teodoro A, Ramirez-Cando L. Spatial estimation of surface ozone concentrations in Quito Ecuador with remote sensing data, air pollution measurements and meteorological variables. Environmental Monitoring and Assessment. 2019;191(3):155.

33. Masiol M, Squizzato S, Chalupa D, Rich DQ, Hopke PK. Spatial-temporal variations of summertime ozone concentrations across a metropolitan area using a network of low-cost monitors to develop 24 hourly land-use regression models. Science of The Total Environment. 2019;654:1167-78.

34. Mukerjee S, Smith LA, Johnson MM, Neas LM, Stallings CA. Spatial analysis and land use regression of VOCs and NO2 from school-based urban air monitoring in Detroit/Dearborn, USA. Science of The Total Environment. 2009;407(16):4642-51.

35. Hjort J, Hugg TT, Antikainen H, Rusanen J, Sofiev M, Kukkonen J, et al. Fine-Scale Exposure to Allergenic Pollen in the Urban Environment: Evaluation of Land Use Regression Approach. Environ Health Perspect. 2016;124(5):619-26.

36. Weinberger KR, Kinney PL, Robinson GS, Sheehan D, Kheirbek I, Matte TD, et al. Levels and determinants of tree pollen in New York City. Journal of Exposure Science & Environmental Epidemiology. 2018;28(2):119-24.

37. Amini H, Yunesian M, Hosseini V, Schindler C, Henderson SB, Künzli N. A systematic review of land use regression models for volatile organic compounds. Atmospheric Environment. 2017;171:1-16.

38. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Third ed. Boca Raton: CRC Press; 2014.

39. McElreath R, O'Reilly for Higher E, Safari aORMC. Statistical Rethinking, 2nd Edition: Chapman and Hall/CRC; 2020.

40. Clark JS, Gelfand AE. A future for models and data in environmental science. Trends in Ecology & evolution. 2006;21(7):375-80.

41. Berliner LM. Hierarchical Bayesian time series models. Maximum entropy and Bayesian methods: Springer; 1996. p. 15-22.

42. Wikle CK. Hierarchical models in environmental science. International Statistical Review. 2003;71(2):181-99.

43. Gamerman D, Lopes HF. Markov chain Monte Carlo : stochastic simulation for Bayesian inference. 2nd ed. / ed. Boca Raton: Taylor & Francis; 2006.

44. de Valpine PaP, C. and Turek, D. and Michaud, N. and Anderson-Bergman, C. and Obermeyer, F. and Wehrhahn Cortes, C. and Rodriguez, A. and Temple Lang, D. and Paganin,

S. NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. 0.11.1 ed2021.

45. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. 2.29 ed2022.

46. Banerjee S, Gelfand AE, Carlin BP. Hierarchical modeling and analysis for spatial data. Boca Raton, Florida Chapman & Hall/CRC Press; 2015.

47. Petris G, Petrone S, Campagnoli P. Dynamic linear models. Dynamic Linear Models with R: Springer; 2009. p. 31-84.

48. Schmidt AM, Lopes HF. Dynamic models. Handbook of Environmental and Ecological Statistics: CRC Press; 2019. p. 57-80.

49. West M, Harrison J. Bayesian forecasting and dynamic models. New York: Springer; 1997.

50. Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986;33(3):22.

51. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34(1):1-14.

52. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol. 2005;15(2):185-204.

53. Su JG, Jerrett M, Beckerman B, Verma D, Arain MA, Kanaroglou P, et al. A land use regression model for predicting ambient volatile organic compound concentrations in Toronto, Canada. Atmospheric Environment. 2010;44(29):3529-37.

54. Woodruff TJ, Axelrad DA, Caldwell J, Morello-Frosch R, Rosenbaum A. Public health implications of 1990 air toxics concentrations across the United States. Environmental Health Perspectives. 1998;106(5):245-51.

55. Wheeler AJ, Smith-Doiron M, Xu X, Gilbert NL, Brook JR. Intra-urban variability of air pollution in Windsor, Ontario--measurement and modeling for human exposure assessment. Environ Res. 2008;106(1):7-16.

56. Marć M, Bielawska M, Wardencki W, Namieśnik J, Zabiegała B. The influence of meteorological conditions and anthropogenic activities on the seasonal fluctuations of BTEX in the urban air of the Hanseatic city of Gdansk, Poland. Environmental Science and Pollution Research. 2015;22(15):11940-54.

57. Deville Cavellin L, Weichenthal S, Tack R, Ragettli MS, Smargiassi A, Hatzopoulou M. Investigating the Use Of Portable Air Pollution Sensors to Capture the Spatial Variability Of Traffic-Related Air Pollution. Environmental Science & Technology. 2016;50(1):313-20.

58. Gaeta A, Cattani G, Di Menno di Bucchianico A, De Santis A, Cesaroni G, Badaloni C, et al. Development of nitrogen dioxide and volatile organic compounds land use regression models to estimate air pollution exposure near an Italian airport. Atmospheric Environment. 2016;131:254-62.

59. Aguilera I, Sunyer J, Fernández-Patier R, Hoek G, Aguirre-Alfaro A, Meliefste K, et al. Estimation of Outdoor NOx, NO2, and BTEX Exposure in a Cohort of Pregnant Women Using Land Use Regression Modeling. Environmental Science & Technology. 2008;42(3):815-21.

60. Amini H, Hosseini V, Schindler C, Hassankhany H, Yunesian M, Henderson SB, et al. Spatiotemporal description of BTEX volatile organic compounds in a Middle Eastern megacity:

Tehran Study of Exposure Prediction for Environmental Health Research (Tehran SEPEHR). Environmental Pollution. 2017;226:219-29.

61. Liu K, Zhang C, Cheng Y, Liu C, Zhang H, Zhang G, et al. Serious BTEX pollution in rural area of the North China Plain during winter season. Journal of Environmental Sciences. 2015;30:186-90.

62. Weber RW. Floristic zones and aeroallergen diversity. Immunology and Allergy Clinics of North America. 2003;23(3):357-69.

63. Ellis AK, North ML, Walker T, Steacy LM. Environmental exposure unit: a sensitive, specific, and reproducible methodology for allergen challenge. Annals of Allergy, Asthma & Immunology. 2013;111(5):323-8.

64. Ellis AK, Soliman M, Steacy L, Boulay M-È, Boulet L-P, Keith PK, et al. The Allergic Rhinitis – Clinical Investigator Collaborative (AR-CIC): nasal allergen challenge protocol optimization for studying AR pathophysiology and evaluating novel therapies. Allergy, Asthma & Clinical Immunology. 2015;11(1):16.

65. White JF, Bernstein DI. Key pollen allergens in North America. Annals of Allergy, Asthma & Immunology. 2003;91(5):425-35.

66. Portnoy J, Barnes C. Clinical relevance of spore and pollen counts. Immunology and Allergy Clinics of North America. 2003;23(3):389-410.

67. Martorano L, Erwin EA. Aeroallergen Exposure and Spread in the Modern Era. The Journal of Allergy and Clinical Immunology: In Practice. 2018;6(6):1835-42.

68. Werchan B, Werchan M, Mucke HG, Gauger U, Simoleit A, Zuberbier T, et al. Spatial distribution of allergenic pollen through a large metropolitan area. Environ Monit Assess. 2017;189(4):169.

69. Werchan B, Werchan M, Mücke H-G, Bergmann K-C. Spatial distribution of pollen-induced symptoms within a large metropolitan area—Berlin, Germany. Aerobiologia. 2018;34(4):539-56.

70. Peel RG, Hertel O, Smith M, Kennedy R. Personal exposure to grass pollen: relating inhaled dose to background concentration. Annals of Allergy, Asthma & Immunology. 2013;111(6):548-54.

71. Weinberger KR, Kinney PL, Lovasi GS. A review of spatial variation of allergenic tree pollen within cities. Arboriculture & Urban Forestry. 2015;41(2):57-68.

72. Rojo J, Oteros J, Pérez-Badia R, Cervigón P, Ferencova Z, Gutiérrez-Bustillo AM, et al. Near-ground effect of height on pollen exposure. Environmental Research. 2019;174:160-9.

73. Levetin E. Methods for aeroallergen sampling. Current allergy and asthma reports. 2004;4(5):376-83.

74. Weber RW. Outdoor aeroallergen sampling: not all that simple. Annals of Allergy, Asthma & Immunology. 2007;98(6):505-6.

75. Katz DSW, Dzul A, Kendel A, Batterman SA. Effect of intra-urban temperature variation on tree flowering phenology, airborne pollen, and measurement error in epidemiological studies of allergenic pollen. Sci Total Environ. 2019;653:1213-22.

76. Maya-Manzano JM, Sadyś M, Tormo-Molina R, Fernández-Rodríguez S, Oteros J, Silva-Palacios I, et al. Relationships between airborne pollen grains, wind direction and land cover using GIS and circular statistics. Science of The Total Environment. 2017;584-585:603-13.

77. Katz DSW, Batterman SA. Allergenic pollen production across a large city for common ragweed (Ambrosia artemisiifolia). Landsc Urban Plan. 2019;190.

78. Charalampopoulos A, Damialis A, Lazarina M, Halley JM, Vokou D. Spatiotemporal assessment of airborne pollen in the urban environment: The pollenscape of Thessaloniki as a case study. Atmospheric Environment. 2021;247:118185.

79. Rojo J, Rapp A, Lara B, Fernandez-Gonzalez F, Perez-Badia R. Effect of land uses and wind direction on the contribution of local sources to airborne pollen. Sci Total Environ. 2015;538:672-82.

80. Hugg TT, Hjort J, Antikainen H, Rusanen J, Tuokila M, Korkonen S, et al. Urbanity as a determinant of exposure to grass pollen in Helsinki Metropolitan area, Finland. PLOS ONE. 2017;12(10):e0186348.

81. Ríos B, Torres-Jardón R, Ramírez-Arriaga E, Martínez-Bernal A, Rosas I. Diurnal variations of airborne pollen concentration and the effect of ambient temperature in three sites of Mexico City. International Journal of Biometeorology. 2016;60(5):771-87.

3 Spatial modelling of ambient concentrations of volatile organic compounds in Montreal, Canada

3.1 Preface to Manuscript 1

In this chapter, a spatial Bayesian hierarchical model is proposed to study the dispersion of VOCs in Montreal. Land-use regression is extended by considering a potential spatial structure left after accounting for the land-use variables. The data consists of VOC concentration measurements made across three different monitoring campaigns: December 2005, April 2006, and August 2006.

The differences across campaigns are captured by the spatial structure, an indicator variable, and by allowing the coefficients associated with the land-use variables to change across campaigns. Taking advantage of the hierarchical structure, each campaign is modelled separately allowing decision-makers to learn how the concentration of different VOCs varies across campaigns. Additionally, the mean concentration across campaigns is obtained at a set of unobserved locations of interest.

The predicted surfaces obtained in this study will be used in future health studies to investigate if there are any associations between the concentration of a certain VOC over an area and the relative risk of breast or prostate cancer.

At the time of writing this thesis, this manuscript is under review in the *Environmental Epidemiology* journal.

3.2 Authors contribution

Sara Zapata-Marin: Formal analysis, Methodology, Writing - Original Draft

Alexandra M. Schmidt: Methodology, Supervision, Writing - Review & Editing

Dan Crouse: Data Collection, Writing - Review & Editing

Vikki Ho: Writing - Review & Editing

France Labrèche: Writing - Review & Editing

Eric Lavigne: Funding acquisition, Writing - Review & Editing

Marie-Élise Parent: Writing - Review & Editing

Mark S. Goldberg: Conceptualization, Funding acquisition, Writing - Original Draft

Spatial modelling of ambient concentrations of volatile organic compounds in Montreal, Canada

Sara Zapata-Marin [a*], Alexandra M. Schmidt [a,b], Dan Crouse [c], Vikki Ho [d,e], France Labrèche [f], Eric Lavigne [g,h], Marie-Élise Parent [e,i], Mark S. Goldberg [b, j-l]

[a] Quantitative Life Sciences, McGill University, Montreal, QC, Canada

[b] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada

[c] Health Effects Institute, Boston, MA, United States

[d] Health Innovation and Evaluation Hub, Université de Montréal, Hospital Research Centre (CRCHUM), Montreal, QC, Canada

[e] Department of Social and Preventive Medicine, École de santé publique de l'Université de Montréal (ESPUM), Montreal, QC, Canada

[f] Department of Environmental and Occupational Health, École de santé publique de l'Université de Montréal (ESPUM), Montréal, QC, Canada

[g] Air Health Science Division and Population Studies Division, Health Canada, Ottawa, ON, Canada

[h] School of Epidemiology & Public Health, University of Ottawa, Ottawa, ON, Canada

[i] Epidemiology and Biostatistics Unit, Centre Armand-Frappier Santé Biotechnologie, Institut national de la recherche scientifique, Université du Québec, Laval, QC, Canada

[j] Department of Medicine, McGill University Health Center, Montreal, QC, Canada

[k] Gerald Bronfman Department of Oncology, McGill University, Montreal, QC, Canada

[l] Centre for Outcomes Research and Evaluation, Research Institute of the McGill University Hospital Centre, Montreal, QC, Canada

3.3 Abstract

Background: Volatile organic compounds are components of the complex mixture of air pollutants within cities and can cause various adverse health effects. Therefore, it is necessary to understand their spatial distribution for exposure assessment in epidemiological studies.

Objectives: The objective was to model measured concentrations of five VOCs within the city of Montreal, Canada, developing spatial prediction models that can be used in health studies.

Methods: We measured concentrations using 3M 3500 Organic Vapour Monitors, over two-week periods, for three monitoring campaigns between 2005 and 2006 in over 130 locations in the city. Using GC/MSD, we measured concentrations of benzene, n-decane, ethylbenzene, hexane, and trimethylbenzene. We fitted four different models that combine land-use regression and geostatistical methods to account for the potential spatial structure that remains after accounting for the land-use variables. The fitted models also accounted for possible variations in the concentration of air pollutants across campaigns.

Results: The highest concentrations for all VOCs were found in December with hexane being the most abundant followed by ethylbenzene. We obtained predicted surfaces for the VOCs for the three campaigns as well as mean surfaces across campaigns. We found higher concentrations of some VOCs along highways, and in the Eastern part of Montreal which is a highly industrialized area.

Conclusions: Each of the fitted models captured the spatial and across-campaigns variability for each VOC, and we found that different VOCs required different model structures.

Keywords: air pollution; monitoring; spatial statistics; volatile organic compounds

What this study adds

It is critical to study the spatial distribution of Volatile Organic Compounds (VOCs) to understand the population health risks associated with their exposure. Bayesian hierarchical models were used to account for the spatial and across-campaigns variation of benzene, decane, ethylbenzene, trimethylbenzene, and hexane concentration in Montreal for three monitoring campaigns. Higher concentrations were found during December with hexane being the most abundant VOC. The predicted surfaces also showed higher concentrations along some of the major highways and in the Eastern part of the island where refineries were operating at the time of the study.

3.4 Introduction

Volatile organic compounds (VOCs) are organic compounds that have high vapor pressures (~10 Pa) at room temperature (25°C) [82]. Acute and chronic exposures to these chemical compounds can cause adverse health effects such as irritation of the eyes and upper respiratory tract, and effects on the central nervous system (e.g., loss of coordination), as well as being toxic to the liver and kidneys [83, 84]. Furthermore, benzene, trichloroethylene, and vinyl chloride are accepted carcinogens [84] and ethylbenzene, carbon tetrachloride, 1,2-dichloroethane, chloroform, and other trihalomethanes have been identified as possible carcinogens.

In Canada, it has been estimated that 2.3 Mt of VOCs were released in 2005, a 21% (600 kt) decrease from 1990 [85]. Oil and gas industries were the main source of VOC emissions (29.6% of total emissions; 680 kt), followed by transportation and mobile equipment (27.4%; 630 kt), and paints and solvents (19.13%; 440 kt) [85].

3.4.1 Spatial distribution and seasonality of concentrations of VOCs in urban areas

To identify sources of atmospheric pollutants, governments use annual emission inventories of greenhouse gases and air pollutants. Given restrictions related to data, time, staff, funding, and the lack of a systematic assessment, it has been shown that emission inventories often underestimate concentrations of air pollutants [86-88]. Although most emission inventories have a 1^{o} by 1^{o} spatial resolution and a temporal resolution of one year, this level of spatial resolution is not sufficient to characterize the variability of concentrations of some pollutants within cities, as some studies have shown that, for pollutants related to traffic, intra-urban variability exceeds

inter-city variability [22, 52, 89]. Furthermore, recent studies found differences in the distribution of sources as well as precursors of ozone, especially VOCs, between urban, suburban, rural, and industrial areas [27, 90-92].

Land-use regression (LUR) methods have been extensively used to estimate the spatial variability of air pollutants and their relationship with environmental factors in urban settings. Although there is no consensus on the monitoring process [37] (e.g., number of monitoring sites, monitoring period, the distance between sites) a study on NO_2 found that, studies using LUR should be based on a large number of sites (>80) for better performance [93].

3.4.2 Objectives

To support two population-based case-control studies of postmenopausal breast cancer [94, 95] and one prostate cancer case-control study [96] that we conducted in the mid-1990s and early 2000s in Montreal, Quebec, we conducted a dense monitoring program of NO_2 [27] and selected VOCs in order to link these to the residential addresses of participants in these studies.

The main objective of the present study was to determine the spatial distribution of ambient concentrations of selected VOCs from our monitoring campaign conducted in 2005 and 2006. Using a combination of land-use regression and geostatistical methodswe analyzed benzene, ethylbenzene, and trimethylbenzene which are aromatic hydrocarbons found in fossil fuels and urban air masses, predominantly emitted by vehicle exhausts, fuel evaporation, and spoilage [7].

We also analyzed n-decane and hexane which are alkane hydrocarbons found in fossil fuels and solvents that become airborne by evaporation or combustion.

Additionally, we obtained predicted surfaces by interpolating concentrations at locations where measurements were not made while accounting for local variations in concentrations so that, in future studies, we can link the concentration predictions with residential addresses of participants in the three cancer case-control studies [94-96], and hence estimate risks associated with these exposures.

3.5 Materials and methods

The greater Montreal area is the second most populated city in Canada, with a population in 2016 of over four million inhabitants [97]. From 1981 to 2010, the mean daily temperature in April was around 6.4ºC (temperature range 1.2 to 11.6ºC), 20.1ºC in August (temperature range 14.8 to 25.3ºC), and -5.4ºC in December (temperature range -9.3 to -1.4ºC) [98]. The average annual concentration of NO_2 in 2018 was of 10.4 parts per billion (ppb), while the three-year average from 2016 to 2018 was 7.4 $\mu g/m^3$ for fine particulate matter and 57 ppb for ozone [99].

The east end of Montreal is of particular interest as it is an industrial area with refineries and various other heavy industries. Between 2005 and 2006, refineries included Shell Canada Montreal East Refinery (closed in 2010), the Petro Canada Montreal Refinery (now Suncor), and petrochemical plants like Parachem Petrochemical and Petromont (closed in 2008) [100].

3.5.1 Data Collection

The location of the samplers was chosen using a population-weighted location-allocation model that placed 133 samplers in areas likely to have high spatial variability of traffic-related pollution and in areas with high population densities [27]. In addition, we added about 20 samplers to capture concentrations in residential areas that were under-represented by the initial allocation scheme. The minimum distance between any two neighbouring samplers was approximately 100 m and the maximum distance was just over three kilometres. The samplers were deployed in three monitoring campaigns: December 2005 ("cold" weather), April 2006 ("temperate" weather), and August 2006 ("hot" weather).

In addition to the Ogawa samplers that measured concentrations of NO_2 , [27] we co-located passive 3M 3500 Organic Vapour Monitors (3M Company, Saint Paul, MN, USA). After a two-week uninterrupted sampling period, we retrieved each monitor, snapped the shipping cap onto the monitor, ensuring that the two-port plugs were sealed firmly, and then recorded the date and time. We placed the sampler in the shipping container, closed the container with its plastic lid, and then sealed it immediately with Teflon tape. These were then shipped to a commercial laboratory that conducted all of the analyses (Airzone, Mississauga, ON [101]).

Samples were extracted with 2 ml of solvent (carbon disulfide) and concentrations using GC/MSD were estimated, using NIOSH methods 1003, 1500 and 1501 with a detection limit of 0.2 $\mu g/m^3$. We had three field blanks per sampling survey, and all sample results were corrected

with the blanks, deuterated internal standard and recovery. The multipoint calibration curve had a $R^2 > 0.999$ and the detection limits, based on the U.S. Federal Register Code of Federal Regulations (CFR) 40 method, are shown in (see Supplementary Material Table 1).

VOLATILE ORGANIC COMPOUNDS DATA

We analysed concentrations of five exhaust-related VOCs (n-decane, hexane, ethylbenzene, benzene, and 1,2,4-trimethylbenzene) for three, continuous 2-week monitoring campaigns in December 2005, April 2006, and August 2006. There were 133 monitoring locations for the December and April campaigns, and 131 for the August campaign.

LAND-USE VARIABLES

Potential predictors of the different VOCs were obtained using circular buffers at 50-, 100-, 200-, 500-, 1000-m radii around each monitoring location (see Table 2 in the Supplementary Material). Land-use variables were available as a proportion of the area of each buffer covered by each specific variable. The available land-use variables for each VOC were obtained from DMTI Spatial Inc.[102] and included buildings, open areas, residential, industrial, commercial, waterbody, parks and recreational, governmental and institutional, commercial, and roads land-use, which are common predictors for local variability of urban air pollution [27, 103, 104]. Average and total NOx and total daily traffic volume were obtained from VISSIM [105], a traffic simulation software, and MOVES [106], an emission modeling system for mobile sources. Population density for 2016 was based on Canadian census data [97]. Finally, the easting and northing coordinates were also included in the mean structure of the models.

After obtaining the spatial variables at the different buffer sizes, we selected the appropriate variables and buffer sizes for each VOC in each campaign by using the procedure of least absolute shrinkage and selection operator (LASSO) [107]. This regression analysis method shrinks the coefficients towards zero to reduce the set of covariates used in the model. The goal is to minimize the sum of squared errors with a bound on the absolute values of the coefficients. We included the variables selected using LASSO for each campaign so that we would have the same set of variables for all three campaigns.

Because our main goal was to predict concentrations where measurements were not made rather than to find associations between the concentrations and the land-use variables, we excluded variables that were highly correlated (>0.99) with each other (see tables 3-9 in the Supplementary Material).

3.5.2 Statistical Analysis

For each VOC, we fitted four different regression models (see below). These models considered possible variations between monitoring campaigns and a possible spatial structure between the monitoring locations after accounting for the land-use variables. To select the best model among the fitted ones, for each VOC the Watanabe-Akaike information criterion (WAIC) was used. WAIC is a measure of the predictive accuracy that allows us to measure the performance of a model and to compare multiple models accounting for both goodness of fit and model complexity [108]. Smaller values indicate the optimal model among the fitted ones.

3.5.3 Model description

To obtain predicted surfaces over a region given the information collected at a set of monitoring stations, we used a combination of land-use regression and geostatistical methods considering possible variations across campaigns and a possible latent (unobservable) spatial structure after accounting for the land-use variables [46].

Specifically, for all our models, let $Y_j(\boldsymbol{s})$ be the natural logarithmic concentration of a VOC at location $\boldsymbol{s}$, and campaign j where j = 1, 2, 3 for the December, April, and August campaigns, respectively. The base model is defined as follows,

$$Y_j(\boldsymbol{s}) = Z_j'\alpha_j + \boldsymbol{X}'(\boldsymbol{s})\boldsymbol{\beta_j} + \omega_j(\boldsymbol{s}) + \epsilon_j(\boldsymbol{s}), \quad (1)$$

where $\boldsymbol{X}(\boldsymbol{s})$ is a q-dimensional vector containing the land-use predictors, an intercept, and the standardized Universal Transverse Mercator (UTM) coordinates at each monitoring site; $\boldsymbol{\beta}_j$ is a vector of coefficients associated with the land-use variables for each campaign; Z_j represents each campaign, such that $Z_1 = 0$, and α_j is the coefficient associated with this indicator variable. Finally, the latent spatial structure $\boldsymbol{\omega}_j$ helps to accommodate for a possible residual spatial structure after accounting for the land-use variables. This spatial residual $\boldsymbol{\omega}_j = (\omega_j(\boldsymbol{s}_1), \cdots, \omega_j(\boldsymbol{s}_n))'$ follows, *a priori*, a Gaussian process with mean 0 and an exponential covariance matrix $\boldsymbol{\Sigma}_j = \sigma_j^2 exp(-\boldsymbol{d}/\phi_j)$ where σ_j^2 is the partial sill at each campaign j, $\boldsymbol{d}$ is an Euclidean distance matrix, and ϕ_j is a parameter that controls how fast the spatial correlation

among $\omega_j(\cdot)$ decays to zero; and the measurement error $\epsilon_j(\boldsymbol{s})$ follows a zero mean normal distribution with variance τ^2 (nugget effect).

We fitted four models that are particular cases of the general structure in equation (1). For Model 1, we let $\boldsymbol{\beta}_j = \boldsymbol{\beta}$ meaning the effect of the land-use variables is the same across campaigns, in Model 2 we also let $\boldsymbol{\beta}_j = \boldsymbol{\beta}$, and $\omega_j(\boldsymbol{s}) = 0$, so that the spatial variation is fully captured by the land-use variables. In Model 3, we let $\boldsymbol{\alpha} = \mathbf{0}$, so that all the variability across campaigns is captured by $\boldsymbol{\beta}_j$, and $\omega_j(\boldsymbol{s})$, and finally in Model 4 we let $\boldsymbol{\alpha} = \mathbf{0}$ and $\omega_j(\boldsymbol{s}) = 0$, meaning there is no additional spatial structure after accounting for the land-use variables.

Under the Bayesian paradigm, model specification is complete after assigning a prior distribution to the parameter vector. In our case, all parameters were assumed to be independent *a priori.* We assigned a normal distribution with a mean of zero and with large variances for $\boldsymbol{\alpha}$ and for $\boldsymbol{\beta}$ in models 1 and 2, as this reflects our prior ignorance about the association between the land-use variables and the VOCs concentrations; for $\boldsymbol{\sigma}^2$ and τ^2 we assigned an inverse gamma prior with mean fixed at 1 and an infinite variance; for the spatial range $\boldsymbol{\phi}$ we assigned an exponential prior with mean equal to the practical range (assuming the correlation between sites decreases to 0.05 at half of the maximum observed distance); and for model 4, the regression coefficients $\boldsymbol{\beta}_j$ follow a normal prior with mean $\boldsymbol{\gamma}$, and variance ψ^2 where $\boldsymbol{\gamma} \sim N(0,10)$, and ψ^2 follows an inverse gamma distribution with mean equals to 1 and an infinite variance. The parameter vector for each model is defined as $\theta_1 = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau, \boldsymbol{\sigma}, \boldsymbol{\phi}\}$ for model 1, $\theta_2 = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau\}$ for model 2, $\theta_3 = \{\boldsymbol{\beta}, \tau, \boldsymbol{\sigma}, \boldsymbol{\phi}\}$ for model 3, and $\theta_4 = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau, \psi^2\}$ for model 4.

We used Markov Chain Monte Carlo methods[43] to obtain samples from the posterior distribution. The statistical analysis was conducted using the package Nimble in the software R (version 3.4.5, [44]). The code for fitting these models is available at https://github.com/SaraZM/VOCs.

After fitting all four models we chose the appropriate model for each VOC using the minimum value of the WAIC.

3.5.4 Spatial Interpolation

To predict concentrations of VOCs at unsampled locations across Montreal, we needed the posterior predictive distributions of the different VOCs at unobserved locations of interest. To accomplish this, let $\boldsymbol{Y}_j^u$ be a vector that contains the unknown concentrations at a set of unobserved locations $\mathcal{S}^u = (\boldsymbol{s}_1^u, \dots, \boldsymbol{s}_n^u)$ with associated covariates $\boldsymbol{X}^u = (\boldsymbol{X}(\boldsymbol{s}_1^u), \dots, \boldsymbol{X}(\boldsymbol{s}_n^u))$. Under the Bayesian framework this is achieved through the posterior predictive distribution,

$$p(\boldsymbol{Y}_j^u \,|\boldsymbol{Y}_j, \boldsymbol{X}, \boldsymbol{X}^u, \boldsymbol{\theta}) = \int p(\boldsymbol{Y}_j^u \,|\boldsymbol{Y}_j, \boldsymbol{X}^u, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}_j, \boldsymbol{X})\, d\boldsymbol{\theta} \tag{2}$$

$$= E_{p(\boldsymbol{\theta}|\boldsymbol{Y}_j)}\left[p(\boldsymbol{Y}_j^u \,|\boldsymbol{Y}_j, \boldsymbol{\theta})\right]$$

where $\boldsymbol{\theta}$ is the corresponding parameter vector for each model, and $p(\boldsymbol{Y}_j^u \,|\boldsymbol{Y}_j, \boldsymbol{\theta})$ has a conditional normal distribution from the joint distribution of the process at unobserved locations $\boldsymbol{Y}_j^u$ and the process at sampled sites $\boldsymbol{Y}_j$ which is given by,

$$\begin{pmatrix} \boldsymbol{Y}_j^u \\ \boldsymbol{Y}_j \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_j^u \\ \boldsymbol{\mu}_j \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_j^u & \boldsymbol{\Psi}_j' \\ \boldsymbol{\Psi}_\mathrm{j} & \boldsymbol{\Sigma}_j \end{pmatrix}\right], \tag{3}$$

where $\boldsymbol{\mu}_j^u$ and $\boldsymbol{\mu}_j$ are the means of the unobserved and observed locations respectively; $\boldsymbol{\Sigma}_j^u$ is the covariance matrix among unobserved sites, $\boldsymbol{\Sigma}_j$ is the covariance matrix among observed sites, and $\boldsymbol{\Psi}_\mathrm{j}$ is the covariance matrix between observed and unobserved sites. From the properties of the partition of the multivariate normal distribution, it follows that

$$\boldsymbol{Y}_j^u|\boldsymbol{Y}_j, \boldsymbol{\theta} \sim N\left(\boldsymbol{\mu}_j^u + \boldsymbol{\Psi}_\mathbf{j}'\left(\boldsymbol{\Sigma}_j\right)^{-1}\left(\boldsymbol{Y}_j - \boldsymbol{\mu}_j\right);\ \boldsymbol{\Sigma}_j^u - \boldsymbol{\Psi}_j'\left(\boldsymbol{\Sigma}_j\right)^{-1}\boldsymbol{\Psi}_\mathbf{j}\right). \tag{4}$$

For models 2 and 4 where there is no spatial component, this distribution is reduced to $\boldsymbol{Y}_j^u|\,\boldsymbol{\theta} \sim N\left(\boldsymbol{\mu}_j^u;\ \boldsymbol{\Sigma}_j^u\right)$, where $\boldsymbol{\Sigma}_j^u$ is a diagonal matrix.

We sampled from $p(\boldsymbol{Y}_j^u|\boldsymbol{Y}_j, \boldsymbol{\theta}^{(g)})$ for each $\boldsymbol{\theta}^{(g)}$ drawn from the posterior distribution $p\left(\boldsymbol{\theta}\middle|\boldsymbol{Y}_j, \boldsymbol{X}\right)$, and used composition sampling to obtain samples from the posterior predictive distribution.

To obtain the predictive surfaces, we first predicted at the centroids of a coarser grid with a 1 km by 1 km grid cell. Additionally, after obtaining the posterior mean for each campaign at each grid cell, we computed the mean across campaigns for each grid cell, a similar procedure can be followed to obtain the median. Then, since computing the predicted values for finer grids can be computationally demanding, to obtain smoother surfaces, we fitted a general additive model (GAM) with the posterior means at each cell as a response variable and the coordinates of the

centroid of each grid cell as the predictors. Finally, using the results of these GAMs we predicted the values of the process at the centroids of a finer grid (0.25 km by 0.25 km).

3.6 Results

3.6.1 Volatile Organic Compounds

For the December 2005 campaign, there were 8 measurements below the detection limit for n-decane, and in the April and August 2006 campaigns, there were 3 and 2 measurements below the detection limit for n-decane and hexane, respectively. Given that there are only few measurements below the limit of detection, these measurements were excluded from the analysis. Additionally, only the locations that were present in all three campaigns were analyzed. Therefore, the analysis included 127 monitoring locations for hexane, 121 for n-decane, and 129 for ethylbenzene, 1,2,4-trimethylbenzene, and benzene.

The concentration of benzene ranged from 0.4 to 5.27 $\mu g/m^3$, n-decane ranged between 0.13 and 5.25 $\mu g/m^3$, ethylbenzene ranged between 0.59 and 27.47 $\mu g/m^3$, hexane ranged between 0.4 and 32.03 $\mu g/m^3$, and 1,2,4-trimethylbenzene between 0.36 and 2.22 $\mu g/m^3$ (Table 1). The largest variability for all the VOCs was found in December, except for ethylbenzene which had a greater variability in August. Hexane had the highest concentrations across campaigns, followed by ethylbenzene.

Table 1 Selected moments of the distributions of benzene, n-decane, ethylbenzene, hexane and 1,2,4-trimethylbenzene levels (in $\mu g/m^3$) across three sampling campaigns in Montreal, between 2005 and 2006

	Benzene				n-Decane			
	December	April	August	Average	December	April	August	Average
Mean	1.35	1.30	0.56	1.07	2.08	2.05	0.95	1.69
Median	1.28	1.18	0.49	0.98	1.92	1.96	0.88	1.61
Std. dev.	0.61	0.60	0.31	0.44	0.73	0.66	0.44	0.42
Minimum	0.40	0.68	0.18	0.54	0.25	0.89	0.13	1.01
Maximum	4.72	5.27	2.51	3.35	4.27	5.25	3.31	3.27
	Ethylbenzene				Hexane			
	December	April	August	Average	December	April	August	Average
Mean	3.63	2.77	2.03	2.81	14.25	6.35	1.57	7.39
Median	3.23	2.65	1.82	2.62	13.78	5.83	1.42	7.07
Std. dev.	2.15	0.92	2.3	1.62	5.21	3.8	0.83	2.11
Minimum	1.12	1.22	0.59	1.30	2.24	2.32	0.40	3.32
Maximum	23.70	8.90	27.47	20.02	30.77	32.03	5.56	16.79

	1,2,4-Trimethylbenzene			
	December	April	August	Average
Mean	1.14	0.98	1.00	1.04
Median	1.10	0.98	0.97	0.99
Std. dev.	0.38	0.22	0.26	0.23
Minimum	0.36	0.53	0.53	0.63
Maximum	2.16	1.85	2.22	1.85

3.6.2 Model Comparison and diagnostics

To choose the best model among the proposed ones for each VOC, we used the minimum WAIC (Table 2). For benzene and n-decane, the selected model was one with spatial structure and an indicator variable to capture the variability across campaigns (Model 1). For ethylbenzene, hexane and 1,2,4-trimethylbenzene the selected model had different coefficients for the land-use variables across campaigns and did not have an additional spatial structure (Model 4).

We also computed the observed versus the predicted values for each VOC and we did not identify important outliers (Figure 1). Although not shown here, we found no important patterns in the residuals for each campaign after accounting for the covariates and spatial structure.

Table 2 WAIC of the fitted models for each VOC. Bold values (minimum WAIC) identify the selected models.

Model	Benzene	n-Decane	Ethylbenzene	Hexane	1,2,4-trimethylbenzene
Model 1	***119.64***	***331.18***	235.37	448.28	-19.80
Model 2	190.47	363.36	269.78	417.39	-13.58
Model 3	151.23	382.25	260.87	496.53	32.20
Model 4	187.05	348.87	***219.44***	***415.02***	***-63.02***

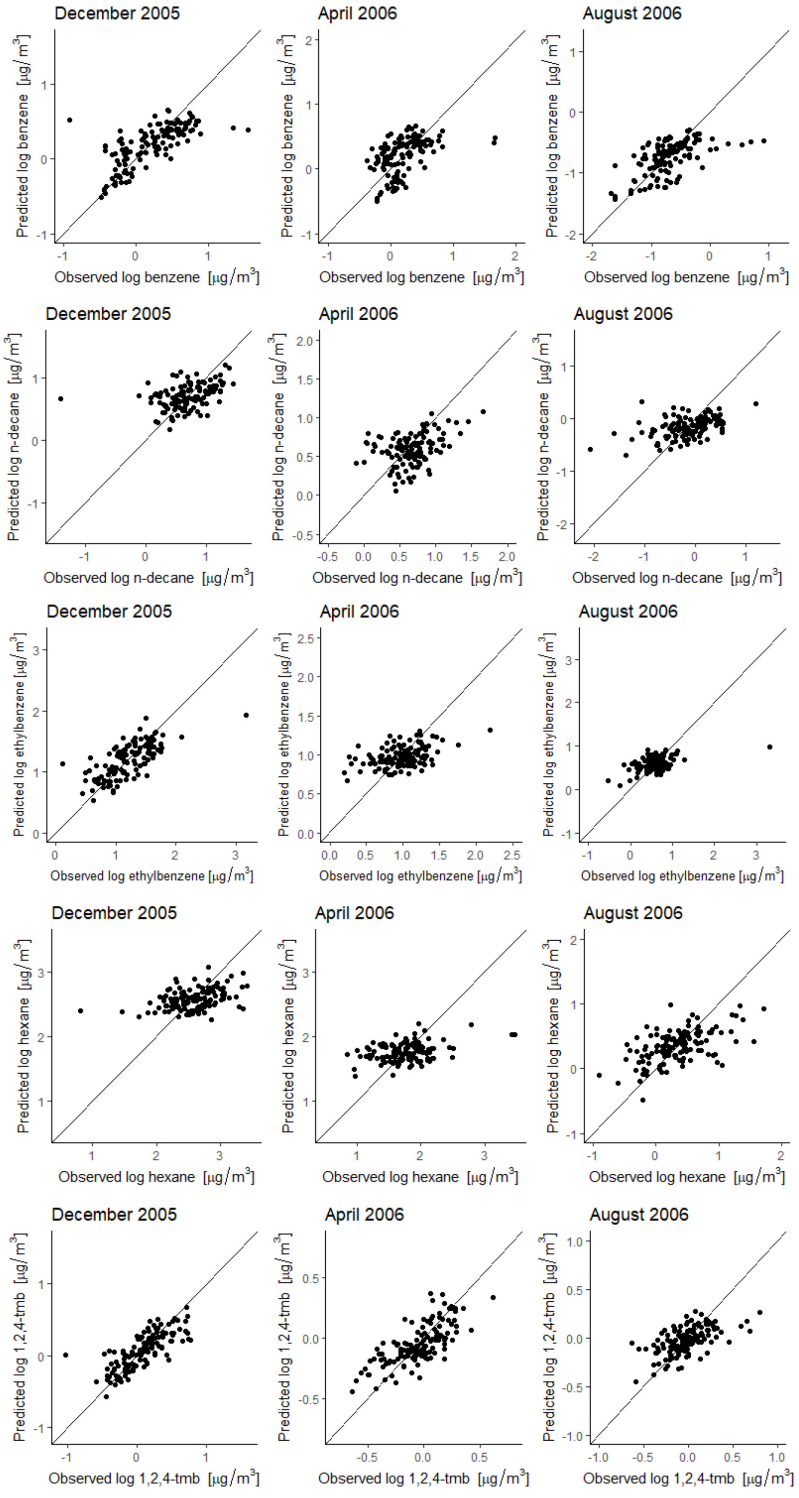


Figure 1 Scatter plots of the observed versus fitted values for benzene, n-decane, ethylbenzene, hexane, and 1,2,4-trimethylbenzene using the selected models (Table 2). The straight line represents perfect prediction.

For benzene and n-decane (see Supplementary Material Tables 3 - 6), we obtained different intercepts for each campaign, with August having the lowest values. For benzene (see Supplementary Material Tables 3 and 4), the spatial variance and practical range also changed across campaigns, suggesting that the spatial structure was different for each of the campaigns, with the highest spatial variance in the August campaign and similar values for the December and April campaigns. In the case of n-decane (see Supplementary Material Tables 5 and 6), we obtained similar values for the spatial variance but different values for the practical range, with a posterior mean of 8.82 km and 1.53 km for the April and December campaigns respectively.

For ethylbenzene, hexane, and 1,2,4-trimethylbenzene (see Supplementary Material Tables 7 - 9), some of the coefficients associated with the land-use variables differed in magnitude and direction across campaigns. This might be due to seasonal effects that can affect the relationship between the land-use variables and the pollutants levels.

Figure 2 shows the predicted surfaces obtained for a grid with a 0.25 by 0.25 km cell size. We also obtained the standard deviation of the posterior predictive distribution showing higher uncertainty in areas where no monitors were located (see Supplementary Material Figures 1-5).

The predicted surfaces for all the analyzed VOCs resulted in higher concentrations across the island for the December campaign, especially in the north part of the island. The predicted surfaces for benzene showed higher concentrations in the northeast part of the island with December and April showing similar levels. The mean predicted surface also showed the highest values at the northeast part of Montreal. For benzene and n-decane higher values were

found during the April campaign.We also found higher concentrations in the central part of the island and along some of the most important highways in Montreal for n-decane and 1,2,4-trimethylbenzene. The predicted surface for hexane shows the highest concentrations for the December campaign with little variability across the island during all three campaigns.

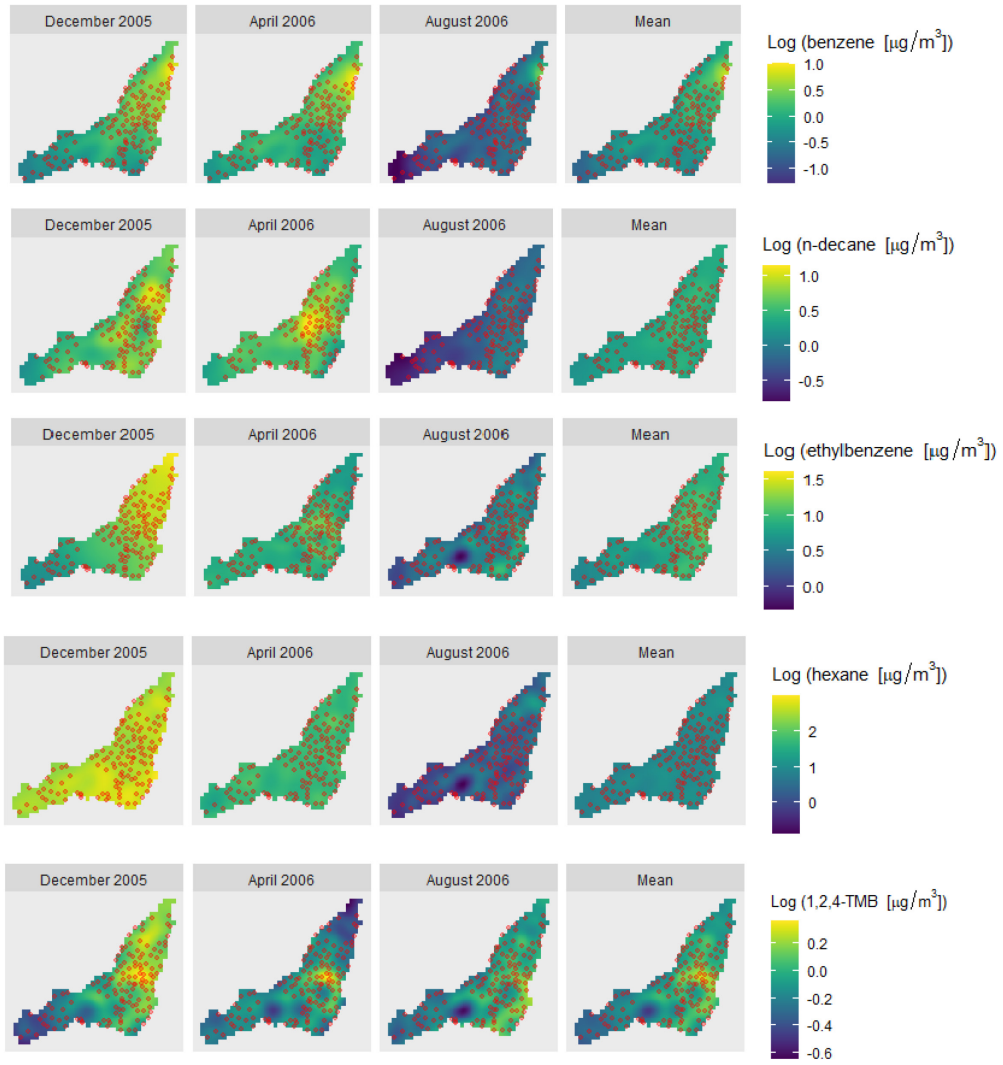


Figure 2 Posterior mean of the predicted surfaces in the log scale for benzene, n-decane, ethylbenzene, hexane, and 1,2,4-trimethylbenzene concentration at each campaign. Red solid circles represent the locations of the monitors.

3.7 Discussion

We fitted spatial regression models to data from three dense sampling campaigns in Montreal. Land-use variables were used to determine the predicted concentrations of the five selected VOCs. For each VOC, we selected the model that explained the most variability in the data, accounting for the complexity of the model, and found reasonably good fits to the data (Figure 1). From these models, we then predicted the concentrations of each VOC across the island at a resolution of 0.25 by 0.25 km cell grid size.

We found that the highest concentrations for all five VOCs were during the December campaign, followed by April and August. Meteorological conditions such as anticyclones in colder weather, leading to stagnant meteorological conditions, might have facilitated the accumulation of these air pollutants during the December campaign. Higher concentrations during winter have also been found in other cities for ozone and different VOCs [90, 91, 109, 110]. Additionally, higher levels of benzene and hexane were observed during this period compared to levels in other Canadian cities [53, 55].

We also found that the spatial distribution of the different VOCs changed by season, therefore, it is not recommended to use one season as representative of the annual exposure, as shown in a previous study [55].

There is face validity to our results. Specifically, for benzene we found that areas located in the northeast part of Montreal where the highest benzene levels were predicted correspond to an area with oil refineries operating at the time the study was conducted. Additionally, we found that, for

n-decane the highest levels were predicted in sections of a major highway (Autoroute 40), especially in the North. As well, for 1,2,4-trimethylbenzene, the predicted areas with the highest levels corresponded to multiple sections of several highways (Autoroute 40, Autoroute 136, and Autoroute 15).

3.7.1 Strengths and limitations

We used 3M passive monitors because of ease in installing on fixed city poles at 10-feet heights, they did not require electricity or pumps, if stolen they would not be costly, and they are thought to be sufficiently accurate and precise. Passive samplers have been shown to be reliable in measuring VOCs over extended periods of time [111]. Furthermore, an analysis of 3M organic passive dosimeters outdoors using a sampling duration of 72 hours was comparable to automated continuous gas chromatography measurements [112]. Other methods could have been used that could lead to more accurate estimates, such as passivated Summa canisters and flame and photoionization detectors, but they are not suitable for remote sites without electricity, their operation is difficult in cold weather, they require knowledge of the proportions of concentrations of the different VOCs, and these methods are expensive.

The present study sampled a considerably larger number of sites than some previous studies [22, 26, 30, 53, 55, 90-92, 110]. Additionally, both the spatial and campaign variability was accounted for by the model instead of averaging the data across campaigns. By obtaining the predicted surface for several VOCs we not only facilitate the comparison of pollutant levels across campaigns but also across VOCs.

As the main goal of the present study was prediction, the results presented here should not be used to identify the relationship between land-use variables and VOC concentrations. To meet this goal, one needs to adjust the method for variable selection by testing for collinearity and confounding, and possibly changing the set of predictors for each campaign. Additionally, it would also be valuable to account for meteorological covariates such as temperature, wind speed, humidity, and atmospheric pressure.

Finally, this study shares the same limitations of other LUR methods namely that the results are case and area specific, therefore they are only valid for this area of study [37].

3.7.2 Conclusions

In the present study, we obtained predicted surfaces showing the spatial variability of each of the studied VOCs. We found higher concentrations of VOCs in the east and central part of Montreal with higher concentrations during the winter campaign.

We proposed four models, each of which accounted for spatial and campaign variability. The model that fitted the best, according to WAIC, for benzene and n-decane accounted for the spatial structure after adjusting for the land-use variables, and for seasonal variability through the intercept. For hexane, ethylbenzene, and 1,2,4-trimethylbenzene, land-use covariates alone accounted for the spatial variability, and the campaign variability was accounted for through the coefficients associated with the land-use variables.

Inference was performed under the Bayesian framework, therefore it was straightforward to obtain summaries of the predictive posterior distribution such that the spatial interpolation to unobserved locations of interest naturally accounted for the uncertainty in the estimation of the unknowns in the model.

The proposed models are flexible to adjust for any set of land-use variables or air pollutants concentrations, and the methods are easily reproducible. The predicted surfaces obtained here, and the spatial interpolation methods used in this study, can help estimate the air pollutant levels at residential addresses of participants for health studies.

References

7. Reimann S, Lewis AC. Anthropogenic VOCs. Volatile Organic Compounds in the Atmosphere2007. p. 33-81.

22. Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, et al. Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science. 1997;11(7):699-718.

26. Oiamo TH, Johnson M, Tang K, Luginaah IN. Assessing traffic and industrial contributions to ambient nitrogen dioxide and volatile organic compounds in a low pollution urban environment. Science of The Total Environment. 2015;529:149-57.

27. Crouse DL, Goldberg MS, Ross NA. A prediction-based approach to modelling temporal and spatial variability of traffic-related air pollution in Montreal, Canada. Atmospheric Environment. 2009;43(32):5075-84.

30. Atari DO, Luginaah IN. Assessing the distribution of volatile organic compounds using land use regression in Sarnia," Chemical Valley", Ontario, Canada. Environmental Health. 2009;8(1):1-14.

37. Amini H, Yunesian M, Hosseini V, Schindler C, Henderson SB, Künzli N. A systematic review of land use regression models for volatile organic compounds. Atmospheric Environment. 2017;171:1-16.

43. Gamerman D, Lopes HF. Markov chain Monte Carlo : stochastic simulation for Bayesian inference. 2nd ed. / ed. Boca Raton: Taylor & Francis; 2006.

44. de Valpine PaP, C. and Turek, D. and Michaud, N. and Anderson-Bergman, C. and Obermeyer, F. and Wehrhahn Cortes, C. and Rodriguez, A. and Temple Lang, D. and Paganin,

S. NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. 0.11.1 ed2021.

46. Banerjee S, Gelfand AE, Carlin BP. Hierarchical modeling and analysis for spatial data. Boca Raton, Florida Chapman & Hall/CRC Press; 2015.

52. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol. 2005;15(2):185-204.

53. Su JG, Jerrett M, Beckerman B, Verma D, Arain MA, Kanaroglou P, et al. A land use regression model for predicting ambient volatile organic compound concentrations in Toronto, Canada. Atmospheric Environment. 2010;44(29):3529-37.

55. Wheeler AJ, Smith-Doiron M, Xu X, Gilbert NL, Brook JR. Intra-urban variability of air pollution in Windsor, Ontario--measurement and modeling for human exposure assessment. Environ Res. 2008;106(1):7-16.

82. Williams J, Koppmann R. Volatile Organic Compounds in the Atmosphere: An Overview. Volatile Organic Compounds in the Atmosphere2007. p. 1-32.

83. United States Environmental Protection Agency. Volatile organic compounds' impact on indoor air quality 2014. Available from https://www.epa.gov/indoor-air-quality-iaq/volatile-organic-compounds-impact-indoor-air-quality.

84. Centre international de recherche sur le cancer, IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Benzene. Lyon France; 2019.

85. Canada's air pollutant emissions inventory report. Environment and Climate Change Canada; 2021.

86. Gurney KR, Liang J, Roest G, Song Y, Mueller K, Lauvaux T. Under-reporting of greenhouse gas emissions in U.S. cities. Nature Communications. 2021;12(1):553.

87. Ren X, Salmon OE, Hansford JR, Ahn D, Hall D, Benish SE, et al. Methane Emissions From the Baltimore-Washington Area Based on Airborne Observations: Comparison to Emissions Inventories. Journal of Geophysical Research: Atmospheres. 2018;123(16):8869-82.

88. Turner AJ, Jacob DJ, Benmergui J, Wofsy SC, Maasakkers JD, Butz A, et al. A large increase in U.S. methane emissions over the past decade inferred from satellite data and surface observations. Geophysical Research Letters. 2016;43(5):2218-24.

89. Yifang Zhu WCH, Seongheon Kim, Si Shen, Constantinos Sioutas. Study of ultrafine particles near a major highway with heavy-duty diesel traffic. Atmospheric Environment. 2002;36(27):12.

90. Bozkurt Z, Üzmez ÖÖ, Döğeroğlu T, Artun G, Gaga EO. Atmospheric concentrations of SO2, NO2, ozone and VOCs in Düzce, Turkey using passive air samplers: Sources, spatial and seasonal variations and health risk estimation. Atmospheric Pollution Research. 2018;9(6):1146-56.

91. Kumar A, Singh D, Kumar K, Singh BB, Jain VK. Distribution of VOCs in urban and rural atmospheres of subtropical India: Temporal variation, source attribution, ratios, OFP and risk assessment. Sci Total Environ. 2018;613-614:492-501.

92. Li B, Ho SSH, Qu L, Gong S, Ho KF, Zhao D, et al. Temporal and spatial discrepancies of VOCs in an industrial-dominant city in China during summertime. Chemosphere. 2021;264(Pt 2):128536.

93. Basagaña X, Rivera M, Aguilera I, Agis D, Bouso L, Elosua R, et al. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmospheric Environment. 2012;54:634-42.

94. Crouse DL, Goldberg MS, Ross NA, Chen H, Labrèche F. Postmenopausal Breast Cancer Is Associated with Exposure to Traffic-Related Air Pollution in Montreal, Canada: A Case Control Study. Environmental Health Perspectives. 2010;118(11):1578-83.

95. Goldberg MS, Labrèche F, Weichenthal S, Lavigne E, Valois M-F, Hatzopoulou M, et al. The association between the incidence of postmenopausal breast cancer and concentrations at street-level of nitrogen dioxide and ultrafine particles. Environmental Research. 2017;158:7-15.

96. Parent M-É, Goldberg MS, Crouse DL, Ross NA, Chen H, Valois M-F, et al. Traffic-related air pollution and prostate cancer risk: a case–control study in Montreal, Canada. Occupational and Environmental Medicine. 2013;70(7):511-8.

97. Statistics Canada. Census Profile, 2016 Census 2019. Available from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CMACA&Code1=462&Geo2=PR&Code2=01&Data=Count&SearchText=Montreal&SearchType=Begins&SearchPR=01&TABID=1&B1=All.

98. Government of Canada. Canadian Climate Normals 1981-2010 Station Data 2021. Available from https://climate.weather.gc.ca/climate_normals/results_1981_2010_e.html?searchType=stnProv&lstProvince=QC&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&stnID=5415&dispBack=0.

99. Environmental Assessment Report Air Quality Montreal. Ville de Montréal, Service de l'environnement, Division de la planification et du suivi environnemental, Réseau de surveillance de la qualité de l'air (RSQA); 2018.

100. Boulet D, Melançon S. Environmental Assessment Report Air Quality Montreal. Ville de Montréal, Service des infrastructures, du transport et de l'environnement, Direction de l'environnement, Division de la planification et du suivi environnemental, RSQA; 2012.

101. Airzone. Available from https://www.airzoneone.com/lab-analysis/

102. DMTI Spatial. CanMap® GIS Data for GIS Mapping Software. Available from https://www.dmtispatial.com/canmap/

103. Xie X, Semanjski I, Gautama S, Tsiligianni E, Deligiannis N, Rajan R, et al. A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods. ISPRS International Journal of Geo-Information. 2017;6(12).

104. Ramos Y, Requia WJ, St-Onge B, Blanchet JP, Kestens Y, Smargiassi A. Spatial modeling of daily concentrations of ground-level ozone in Montreal, Canada: A comparison of geostatistical approaches. Environ Res. 2018;166:487-96.

105. PTV Vissim. Traffic Simulation Software: PTV Group. Available from https://www.ptvgroup.com/en/solutions/products/ptv-vissim/

106. United States Environmental Protection Agency. MOVES and Other Mobile Source Emissions Models. Available from http://www.epa.gov/moves.

107. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning : with applications in R: New York : Springer, [2013] ©2013; 2013.

108. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing. 2014;24(6):997-1016.

109. Gu P, Dallmann TR, Li HZ, Tan Y, Presto AA. Quantifying Urban Spatial Variations of Anthropogenic VOC Concentrations and Source Contributions with a Mobile Sampling Platform. Int J Environ Res Public Health. 2019;16(9).

110. Yang Y, Liu X, Zheng J, Tan Q, Feng M, Qu Y, et al. Characteristics of one-year observation of VOCs, NOx, and O3 at an urban site in Wuhan, China. J Environ Sci (China). 2019;79:297-310.

111. Shields HC, Weschler CJ. Analysis of Ambient Concentrations of Organic Vapors with a Passive Sampler. JAPCA. 1987;37(9):1039-45.

112. Stock TH, Morandi MT, Afshar M, Chung KC. Evaluation of the Use of Diffusive Air Samplers for Determining Temporal and Spatial Variation of Volatile Organic Compounds in the Ambient Air of Urban Communities. Journal of the Air & Waste Management Association. 2008;58(10):1303-10.

4 Within city spatiotemporal variation of pollen concentration in the city of Toronto, Canada

4.1 Preface to Manuscript 2

In this chapter, Bayesian hierarchical models are used to model the weekly concentration of grass, weed, tree, and total pollen within a city.

Land-use regression methods are extended by considering the overall mean variability of pollen concentration across the city. Additionally, the high number of measurements equal to zero are considered using a hurdle model.

The predicted surfaces obtained in this study will be used in future health studies.

This manuscript has been published in the *Environmental Research Journal*.

4.2 Authors contribution

Sara Zapata-Marin: Formal analysis, Methodology, Writing - Original Draft

Alexandra M. Schmidt: Methodology, Supervision, Writing - Review & Editing

Scott Weichenthal: Conceptualization, Writing - Review & Editing

Daniel S.W. Katz: Data acquisition, Writing - Review & Editing

Tim Takaro: Writing - Review & Editing

Jeffrey Brook: Writing - Review & Editing

Eric Lavigne: Funding acquisition, Data acquisition, Writing - Review & Editing

Within city spatiotemporal variation of pollen concentration in the city of Toronto, Canada

Sara Zapata-Marin [a], Alexandra M. Schmidt [a,b], Scott Weichenthal [b], Daniel S.W. Katz [c], Tim Takaro [d], Jeffrey Brook [e] and Eric Lavigne [f]

[a] Quantitative Life Sciences Program, McGill University, Montreal, QC, Canada.

[b] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada

[c] Dell Medical School, University of Texas at Austin, Austin, Texas, USA

[d] Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada

[e] Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada

[f] Air Health Science Division and Population Studies Division, Health Canada, Ottawa, Ontario, Canada

Corresponding author: Sara Zapata-Marin [a] sara.zapata-marin@mail.mcgill.ca

4.3 Abstract

Background: The exacerbation of asthma and respiratory allergies has been associated with exposure to aeroallergens such as pollen. Within an urban area, tree cover, level of urbanization, atmospheric conditions, and the number of source plants can influence spatiotemporal variations in outdoor pollen concentrations.

Objective: We analyze weekly pollen measurements made between March and October 2018 over 17 sites in Toronto, Canada. The main goals are: to estimate the concentration of different types of pollen across the season; estimate the association, if any, between pollen concentration and environmental variables, and provide a spatiotemporal surface of concentration of different types of pollen across the weeks in the studied period.

Methods: We propose an extension of the land-use regression model to account for the temporal variation of pollen levels and the high number of measurements equal to zero. Inference is performed under the Bayesian framework, and uncertainty of predicted values is naturally obtained through the posterior predictive distribution.

Results: Tree pollen was positively associated with commercial areas and tree cover, and negatively associated with grass cover. Both grass and weed pollen were positively associated with industrial areas and TC brightness and negatively associated with the northing coordinate. The total pollen was associated with a combination of these environmental factors. Predicted surfaces of pollen concentration are shown at some sampled weeks for all pollen types.

Significance: The predicted surfaces obtained here can help future epidemiological studies to find possible associations between pollen levels and some health outcome like respiratory allergies at different locations within the study area.

Keywords: Bayesian inference, spatial distribution, temporal variation, land-use regression

4.4 Introduction

Aeroallergens and air pollutants have been linked to the exacerbation and prevalence of allergic rhinitis and allergic asthma [18, 113-115], particularly in urban areas [116, 117]. It has been estimated that 3.8 million Canadians suffer from asthma [118], 75% of whom also suffer from respiratory allergies [110]. Increases in outdoor aeroallergen levels were found to be associated with the number of hospital admissions for asthma, additionally enhanced by higher air pollution levels [119-124]. Higher CO_2 levels and warmer temperatures were shown to affect the growth and pollen production levels in different plants [125]. Additionally, it was suggested that climate change can affect the start, duration, and intensity of pollen season [75, 126-128]. For example, it was shown that urbanization-induced environmental change can modify the characteristics of tree [75] and ragweed [129] pollen seasons. These environmental changes are likely to affect the epidemiology of both asthma and respiratory allergies [127, 130]. Moreover, living in urban areas is a known risk factor for the development of respiratory allergies induced by aeroallergens [16, 126]. Little is known about the spatial distribution of pollen concentrations within a city as pollen levels are typically recorded at a single monitoring site. However, previous studies have argued that the information from one site is not representative of the pollen levels across the city, and relying on this data might lead to measurement error in any subsequent epidemiological analysis [25, 75].

In recent years, several studies have analyzed spatial variations in pollen concentrations within cities using networks of pollen monitoring sites. The number of monitoring sites used varies from five or fewer [131, 132] up to 25 sites [25]. For example, grass pollen was monitored in a network of 16 locations in the Metropolitan Helsinki Area to study the relationship between the

level of urbanization and pollen concentrations [35]. Another network of 14 sites was used in Berlin to monitor spatiotemporal variations in birch, grass, and mugwort pollen and to evaluate the influence of local sources and impacts on allergy symptoms and severity [68, 69]. More recently, a study of spatiotemporal variations in pollen concentrations in Detroit with 25 monitoring sites showed considerable heterogeneity of pollen levels across locations. All these studies support the idea that a single monitoring site does not reflect outdoor concentrations across a city [25]. However, to our knowledge, none of the studies conducted to date have developed models to account for spatiotemporal variations in pollen concentrations throughout the pollen season as well as the influence of land-use variables on pollen concentration at a local scale.

This study aims to capture both the spatial and temporal variability of aeroallergens concentration in Toronto, Canada for different types of pollen. Data were weekly collected from 17 sites across Toronto. The proposed model developed from these data expands the land use regression (LUR) methodology by adding a time-varying mean to capture the overall temporal variability of pollen concentration across weeks during the 2018 pollen season. The main goals of this study are: 1) to estimate spatial variations of pollen concentrations across unobserved locations within Toronto; 2) to identify important predictors that might be associated with the intra-urban variation of grass, weed, tree, and total pollen concentration; 3) to capture the temporal trend across weeks over the studied region; and 4) to account for the high number of zeros that help identify the temporal windows with higher concentrations of each pollen type for the sampled period.

4.5 Materials and methods

4.5.1 Study Area

The study was conducted in the city of Toronto (2,731,571 inhabitants), the provincial capital of Ontario (43°44′30″N 79°22′24″W). One of the characteristic features of the city is the Toronto ravine system, which forms a large urban forest that runs throughout the city. Toronto has the characteristics of a humid continental climate. The city experiences four distinct seasons with temperatures below 0°C during the winter, above 20°C during the summer, annual precipitation of 831 mm, and an annual snowfall of 1,220 mm.

Plants of allergenic concern that can be found in this zone include different types of weeds, grasses, and trees. Some allergenic weeds in this region are mugwort (*Artemisia*) and ragweed (*Ambrosia*) which have been identified as allergenic sources [62, 133]. Mugwort thrives in partial shade areas and ragweed requires abundant sunlight to thrive, therefore they can be found along roadside edges, garbage dumps, agricultural croplands, and areas with low tree cover. Weeds from the Amaranthaceae family (*Chenopodiaceae*, *Amaranthaceae*) and plantain (*Plantago*) were also found in the samples.

Two types of grasses were also sampled, the true grasses (*Graminae*), and the sedge family grasses (*Cyperaceae*). Some of the true grasses include Ryegrass (*Lolium*) and orchard (*Dactylis*) grass, which are two common types of grass that can be found in this area due to their high tolerance to cool weather. Other grass types from this family include fescue (*Festuca*), brome and chess (*Bromus*), timothy (*Phleum*), and sweet vernal (*Anthoxanthum*) [62].

Some deciduous trees that were sampled include birch (*Betula*), alder (*Alnus*), aspen (*Populus*), willow (*Salix*), maple (*Acer*), white and red oaks (*Quercus*), beech (*Fagus*), elms (*Ulmus*), ash (*Fraxinus*), sycamore (*Platanus*), among others.[18, 62].

4.5.2 Data Collection

Aeroallergen data: Pollen grains were monitored at 17 different sites (1.5 m above ground level) across the city of Toronto during the 2018 pollen season (Figure 1). The monitoring locations differed from each other with respect to land use and vegetation (see Figures 1 to 5 in the Supplementary Material).

Potential monitoring locations were identified using location attribution in order to obtain locations with higher variability based on land-use predictors. Ultimately, the chosen locations relied on volunteers in these identified areas. For 16 of these locations, sampling was done in the volunteers backyards, and one monitor was located at Environment Canada.

Sampling was conducted daily for 10 sites and weekly for the remaining 7 sites. The weekly mean concentration for the 10 daily sites was obtained by averaging over the available observations for the daily monitoring sites. The data were collected between March 11 and October 7, 2018.

Rotation impaction samplers provided by Aerobiology Research Laboratory [77] were used to collect pollen. To avoid oversampling, for the daily measurements the sampler would collect pollen for a full minute every 10 minutes over 24 hours, therefore each measurement is from 144 minutes of sampling per day. For the weekly measurements, 144 minutes of sampling were

conducted over the week. For this type of samplers, over 24 hours, the volume of air sampled is 6.8 m^3 [134].

The collected data included the level of different types of tree, weed, and grass pollen but only the total concentration of each pollen type was analyzed. Additionally, the total pollen concentration was computed as the sum of all three pollen types. Pollen measurements were converted into their volumetric equivalents expressing the aeroallergen concentration as pollen grains per cubic meter of air sampled.

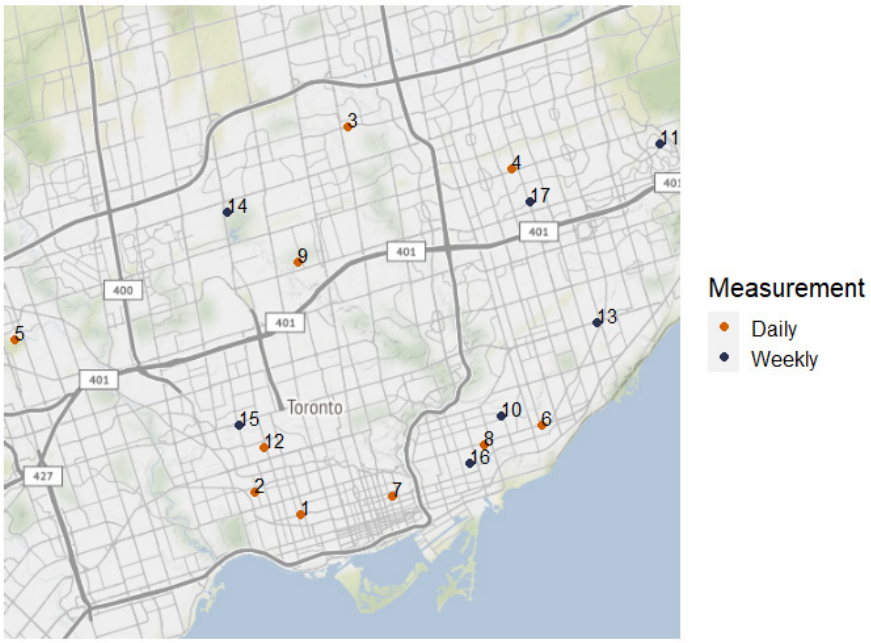


Figure 1 Location of the n=17 pollen-monitoring sites that measured concentration of different types of pollen in the city of Toronto between March 11 and October 7, 2018.

Environmental determinants: We used buffers as a measure of specific land-use covered in a circle of specific radius. Potential predictors of local pollen levels were computed at 50-, 100-, 200-, 500-, and 1000-m buffer sizes around each sampling site (Figure 1). These buffer sizes were selected based on the findings of previous studies [35, 36, 75]. Additionally, we used the

standardized northing and easting coordinates in Universal Transverse Mercator projection (UTM) for each site to account for the spatial structure. The final environmental determinants used for the models included two remote sensing-based indexes, one for greenness, and one for brightness (monthly mean tasseled cap (TC) greenness and tasseled cap brightness) obtained from Landsat data, five land-use variables (commercial, grass, industrial, major roads, and tree cover) obtained from DMTI Spatial Inc. and the City of Toronto tree canopy study in 2018, and three climatic variables (mean humidity, mean precipitation, and mean temperature) retrieved from one monitoring station from Environment Canada (see Supplementary Material Table 1). The values of the land-use determinants were computed using ArcGIS 10.3 and ArcMap 10.7 for the different buffer sizes.

Data on Normalized Difference Vegetation Index (NDVI) was also available but based on a preliminary analysis, TC greenness had a higher correlation with all pollen types than NDVI, and therefore NDVI was not used in the analysis.

4.5.3 Statistical Analysis

Typically, when analyzing pollen concentration data, only the window of time with concentration higher than zero is analyzed. However, accounting for the high number of measurements equal to zero can potentially provide additional information. To this end, we considered a hurdle model [50] that accounts for the probability of observing pollen levels at week t and location s equal to zero. The first part of the hurdle model provides an estimate of the probability of the pollen concentration being equal to zero for each week, allowing the estimation of the start and end of the season for each pollen type. The second part, models the

positive values which are assumed to follow some probability distribution with support on the positive real numbers.

More specifically, the proposed hurdle model is a mixture between a Bernoulli probability function and, in this case, as the concentrations are strictly positive, the logarithm of the pollen levels follow a normal probability density function $\mathrm{p}(\mathrm{y_t(s)}|\mu_t(s), \tau^2)$ where $y_t(s)$ is the particle concentration measured in grains per m^3 at time *t* and location *s.* The probability density function for each pollen concentration at week *t* and location *s* is defined as,

$$p(y_t(s)|\rho_t, \mu_t(s), \tau^2) = \begin{cases} \rho_t, & y_t(s) = 0 \\ (1-\rho_t)p(y_t(s)|\,\mu_t(s), \tau^2), & y_t(s) > 0 \end{cases} \quad (1)$$

where ρ_t is the probability of the measurement being equal to zero at week *t* and location *s* and the logarithm of the positive values follow a normal distribution $\log(y_t(s)) \sim N(\mu_t(s), \tau^2)$ with variance τ^2 and mean $\mu_t(s)$ modeled as the sum of four different components

$$\mu_t(s) = \theta_t + \alpha x(s) + \beta u_t(s) + \zeta z_t. \quad (2)$$

The spatial component of the mean consists of the vector $x(s)$ that contains the coordinates, and the land-use variables at each location *s*; the temporal component of the mean consists of a latent (non-observable) mean varying level θ_t at week *t*, which captures the overall mean pollen concentration across all locations, a temporal component that captures the association of climatic factors at week *t* contained in the vector z_t , and the vector $u_t(s)$ that contains the TC brightness and TC greenness at each location *s* and week *t*. The mean varying level θ_t evolves smoothly with time through a random walk prior distribution, such that $\theta_t \sim N\left(\theta_{t-1}, {W_1}^2\right)$, that is, at week *t,* θ_t follows a normal distribution with mean θ_{t-1} and variance W_1^2.

We fitted four different variants of the hurdle model in equation 1, which differ from each other in the way the probability ρ_t is modeled.

Model 1 considered a constant probability such that $\rho_t = \rho$ for all locations and instants in time. Model 2 considered a polynomial relationship between the logit of the probability ρ_t and a quadratic function of the week *t* such that,

$$\text{logit}(\rho_t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2. \quad (3)$$

In model 3, we considered a random walk structure for the logit ρ_t similar to the one for θ_t, such that,

$$\text{logit}(\rho_t) = \gamma_t, \quad (4)$$

and assumed, a priori, that $\gamma_t \sim N(\gamma_{t-1}, W_2^2)$. Finally, in model 4, we used a logit regression with the climatic factors as predictors of the probability, such that,

$$\text{logit}(\rho_t) = \gamma_0 + \gamma z_t. \quad (5)$$

For all the fitted models the probability ρ_t is assumed the same across locations. The inference procedure followed the Bayesian paradigm. In this case, model specification is complete after assigning a prior distribution to the parameter vector. The parameter vector is defined as $\Omega_1 = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \rho, \boldsymbol{\theta}, \tau^2, W_1\}$ for model 1, $\Omega_{2,4} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \tau^2, W_1\}$ for models 2 and 4, and $\Omega_3 = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \tau^2, W_1, W_2\}$ for model 3, where $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_T)$.We assumed all the parameters to be independent a priori except for θ_t and γ_t in model 3, which depend on θ_{t-1} and γ_{t-1}, respectively. At the initial time $t = 0$, it is assumed that θ_0 follows a normal distribution with know mean m_0 and reasonable large variance C_0, and a similar prior is assigned to γ_0 in model 3. We assigned a zero mean normal prior distribution with relatively large $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}$, to describe our

prior ignorance about the association between the available covariates and pollen concentration; we let $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$ to vary smoothly with time, so the variances $\tau^2, W_1^2,$ and W_2^2 should not assume big values Finally, for model 1, a uniform prior distribution in the interval (0, 1) was assigned to ρ. Additionally, for all pollen types, there were 35 missing values spread across the weeks. Since we followed the Bayesian paradigm throughout our analysis, the missing values, for each of the pollen types, become parameters to be estimated and are naturally incorporated in the inference procedure. Inference follows based on their predictive posterior distribution [135].

In summary, we fit four different models for each pollen type where we assume: a constant probability (Model 1), a polynomial function (Model 2), a random walk model (Model 3), and a regression on temporal covariates (Model 4). Regardless of the fitted model, the resultant posterior distribution does not have a closed form. Therefore, we use Markov Chain Monte Carlo methods [43] to obtain samples from the resultant posterior distribution. This analysis was conducted in R 3.6.3 [136] using the Stan software [137] through the rstan package [138]. The data were visualized using ggplot2 [139].

After fitting all four models, model choice was performed using the Watanabe–Akaike information criterion (WAIC) and leave-one-out cross-validation (LOO) which account for the goodness of fit and the complexity of the model [38, 108]. Smaller values of these criteria indicate the best model among the fitted ones.

The code used to fit these models is available online at https://github.com/SaraZM/WeeklyPollenToronto

4.6 Results

4.6.1 Pollen Data

For all pollen types, there were 35 measurements where data were missing. Most of these measurements were from site 16 due to one of the volunteers changing addresses before the end of the experiment.

Tree pollen: Before estimating the missing values, in our sample from March 11 to October 7, the highest mean tree pollen concentration was recorded at site 2, and the lowest mean tree pollen concentration was recorded at site 4. The difference between the highest and the lowest total tree pollen concentration was 563.21 %. The maximum tree pollen concentration of 8481.271 grains/ m^3 was recorded at site 2 on the week of May 20. Across sites, there were 198 recordings of concentrations equal to zero. The highest number of zeros were recorded at site 10 and the lowest was recorded at site 16. The highest number of zeros across sites were sampled from the week of July 29 to the week of September 30 at the end of the sampling period. The highest tree pollen concentration at all locations was reached between the week of May 13 and May 27. During the sampling period, the tree pollen concentration varied across sites (Figure 2).

Grass pollen: Before estimating the missing values, the highest mean grass pollen concentration was recorded at site 14, which was the monitor located at Environment Canada, whereas the lowest overall grass pollen concentration was recorded at site 8. The difference between the highest and the lowest total grass pollen concentration was 704.03% before estimating the missing values. The maximum total concentration of 325.40 grains/ m^3 was recorded at site 15 on the week of May 27. In total, there were 148 recordings of pollen levels equal to zero. The greatest proportion of zeros was recorded at site 11 and the smallest was recorded at sites 2 and 13. The greatest proportion of zeros across sites were recorded at the beginning of the sampling period, from the week of March

11 to the week of April 22. The highest grass pollen concentration across sites was reached between the week of May 27 and June 10. During the sampled season, the grass pollen concentration varied across sites (Figure 2).

Weed pollen: Before estimating the missing values, the greatest mean weed pollen concentration was recorded at site 14, and the lowest was recorded at site 3. The difference between the highest and the lowest total weed pollen concentration was 338.203%. The maximum weed pollen concentration of 222.4 grains/ m^3 was recorded at site 14 on the week of September 2. Across sites, there were 164 recordings of concentrations equal to zero. The greatest number of zeros were recorded at sites 10 and 15, while the lowest was recorded at sites 9 and 16. The highest amount of zeros happened from the week of March 11 to the week of May 6, at the beginning of the sampled period. The highest weed pollen concentration across sites was reached between the weeks of August 19 to September 2. During the sampled season, the weed pollen concentration varied across the city (Figure 2).

Total pollen: Before estimating the missing values, the highest mean total pollen concentration was recorded at site 2, and the lowest mean total pollen concentration was recorded at site 4. The difference between the highest and the lowest total pollen concentration was 441.42%. The maximum total pollen concentration of 222.4 grains/ m^3 was recorded at site 2 on the week of May 20. In total, there were 8 recordings of concentrations equal to zero. The greatest number of zeros was recorded at site 11 with most locations having none to one zero-measurements across the 31 weeks. The highest amount of zeros across weeks were sampled during the first week of sampling. The highest total pollen concentration at all sites was reached between the weeks of May 13 to May 27. During the entire period, the total pollen concentration was uniform across the city (Figure 2).

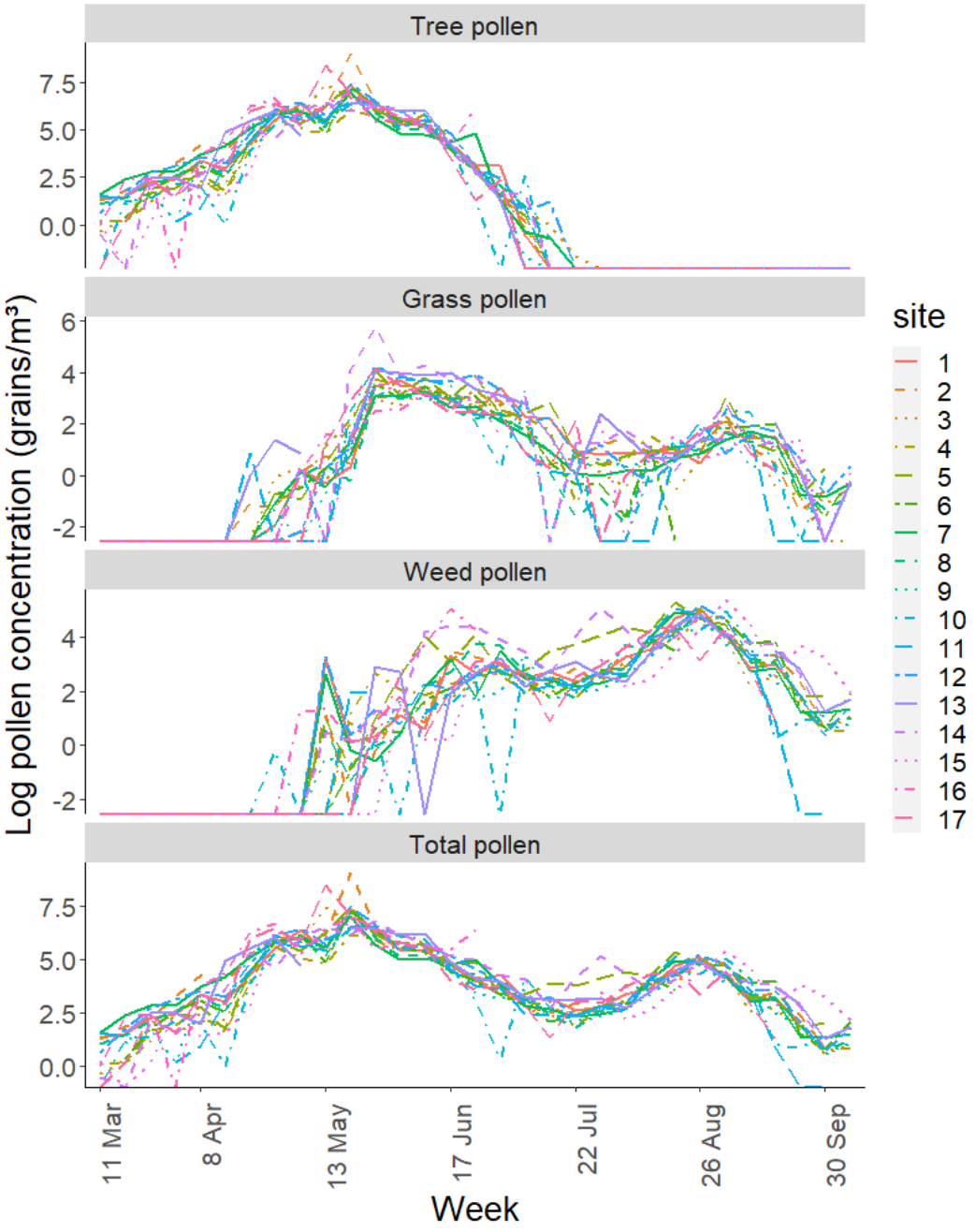


Figure 2 Weekly mean pollen grains per cubic meter for the 17 monitoring stations across 31 weeks during 2018 pollen season.

4.6.2 Optimization of the Buffer Sizes

For the statistical analysis, to facilitate the comparison between pollen types, we decided to use the same predictors and buffer sizes for all types of pollen. For each land-use predictor, we chose a buffer size that was highly correlated with all pollen types. In all four models, the vector $x(s)$ in Equation 2 contains the following scaled variables: commercial (buffer size = 1000m), easting, industrial (1000m), grass cover (500m), major roads (1000m), northing, tree cover (1000m), TC greenness (1000m), TC brightness (1000m), and the mean humidity, precipitation, and temperature. The covariates were not highly correlated with each other.

4.6.3 Model Comparison

After fitting the four models, we chose the best among the fitted ones for each pollen type based on WAIC and LOO (Supplementary material Table 2). For tree and grass pollen, the random walk model (Model 3) was the best among the fitted ones, while for total and weed pollen, the model with a regression on the temporal covariates (Model 4) was the best among the fitted ones.

4.6.4 Model Estimates

Figure 3 shows the posterior estimates (posterior means), 90% posterior credible interval (red line), and 95% posterior credible interval (black line) for the land-use variable coefficients ($\boldsymbol{\alpha}$), climatic coefficients ($\boldsymbol{\beta}$), and coefficients associated with remote sensing–based indices ($\boldsymbol{\zeta}$) for each pollen type.

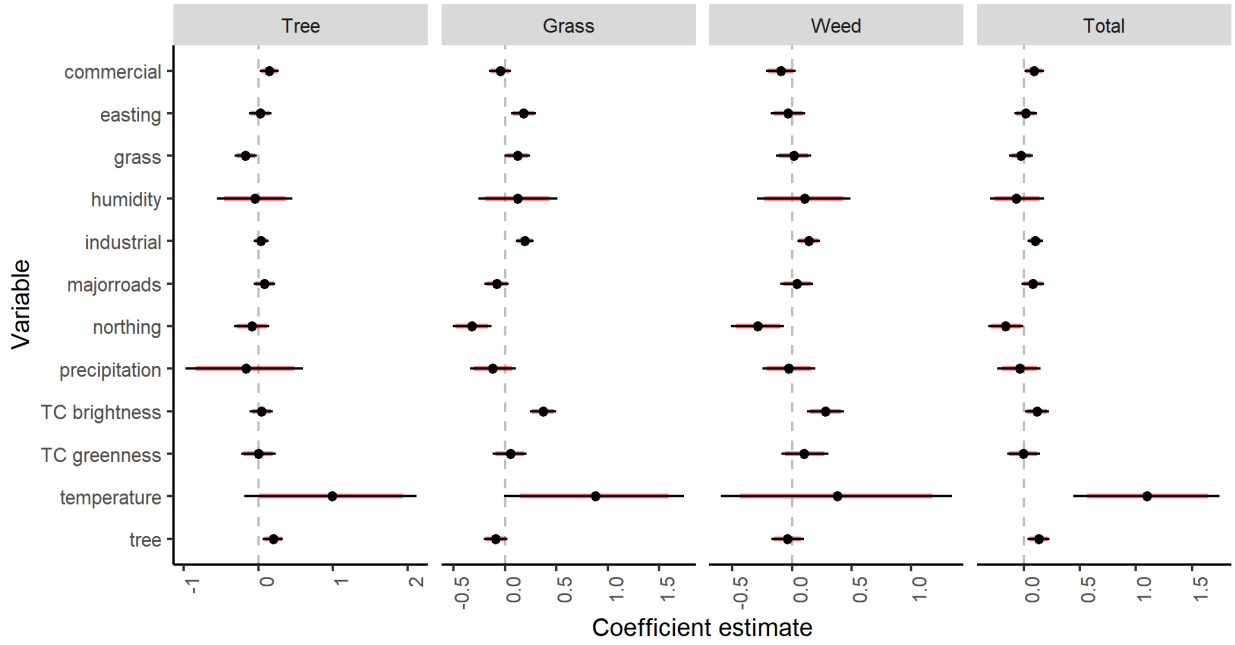


Figure 3 Point estimates, 95% (black line), and 90% (red line) credible intervals for the coefficients of nine land-use variables, three climatic variables, and two remote sensing–based indices for tree, grass, weed, and total pollen. The same buffer size was used for all pollen types.

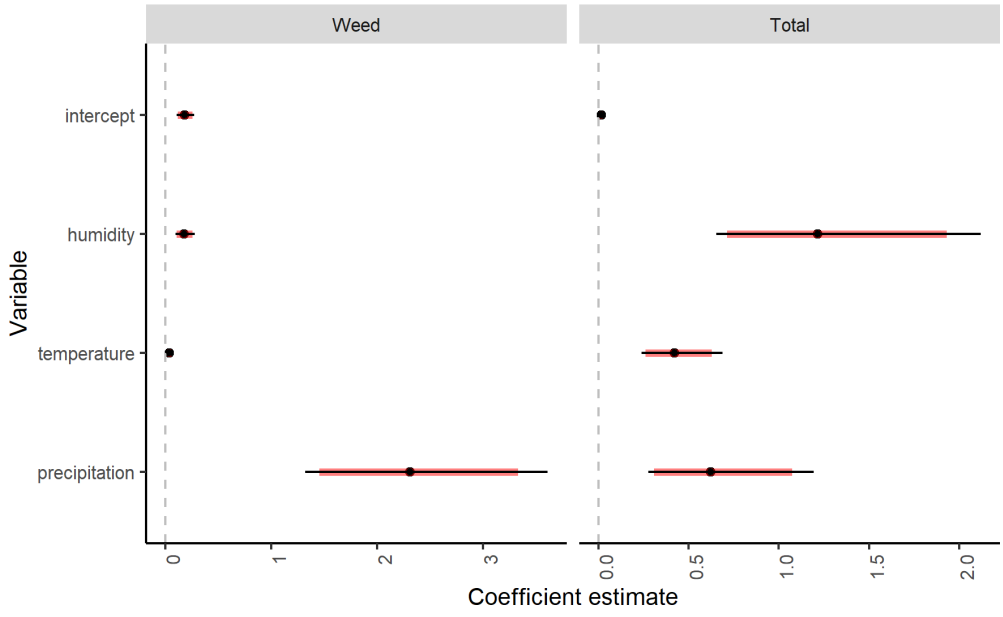


Figure 4 Point estimates (solid circles), together with the 95% (black line), and 90% (red line) posterior credible intervals of the odds of presence of weed and total pollen for the intercept, humidity, temperature and precipitation.

We found that tree pollen concentration was positively associated with commercial areas, and tree cover, and negatively associated with grass cover. We also found a positive association between the grass pollen concentration and the easting coordinate, grass cover, industrial areas, TC brightness, and temperature, and a negative association with the northing coordinate. Weed pollen concentration was positively associated with industrial areas and TC brightness, and negatively associated with northing. The total pollen concentration was positively associated with commercial areas, industrial areas, TC brightness, temperature, and tree cover, and negatively associated with the northing coordinate. For the total pollen the variable with the greatest coefficient was temperature.

Figure 4 shows the estimates (posterior means), 90% credible interval (red line), and 95% credible interval (black line) for the odds of the coefficients of these climatic factors ($\boldsymbol{\gamma}$) for the weed and total pollen.

We found that the increase in humidity and precipitation increases the odds of the probability of weed and total pollen concentration to be equal to zero and, in the case of total pollen, the odds of the probability was also positively associated with temperature. Figure 5 provides the posterior summaries of ρ_t and θ_t, depicting their respective evolution with time during the pollen season.

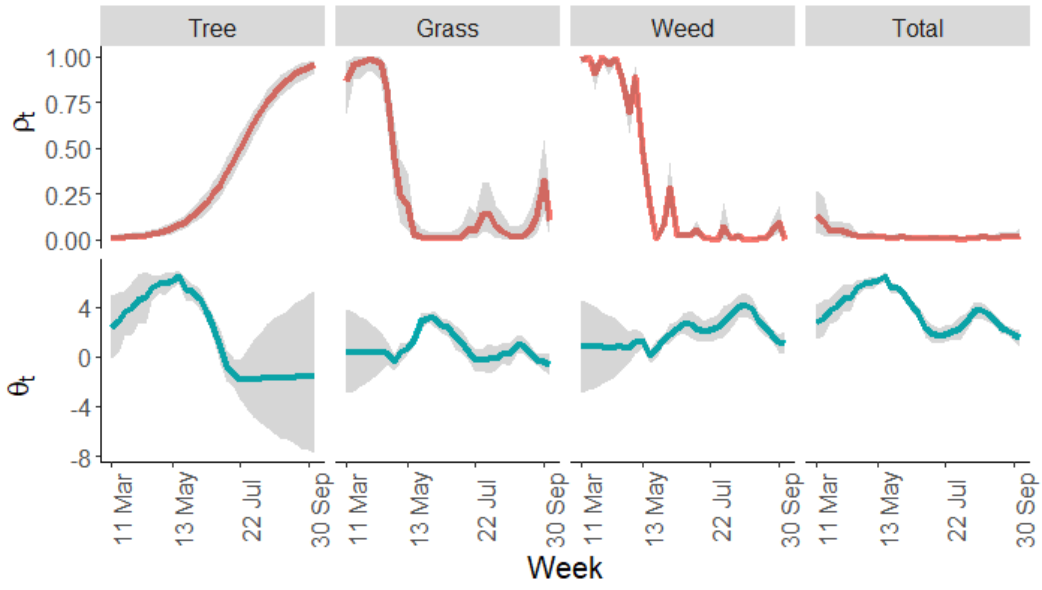


Figure 5 Posterior summaries (mean and limits of the 95% credible interval (shaded grey area)) for the time-varying probability (upper panel), and time-varying level (bottom panel) for the concentration of tree, grass, weed, and total pollen across 31 weeks for the 2018 pollen season in Toronto.

Additionally, we computed the predicted surfaces of pollen concentration for each pollen type on a 1km by 1km grid (Figure 6). The predictors for each grid cell were obtained using the values at the center of each grid cell. The predicted surfaces were computed for weeks with low (first column), medium (second column) and high (third column) pollen levels for each of the pollen types in the logarithmic scale. These weeks were arbitrarily chosen to show the pollen variation across space for weeks with different levels of pollen.

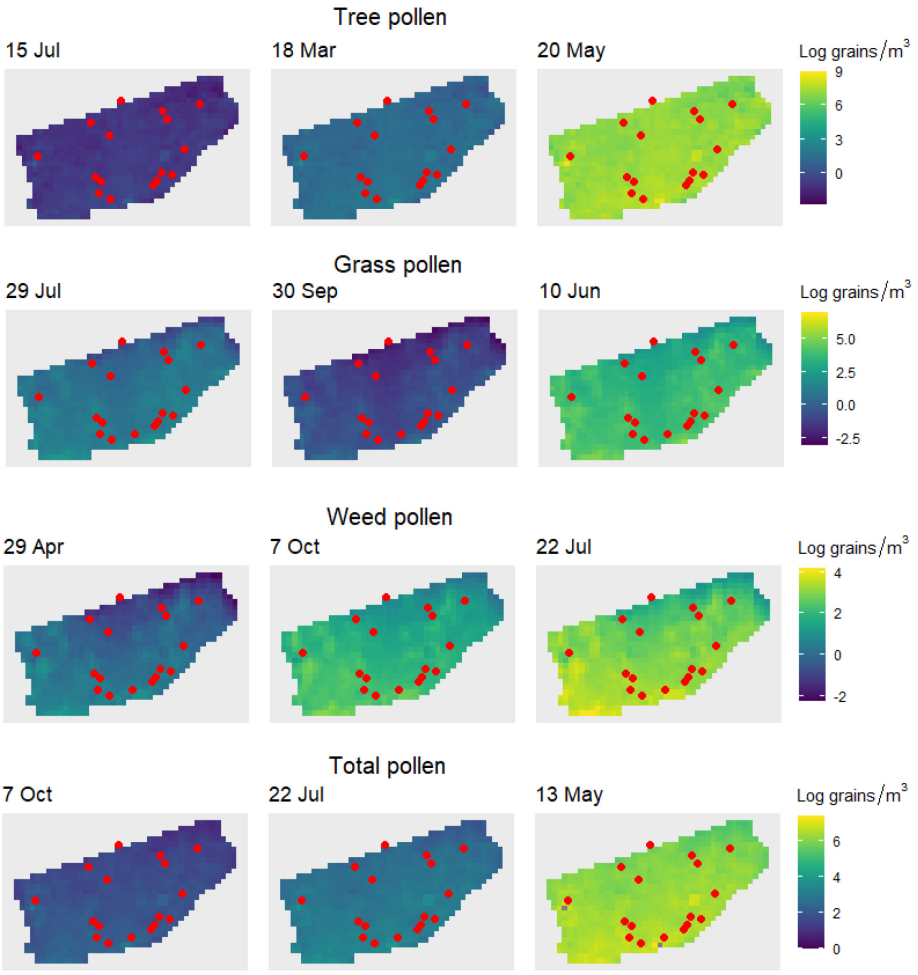


Figure 6 The first column shows a week for which the pollen concentration is relatively low (July 29 for grass pollen, July 15 for tree pollen, April 29 for weed pollen, and October 7 for the total pollen), the second column shows a week with moderate pollen concentration (September 30 for grass pollen, March 18 for tree pollen, October 7 for weed pollen, and July 22 for the total pollen), and the last column shows a week with higher pollen concentration for each pollen type (June 10 for grass pollen, May 20 for tree pollen, July 22 for weed pollen, and May 13 for the total pollen). Note that, for each row, the scales are comparable across the columns.

Although not shown here, we also analyzed the residuals at each site to ensure there was no pattern left after accounting for the covariates and temporal structure. We confirmed there was no temporal autocorrelation in the residuals at different time lags by plotting the autocorrelation function plots of the residuals at each site. Finally, to assess the quality of the fitted values and to

assess if there was any structure left in the residuals we plotted the observed measurements along with the estimated missing and fitted values for some sites and all types of pollen (Figure 6 in the Supplementary Material).

4.7 Discussion

To the best of our knowledge, this is the first study to analyze spatiotemporal variations in aeroallergen concentrations throughout the pollen season for different types of pollen within a city. This approach allowed us to describe the concentration variation across time and evaluate associations with various land-use variables for each pollen type.

Previous studies [35, 36] have used LUR to assess the spatial variation of one pollen type within a city for the whole pollen season. In contrast, we analyzed weekly concentrations and found that pollen type varies widely across the pollen season, with tree and grass pollen types showing the highest levels. By adding a time-varying mean component to the LUR and using a hurdle model, we were also able to capture changes in concentration throughout the pollen season.

The difference in the concentration of the different types of pollen across sites (Figure 2) supports the idea that one monitoring location is not representative of the pollen concentration across the city.

4.7.1 Pollen Data

Tree pollen: In the tree pollen model, we captured higher pollen levels from the start of the sampled period, with a peak the week of May 13, up to mid-July, after which the measurements at all sites were either zero or missing, therefore producing wider credible intervals by the end of July. This shift is also described by the change in the estimated probability from the hurdle

model. This probability was slightly greater than zero at the beginning of the season because, for the first few weeks, only some sites had measurements equal to zero. Then, after mid-May, this probability started to slowly increase up until late September when it reached values close to one. This corresponds to the observed data where the higher number of zeros are observed from late July to the end of the sampling period.

Grass pollen: The posterior summary of the time-varying mean level shows a low grass pollen concentration from March until mid-April. Then, the mean grass pollen concentration across the city increases for the remaining of the sampled period with two peaks in concentration around June and early September. The probability of having a measurement equal to zero was higher for the first few weeks and it smoothly dropped to lower values around mid-May with a small increase around late July. This corresponds with the observed data where the higher grass pollen levels at all sites were reached from late May to mid-July.

Weed pollen: The time-varying mean level for the weed pollen showed a pattern similar to that of the grass pollen. It captured the mean level at zero from the beginning of measurements until late April. After this initial period, the concentration level started to increase as some sites measured small levels of weed pollen and it dropped back again in the last weeks of the sampled period when the concentration levels were lower. The time-varying mean level also captured what was observed at the different locations where the peak in weed pollen concentration was reached from mid-August to early September. The highest probability of the measurements being equal to zero was observed during the first weeks of the sampling period. The fixed effects also showed how weed pollen concentration was positively associated with an increase in the proportion of industrial areas and an increase of TC brightness, and it is negatively associated with the northing coordinate.

Total pollen: The change in concentration levels for all pollen types was captured as a whole in the total pollen time-varying mean level. The first and higher peak around mid-May corresponds to the higher concentrations of tree pollen, while the second lower peak, around late August, corresponds to higher levels of weed and grass pollen. The probability for the total pollen was low for most of the season, except for mid-March due to the higher probabilities in grass pollen and weed pollen and in late September due to higher probabilities in the tree pollen. The covariates that we found to be associated with the total pollen are similar to those of the pollen types individually. For example, industrial areas were positively associated with grass and weed, while commercial areas were positively associated with tree pollen concentration. Aggregating the data to obtain the total pollen narrowed the credible interval for the temperature coefficient. We hypothesize this might be due to the lower number of observed zeros in the data.

In terms of land use, we found that industrial areas and TC brightness were positively associated with both grass and weed pollen. Previous studies have found that ragweed thrives in environments similar to the ones found in industrial areas [23, 77]. However, to the best of our knowledge, no explicit relationship with industrial areas has been reported before. Additionally, other studies have highlighted the importance of remote sensing-based indices in modeling and predicting pollen concentration in human-modified environments [35]. Furthermore, remote sensing variables have been used extensively in the literature for the identification of various land uses or vegetation types [35, 140, 141].

Tree cover and grass cover should be correlated with pollen production and therefore airborne pollen concentrations. Given that tree and grass cover categories are mutually exclusive, it is intuitive that there are negative associations between grass cover and airborne tree pollen and

vice versa. For the weed and total pollen, we also found that an increase in humidity and precipitation increased the probability of measurements being equal to zero, possibly showing how these climatic factors might influence the start and end of the pollen season for weed and total pollen. Previous studies have also reported a negative association of rainfall and humidity with a reduction in pollen release [79].

The predicted surfaces for each pollen type (Figure 6) show the spatial variability of pollen concentration across the city. The different spatial patterns are easier to visualize in the weeks with higher concentration. For example, both grass and weed pollen concentration were lower at the north and east part of Toronto, with higher weed pollen concentration in the old Toronto area. The distribution of tree pollen concentration was uniform across the city with three areas that stood out, the Woodbine Racetrack to the west, the Golden Mile district to the east, and the old Toronto area to the south (see Figure 7 in the Supplementary Material). Both the Woodbine Racetrack and the Golden Mile district are similar to each other in the sense that they are big open-air areas, which might facilitate the dispersion of the different types of pollen. Finally, the predicted total pollen surface reflects the patterns found in all three pollen types.

4.7.2 Statistical Model

Our proposed model accounted for the complex structure of the data, allowing for temporal correlation in the mean structure and using land-use variables to accommodate local information of the measurements. By analyzing the concentration levels at the weekly scale, we were also able to capture the weekly changes in pollen concentration for each type of aeroallergen.

The advantage of following the Bayesian paradigm becomes clearer when estimating the high number of missing values in the dataset, especially during the first few weeks. The wider credible intervals in the state vector θ_t are explained by the zero measurements. When a zero is observed, the value of θ_t solely relies on its prior distribution $\theta_t \sim N(\theta_{t-1}, W_1^2)$.

This same methodology has the potential to be applied to pollen seasons from other years and at different locations. However, the parameter estimates should not be extrapolated to other locations outside of the study area. However, the parameter estimates found in the present study can be used as prior information for future studies for the same region and during the same period, so that estimates of the parameters can be updated as new information becomes available.

A possible extension of this work would be to study the spatiotemporal variation of specific pollen species using a similar methodology. The multiple peaks in concentration in the different types of pollen may be due to an increase in the pollen production of specific tree, weed, and grass species. This framework can help identify if different species are affected by different local factors.

We believe that understanding the temporal and spatial patterns can contribute to future epidemiological studies on asthma and respiratory allergies, particularly in urban areas. With this type of analysis, it would be possible to link spatiotemporal variation in pollen concentration to the spatiotemporal variation in health outcomes data. For instance, one can link the exposure surfaces developed in this project to evaluate fine-scale spatial resolution and critical windows (e.g. during pregnancy, infancy, etc.) exposure to pollen on the incidence of asthma and other allergic diseases. Furthermore, understanding the effect of land use on pollen concentration across a city can better inform urban planners; for instance, urban greening that aims to increase the planting of trees in urban areas should account for the exposure to high concentrations of

aeroallergens [142]. Also, even if some of the findings in the present study are not generalizable beyond the study area, it is still possible to create predictive models of pollen concentration for Toronto, which will be useful for future health related studies and other applications. Although not shown here, we explored models that included a latent spatial structure to, eventually, accommodate spatial structures that were not accounted for by the land-use variables. As the results suggested there was no spatial information left in the data after considering the land-use variables, the results from those models were not included here.

4.8 Acknowledgements

We thank Liu Sun for her assistance with the GIS data.

The first author thanks her sponsors CONACYT, COMECYT and AMEXCID. Schmidt is grateful for financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grants RGPIN-2017-04999). We also wish to thank Aerobiology Research Laboratories.

4.9 Conflict of interest

We have no conflicts of interest to disclose.

4.10 Funding

This study was funded through Addressing Air Pollution Horizontal Initiative program of Health Canada.

References

16. D'Amato G, Bergmann KC, Cecchi L, Annesi-Maesano I, Sanduzzi A, Liccardi G, et al. Climate change and air pollution. Allergo Journal International. 2014;23(1):17-23.

18. Sierra-Heredia C, North M, Brook J, Daly C, Ellis AK, Henderson D, et al. Aeroallergens in Canada: Distribution, Public Health Impacts, and Opportunities for Prevention. Int J Environ Res Public Health. 2018;15(8).

23. Katz DSW, Carey TS. Heterogeneity in ragweed pollen exposure is determined by plant composition at small spatial scales. Sci Total Environ. 2014;485-486:435-40.

25. Katz DSW, Batterman SA. Urban-scale variation in pollen concentrations: a single station is insufficient to characterize daily exposure. Aerobiologia. 2020.

35. Hjort J, Hugg TT, Antikainen H, Rusanen J, Sofiev M, Kukkonen J, et al. Fine-Scale Exposure to Allergenic Pollen in the Urban Environment: Evaluation of Land Use Regression Approach. Environ Health Perspect. 2016;124(5):619-26.

36. Weinberger KR, Kinney PL, Robinson GS, Sheehan D, Kheirbek I, Matte TD, et al. Levels and determinants of tree pollen in New York City. Journal of Exposure Science & Environmental Epidemiology. 2018;28(2):119-24.

38. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Third ed. Boca Raton: CRC Press; 2014.

43. Gamerman D, Lopes HF. Markov chain Monte Carlo : stochastic simulation for Bayesian inference. 2nd ed. / ed. Boca Raton: Taylor & Francis; 2006.

50. Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986;33(3):22.

62. Weber RW. Floristic zones and aeroallergen diversity. Immunology and Allergy Clinics of North America. 2003;23(3):357-69.

68. Werchan B, Werchan M, Mucke HG, Gauger U, Simoleit A, Zuberbier T, et al. Spatial distribution of allergenic pollen through a large metropolitan area. Environ Monit Assess. 2017;189(4):169.

69. Werchan B, Werchan M, Mücke H-G, Bergmann K-C. Spatial distribution of pollen-induced symptoms within a large metropolitan area—Berlin, Germany. Aerobiologia. 2018;34(4):539-56.

75. Katz DSW, Dzul A, Kendel A, Batterman SA. Effect of intra-urban temperature variation on tree flowering phenology, airborne pollen, and measurement error in epidemiological studies of allergenic pollen. Sci Total Environ. 2019;653:1213-22.

77. Katz DSW, Batterman SA. Allergenic pollen production across a large city for common ragweed (Ambrosia artemisiifolia). Landsc Urban Plan. 2019;190.

79. Rojo J, Rapp A, Lara B, Fernandez-Gonzalez F, Perez-Badia R. Effect of land uses and wind direction on the contribution of local sources to airborne pollen. Sci Total Environ. 2015;538:672-82.

108. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing. 2014;24(6):997-1016.

110. Yang Y, Liu X, Zheng J, Tan Q, Feng M, Qu Y, et al. Characteristics of one-year observation of VOCs, NOx, and O3 at an urban site in Wuhan, China. J Environ Sci (China). 2019;79:297-310.

113. Senechal H, Visez N, Charpin D, Shahali Y, Peltre G, Biolley JP, et al. A Review of the Effects of Major Atmospheric Pollutants on Pollen Grains, Pollen Content, and Allergenicity. ScientificWorldJournal. 2015;2015:940243.

114. La Rosa M, Lionetti E, Reibaldi M, Russo A, Longo A, Leonardi S, et al. Allergic conjunctivitis: a comprehensive review of the literature. Italian Journal of Pediatrics. 2013;39(18).

115. Linneberg A, Henrik Nielsen N, Frølund L, Madsen F, Dirksen A, Jørgensen T. The link between allergic rhinitis and allergic asthma: A prospective population-based study. The Copenhagen Allergy Study. Allergy. 2002;57:4.

116. von Mutius E, Weiland SK, Fritzsch C, Duhme H, Keil U. Increasing prevalence of hay fever and atopy among children in Leipzig, East Germany. The Lancet. 1998;351(9106):862-6.

117. Sly M. Changing prevalence of allergic rhinitis and asthma. Ann Allergy Asthma Immunol 1999;82:19.

118. Report from the Canadian Chronic Disease Surveillance System: Asthma and Chronic Obstructive Pulmonary Disease (COPD) in Canada. Public Health Agency of Canada; 2018.

119. Dales RE, Cakmak S, Judek S, Dann T, Coates F, Brook JR, et al. Influence of outdoor aeroallergens on hospitalization for asthma in Canada. J Allergy Clin Immunol. 2004;113(2):303-6.

120. Cakmak S, Dales RE, Coates F. Does air pollution increase the effect of aeroallergens on hospitalization for asthma? J Allergy Clin Immunol. 2012;129(1):228-31.

121. Sun X, Waller A, Yeatts KB, Thie L. Pollen concentration and asthma exacerbations in Wake County, North Carolina, 2006-2012. Sci Total Environ. 2016;544:185-91.

122. Darrow LA, Hess J, Rogers CA, Tolbert PE, Klein M, Sarnat SE. Ambient pollen concentrations and emergency department visits for asthma and wheeze. J Allergy Clin Immunol. 2012;130(3):630-8 e4.

123. Osborne NJ, Alcock I, Wheeler BW, Hajat S, Sarran C, Clewlow Y, et al. Pollen exposure and hospitalization due to asthma exacerbations: daily time series in a European city. Int J Biometeorol. 2017;61(10):1837-48.

124. Erbas B, Jazayeri M, Lambert KA, Katelaris CH, Prendergast LA, Tham R, et al. Outdoor pollen is a trigger of child and adolescent asthma emergency department presentations: A systematic review and meta-analysis. Allergy. 2018;73(8):1632-41.

125. Ziska LH, Beggs PJ. Anthropogenic climate change and allergen exposure: The role of plant biology. J Allergy Clin Immunol. 2012;129(1):27-32.

126. D'Amato G, Holgate ST, Pawankar R, Ledford DK, Cecchi L, Al-Ahmad M, et al. Meteorological conditions, climate change, new emerging factors, and asthma and related allergic disorders. A statement of the World Allergy Organization. World Allergy Organ J. 2015;8(1):25.

127. Zhang Y, Bielory L, Mi Z, Cai T, Robock A, Georgopoulos P. Allergenic pollen season variations in the past two decades under changing climate in the United States. Glob Chang Biol. 2015;21(4):1581-9.

128. Ziska L, Knowlton K, Rogers C, Dalan D, Tierney N, Elder MA, et al. Recent warming by latitude associated with increased length of ragweed pollen season in central North America. Proc Natl Acad Sci U S A. 2011;108(10):4248-51.

129. Ziska LH, Gebhard DE, Frenz DA, Faulkner S, Singer BD, Straka JG. Cities as harbingers of climate change: common ragweed, urbanization, and public health. J Allergy Clin Immunol. 2003;111(2):290-5.

130. Schmidt CW. Pollen Overload: Seasonal Allergies in a Changing Climate. Environ Health Perspect. 2016;124(4):A70-5.

131. Skjøth CA, Ørby PV, Becker T, Geels C, Schlünssen V, Sigsgaard T, et al. Identifying urban sources as cause of elevated grass pollen concentrations using GIS and remote sensing. Biogeosciences. 2013;10(1):541-54.

132. Devadas R, Huete AR, Vicendese D, Erbas B, Beggs PJ, Medek D, et al. Dynamic ecological observations from satellites inform aerobiology of allergenic grass pollen. Sci Total Environ. 2018;633:441-51.

133. Wopfner N, Gadermaier G, Egger M, Asero R, Ebner C, Jahn-Schmid B, et al. The spectrum of allergens in ragweed and mugwort pollen. Int Arch Allergy Immunol. 2005;138(4):337-46.

134. Aerobiology Research Laboratories. GRIPST 2009 Rotation Impaction Sampler Manual

135. Reich BJ, Ghosh SK. Bayesian Statistical Methods. New York Chapman and Hall/CRC 2019.

136. Team RC. R: A Language and Environment for Statistical Computing. 3.6.3 ed. Vienna, Austria: R Foundation for Statistical Computing; 2020.

137. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of Statistical Software. 2017;76(1).

138. Stan Development Team. RStan: the R interface to Stan. R package version 2.17.3. 2018.

139. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016.

140. Liu Q, Huang C, Li H. Mapping plant communities within quasi-circular vegetation patches using tasseled cap brightness, greenness, and topsoil grain size index derived from GF-1 imagery. Earth Science Informatics. 2021;14(2):975-84.

141. Dymond CC, Mladenoff DJ, Radeloff VC. Phenological differences in Tasseled Cap indices improve deciduous forest classification. Remote Sensing of Environment. 2002;80(3):460-72.

142. Eisenman TS, Churkina G, Jariwala SP, Kumar P, Lovasi GS, Pataki DE, et al. Urban trees, air quality, and asthma: An interdisciplinary review. Landscape and Urban Planning. 2019;187:47-59.

5 Modelling temporally misaligned data across space: the case of total pollen concentration in Toronto

5.1 Preface to Manuscript 3

In this chapter, the daily concentration of total pollen is modelled. Similar to Chapter 4, land-use regression is extended by modelling the overall mean pollen concentration, in this case, on a daily scale. However, to be able to model the concentration on a daily scale and to include the data from all the sites in the monitoring network, it is necessary to account for the fact that the measurements were taken at different temporal scales across sites. Therefore, using the properties of the multivariate normal distribution, a temporal misalignment model is proposed.

This manuscript will be submitted to *Environmetrics.*

5.2 Authors contribution

Sara Zapata-Marin: Formal analysis, Methodology, Writing - Original Draft

Alexandra M. Schmidt: Formal analysis, Methodology, Supervision, Writing - Review & Editing,

Scott Weichenthal: Writing - Review & Editing

Eric Lavigne: Funding acquisition, Data acquisition, Writing - Review & Editing

Modelling temporally misaligned data across space: the case of total pollen concentration in Toronto

Sara Zapata-Marin[1], Alexandra M. Schmidt[2], Scott Weichenthal[2], Eric Lavigne[3]

[1]Quantitative Life Sciences, McGill University, Quebec, Canada

[2]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Quebec, Canada

[3]Air Health Science Division and Population Studies Division, Health Canada, Ontario, Canada

5.3 Abstract

Some spatio-temporal processes, particularly in the environmental sciences, face the problem of temporal misalignment in the data. Here, temporal misalignment refers to having measurements at different temporal scales across monitoring locations. For example, there might be daily observations for some sites and weekly observations for a different set of sites. Rather than aggregating the data to the coarser scale, it is possible to account for this difference in scale and take advantage of the fine-scale temporal observations.

A spatio-temporal model is proposed to account for temporal misalignment when one of the scales is the sum or average of the other by using the properties of the multivariate normal distribution. The inference is performed under the Bayesian framework, and uncertainty about the unknown quantities of interest is naturally accounted for. We fit our model to synthetic data for different classes of dynamic linear models with and without spatial structure. Additionally, we use this model to estimate the concentration of total pollen across Toronto, Canada. For some sites, the data was recorded daily whereas, for others, observations were recorded weekly. With the proposed model it is shown how the temporal aggregation of the pollen concentration measurements can impact the associations with the different covariates.

5.4 Introduction

Environmental processes often involve data collected at multiple locations and sometimes across temporal scales to describe complex processes. However, the temporal and spatial scales at which measurements are obtained may impact the conclusions that can be drawn from the data. These variations can occur when different dynamics prevail at different spatial and temporal scales.

An additional challenge arises when data are collected at multiple spatial scales, usually referred to as *spatial misalignment*, a topic that has been broadly studied in spatial statistics. For example, in Berrocal, Gelfand, & Holland (2010a) [143] and Berrocal, Gelfand, & Holland (2010b) [144] a spatio-temporal downscaler model is proposed to relate numerical models with observations from a monitoring network of air pollutants for both the univariate and the bivariate case; and in Lawson et al. (2012) [145] a model is proposed to account for the spatial misalignment between exposure and health data. For a more detailed review regarding spatial misalignment see chapter 7 of Banerjee, Carlin, and Gelfand (2003) [46].

Similarly, it can also be the case that measurements are collected at different temporal scales (e.g. hours and days, days and weeks, weeks and months) at different locations, usually due to the high costs and complex logistics of monitoring over fine temporal scales at multiple sites. Here, the term *temporal misalignment* refers to spatio-temporal data for which measurements across time are obtained at different scales across monitoring sites. More precisely, the focus is on the case where the coarse-scale data is obtained by aggregating over the fine-scale measurements.

Several authors have studied temporal aggregation in dynamic linear models (DLMs) and auto-regressive models. For example, Amemiya and Wu (1972) [146] focus on the aggregation of a p-th order autoregressive system to study the structure of the aggregated sequence; Schmidt and Gamerman (1997) [147] show how, when the fine-scale measurements follow a DLM, under certain constraints, the aggregated series follows the same DLM class as the model for the original scale and Ferreira, Higdon, Lee, and West (2006) [148] propose a multiscale model which links the information across temporal scales via stochastic links. For a more detailed review regarding multiscale time series modelling see chapter 11 of Ferreira and Lee (2007) [149].

In the present study, one of the goals is to recover the unobserved fine-scale measurements at sites where only the coarse-scale measurements are available, assuming that the relevant dynamics of the process rely on the process at the fine scale, in a similar note to Holan, Toth, Ferreira, and Karr (2010) [150] where a *Bayesian multiscale multiple imputation* method is proposed to impute missing observations in a data confidentiality study.

In this work, a model-based approach is proposed to handle temporally misaligned data which allows using the data from sites with different temporal scales when the coarser scale is the sum or average of the lower temporal scale. More specifically, the properties of the multivariate dynamic linear model are used to estimate the temporal trend at a finer time scale over the studied area while using the data from sites with both fine and coarse time scale measurements.

5.4.1 Motivating example

The proposed approach is motivated by data available on the concentration of total pollen across Toronto, Canada. The data were collected between March and October 2018 across $n = 18$ fixed monitoring sites. The total pollen is calculated as the sum of the tree, weed, and grass pollen concentrations at a given period. For 11 of these sites, measurements were taken on a daily basis, while for the remaining 7 sites measurements were taken every week (Figure 1).

The data was collected using rotation impaction samplers. For the daily measurements, the sampler would spin for a full minute every ten minutes over a 24 hour period. The weekly measurements, for a whole minute, samplers would do 20 spins three days a week and 21 spins four days a week, for a total of 144 minutes of sampling per week. The measurements provided are the average over the collected data. Pollen measurements were then converted into their volumetric equivalents expressing the aeroallergen concentration as pollen grains per cubic meter of air sampled.

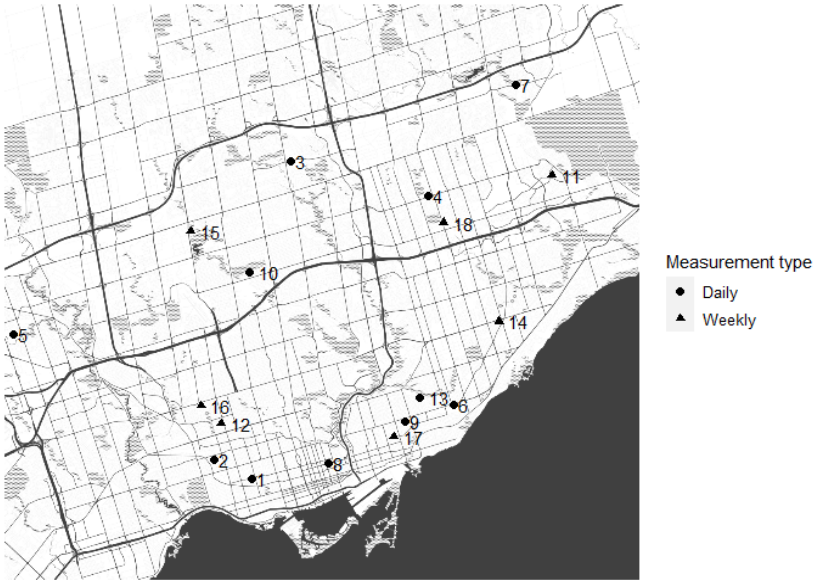


Figure 1 Location of n =18 monitoring stations across Toronto. Eleven sites (circles) captured the daily measurements of pollen, and the seven remaining ones (triangles) captured the weekly concentrations of pollen.

In this study, one of the goals is to learn how different land-use and environmental factors relate to the pollen concentration within a city at the finer temporal scale and to estimate the temporal trend of pollen levels across the city.

Here, the temporally-misaligned data is modelled through a combination of land-use regression and a multivariate dynamic linear model [49]. The aim is to take advantage of the properties of the Gaussian DLM to estimate the measurements at the finer temporal scale for all sites. The proposed model allows borrowing strength from the daily measurements to estimate the daily values at sites where only the weekly mean concentrations were available.

This paper is organized as follows. In section 5.5, the proposed model, as well as some of its properties, are described. The multivariate DLM aggregation approach is revisited in subsection 5.5.1. The inference procedure is described in subsection 5.5.2. Temporal predictions and spatial interpolations are also described through the respective posterior predictive distribution as shown in subsection 5.5.3. Section 5.6 shows an application of the proposed model for a series of simulation studies and the motivating example. Finally, section 5.7 discusses our findings and future directions for the present work.

5.5 Proposed Model

Let $\{Y_t(\mathbf{s}); \mathbf{s} \in D; t = 1,2,\cdots\}$ be a stochastic process in discrete time t, at location $\mathbf{s} \in D$ with $D \subset \mathbb{R}^d$, for $d = 1, 2$ or 3. Assume that $Y_t(\mathbf{s})$ follows a hierarchical model such that

$$Y_t(\boldsymbol{s}) = \mathbf{F}_t'(\boldsymbol{s})\boldsymbol{\theta}_t + \mathbf{X}(\boldsymbol{s})'\boldsymbol{\beta} + \nu_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}), \tag{1}$$

and for the parameters evolving smoothly with time the state equation is defined as

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \mathbf{W}_t), \tag{2}$$

where $\mathbf{F}_t(\boldsymbol{s})$ is a known q-dimensional vector that contains different structures including covariates that vary in space and time, but it could also be covariates that change only in time; $\mathbf{X}(\boldsymbol{s})$ is a p-dimensional vector containing covariates that vary only across space; $\epsilon_t(\boldsymbol{s})$ represents a temporal uncorrelated measurement error, and follows a normal distribution with zero mean and variance τ^2 (nugget effect); the evolution matrix $\mathbf{G}_t$ is a known $q \times q$ matrix; $\boldsymbol{\theta}_t$ is a q-dimensional vector of coefficients associated with $\mathbf{F}_t(\boldsymbol{s})$; and $\mathbf{W}_t$ is a q-dimensional covariance matrix, commonly fixed across time [47], that is, $\mathbf{W}_t = \mathbf{W}$. In particular, here, we suggest a diagonal matrix. Assume that observations are available across n spatial sites and let $\boldsymbol{\nu}_t = \big(\nu_t(\boldsymbol{s}_1), \cdots, \nu_t(\boldsymbol{s}_n)\big)'$ be a latent spatial component effect, which follows a Gaussian Process with mean zero and covariance matrix $\sigma^2\mathbf{R}$, where σ^2 is the partial sill and $\mathbf{R}$ is a correlation matrix whose elements are given by some valid correlation function, e.g. an exponential correlation function such that, $R_{ij} = \exp\big(-d_{ij}/\phi\big)$ where d_{ij} is the Euclidean distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$, and ϕ is the range parameter.

Assuming the $Y_t(\mathbf{s})$'s are conditionally independent across time then, $Y_t(\boldsymbol{s}) \mid \boldsymbol{\theta}_t, \boldsymbol{\beta}, \nu_t(\boldsymbol{s}), \tau \sim N(\mathbf{F}_t'(\boldsymbol{s})\boldsymbol{\theta}_t + \mathbf{X}(\boldsymbol{s})'\boldsymbol{\beta} + \nu_t(\boldsymbol{s}), \tau^2)$, and $\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{W} \sim N(\mathbf{G}_t\boldsymbol{\theta}_{t-1}, \mathbf{W})$. Furthermore, let $\mathbf{Y}_t = \big(Y_t(\boldsymbol{s}_1), \cdots, Y_t(\boldsymbol{s}_n)\big)'$, by marginalizing $\mathbf{Y}_t$ with respect to $\boldsymbol{\nu}_t$, then $\mathbf{Y}_t$ follows a multivariate normal distribution with mean $\mathbf{F}_t'\boldsymbol{\theta}_t + \mathbf{X}'\boldsymbol{\beta}$, and covariance matrix $\boldsymbol{\Sigma} = \tau^2\,\mathbf{I}_n + \sigma^2\boldsymbol{R}$, where $\mathbf{I}_n$ is the n-dimensional identity matrix.

Now, assume there is a collection of aggregated observations of this stochastic process such that $\mathbf{Z}_k = \big(Z_k(\boldsymbol{s}_1), \ldots, Z_k(\boldsymbol{s}_m)\big)'$ at a set of m sites in D each r units of time, that is, assume that

$$\mathbf{Z}_k = \frac{1}{r}\sum_{i=1}^{r} \mathbf{Y}_{rk+i} = \frac{1}{r}[\mathbf{Y}_{rk+1} + \cdots + \mathbf{Y}_{rk+r}] \quad k = 0,1,\cdots,K. \tag{3}$$

Since all the fine-scale measurements $\mathbf{Y}_1, \cdots, \mathbf{Y}_T$ are conditionally independent random variables given $(\boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_T)$ then, by using the properties of the multivariate normal distribution the aggregated measurements follow a multivariate normal distribution, such that, $\mathbf{Z}_k \mid$ $\boldsymbol{\theta}_{rk+1}, \ldots, \boldsymbol{\theta}_{rk+r}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim N\left(\frac{1}{r}\sum_{i=1}^{r} \boldsymbol{F}'_{rk+i}\ \boldsymbol{\theta}_{rk+i} + \mathbf{X}'\boldsymbol{\beta}, \frac{1}{r}\boldsymbol{\Sigma}\right)$.

Additionally, when $\mathbf{Y}_t$ is univariate, Schmidt and Gamerman (1997) [147] showed that under certain constraints, the aggregated process, in this case $\mathbf{Z}_k$, can be described by the same class of DLM of the original time series. The next section discusses these findings for a first-order multivariate DLM. This is to make the connection between the aggregation in the univariate case described in Schmidt and Gamerman (1997) [147] and the multivariate case clearer.

5.5.1 Temporal aggregation in Multivariate Dynamic Linear Models

Let $\mathbf{Y}_t^* = \mathbf{Y}_t - \mathbf{X}'\boldsymbol{\beta}$ and $\mathbf{Z}_k^* = \mathbf{Z}_k - \mathbf{X}'\boldsymbol{\beta}$. Then, $\mathbf{Y}_t^*$ follows a multivariate DLM defined by the quadruple $\{\mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}, \mathbf{W}\}_t$ which, in its recursive representation leads to,

$$\mathbf{Y}_t^* = \mathbf{F}_t{}'\boldsymbol{\theta}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

and with a system equation defined as in equation (2). For the aggregated measurements $\mathbf{Z}_k^*$, it might be possible to represent the aggregated process as a DLM.

5.5.1.1 Local level model

Let $\mathbf{Y}_{rk+i}^*$ follow a local level model such that the fine-scale measurements are modeled as noisy observations of the level θ_{rk+i} which in turn is modeled as a random walk. This model is

described by the quadruple $\{\mathbf{1}_n, 1, \boldsymbol{\Sigma}, W\}$, where $\mathbf{1}_n$ is an n-dimensional column vector where each element is equal to 1. Then, the observation equation in its recursive form is defined as,

$$\mathbf{Y}^*_{rk+i} = \mathbf{1}_n \theta_{rk+1} + \mathbf{1}_n \sum_{j=2}^{i} \omega_{rk+j} + \mathbf{v}_{rk+i}, \quad \mathbf{v}_{rk+i} \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

for $i = 1, \dots, r$. Summing over i, the coarse-scale measurements are,

$$\mathbf{Z}^*_k = \frac{1}{r}\sum_{i=1}^{r} \mathbf{Y}^*_{rk+i} = \mathbf{1}_n \theta_{rk+1} + \mathbf{1}_n \frac{1}{r}\sum_{i=2}^{r} (r - i + 1)\, \omega_{rk+i} + \frac{1}{r}\sum_{i=1}^{r} \mathbf{v}_{rk+i}.$$

Defining,

$$\delta_k = \theta_{rk+1} \qquad \text{and} \qquad \mathbf{v}^*_k = \mathbf{1}_n \frac{1}{r}\sum_{i=2}^{r} (r - i + 1)\, \omega_{rk+i} + \frac{1}{r}\sum_{i=1}^{r} \mathbf{v}_{rk+i},$$

then it follows that,

$$\mathbf{Z}^*_k = \mathbf{1}_n \delta_k + \mathbf{v}^*_k.$$

To prove that $\mathbf{Z}^*_k$ follows a multivariate DLM, it is necessary to write δ_k as a linear function of δ_{k-1} plus a disturbance term. Following Schmidt and Gamerman (1997) [147], by recursively substituting θ_{rk+1} in the system equation, δ_k can be rewritten as

$$\theta_{rk+1} = \theta_{rk-r+1} + \sum_{i=2}^{r} \omega_{rk-r+i},$$

which results in,

$$\begin{aligned}\delta_k &= \theta_{rk-r+1} + \sum_{i=2}^{r} \omega_{rk-r+i} \\ &= \delta_{k-1} + \omega_k^*,\end{aligned}$$

where $\omega_k^* = \sum_{i=2}^{r} \omega_{rk-r+i}$.

In the multivariate DLM structure, it is assumed that $\mathbf{Z}_{k-1}^*$ and δ_k are independent. However, given δ_{k-1}, $\mathbf{Z}_{k-1}^*$ and δ_k depend on ω_{rk-r+i} for $i = 2,3, \ldots, r$ and are not independent. Therefore, their covariance is given by,

$$\mathbf{C} = \mathrm{Cov}(\delta_k, \mathbf{Z}_{k-1}^* | \delta_{k-1}) = \mathrm{Cov}(\omega_k^*, \mathbf{v}_{k-1}^*) = W\mathbf{1}_n \sum_{i=2}^{r} (r - i + 1) = \mathbf{1}_n \frac{r(r-1)W}{2}.$$

Then, using the mutivariate normal theory,

$$\begin{pmatrix} \mathbf{Z}_{k-1}^* \\ \delta_k \end{pmatrix} \mid \delta_{k-1} \sim N\left[\begin{pmatrix} \mathbf{1}_{\mathrm{n}} \delta_{k-1} \\ \delta_{k-1} \end{pmatrix}; \begin{pmatrix} \boldsymbol{\Sigma}^* & \boldsymbol{C} \\ \boldsymbol{C}' & W^* \end{pmatrix} \right]$$

so that the equation for δ_k can be written as,

$$\delta_k = \delta_{k-1} + \mathbf{C}'(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{z}_k - \mathbf{1}_n \delta_{k-1}) + \omega_k^{**}$$

where,

$$\omega_k^{**} \sim N(0, W^* - \mathbf{C}'(\boldsymbol{\Sigma}^*)^{-1}\mathbf{C}),$$

where, according to normal theory, ω_k^{**} and $\mathbf{v}_k^*$ are independent across time and mutually independent. Then, the aggregated model is within a DLM structure but with a non-zero mean system's disturbance. Furthermore, when W is smaller than τ^2 and σ^2, and for small values of r, the structure approximates that of a DLM such that,

$$\begin{aligned} \mathbf{Z}_k^* &= \mathbf{1}_n\delta_k + \mathbf{v}_k^* \qquad & \mathbf{v}_k^* \sim N(\mathbf{0}, \mathbf{\Sigma}^*), \\ \delta_k &= \delta_{k-1} + \omega_k^* \qquad & \omega_k^* \sim N(0, W^*). \end{aligned}$$

The quadruple that defines the DLM for $\mathbf{Z}_k^*$ is given by $\{\mathbf{1}_n, 1, \mathbf{\Sigma}^*, W^*\}$, where $\mathbf{\Sigma}^* = V[\mathbf{v}_k^*]$ and $W^* = V[\omega_k^*]$. Assuming $\mathbf{v}_t$ and $\boldsymbol{\omega}_t$ are independent across time and mutually independent, it follows that,

$$\mathbf{\Sigma}^* = \mathbf{1}_n W\left(\frac{1}{3}r - \frac{1}{2} + \frac{1}{6r}\right)\mathbf{1}'_n + \frac{1}{r}\mathbf{\Sigma} \qquad \text{and} \qquad W^* = (r-1)W.$$

This is just to show that aggregation of multivariate DLMs follows similar structures as those discussed in Schmidt and Gamerman (1997) [147].

In the present study, the fact that $\mathbf{Z}_k^*$ does not always follow the structure of a multivariate DLM is not a concern since the samples of the posterior distribution are obtained using Markov chain Monte Carlo (MCMC) methods and the daily observations at the weekly measured sites are considered as missing observations. This is described in detail in the following section.

5.5.2 Inference procedure

Let $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_T)'$ be the collection of observations at each unit of time, and let $\mathbf{y}_t = \big(y_t(\mathrm{s}_1), \cdots, y_t(\mathrm{s}_n)\big)'$ be the collection of measurements at time t at a set of n sites $\boldsymbol{\mathcal{S}} = (\mathbf{s}_1, \cdots, \mathbf{s}_n)$ in D. Additionally, let $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_K)'$ be the collection of K mean measurements at each r units of time, and let $\mathbf{z_k} = \big(z_k(\mathbf{u}_1), \cdots, z_k(\mathbf{u}_m)\big)'$ be the collection of temporally aggregated data at a set of m sites $\boldsymbol{\mathcal{U}} = (\mathbf{u}_1, \cdots, \mathbf{u}_m)$ in D, such that $\boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{S}} = \varnothing$. From here on $\mathcal{S}$ will be referred as the fine-scale sites, and $\mathcal{U}$ as the aggregated sites.

Let $\mathbf{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \tau, \sigma, \phi, \mathbf{W})'$, be the parameter vector where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T)'$. Following equations (1) and (3), the likelihood function for $\mathbf{\Theta}$ can be written as,

$$l(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{z}) = \prod_{t=1}^{T} f_{\mathbf{y}}\left(\mathbf{y}_t | \boldsymbol{\Theta}\right) \prod_{k=0}^{K} g_{\mathbf{z}}\left(\mathbf{z}_k | \boldsymbol{\Theta}, \mathbf{y}\right), \tag{4}$$

where $f_{\mathbf{y}}(\mathbf{y}_t|\boldsymbol{\Theta})$ is the pdf of a multivariate normal distribution with mean $\mathbf{F'}_t\boldsymbol{\theta}_t + \mathbf{X'}\boldsymbol{\beta}$, and covariance matrix $\boldsymbol{\Sigma} = \tau^2\mathbf{I_n} + \sigma^2\mathbf{R}$; $g_{\mathbf{z}}(\mathbf{z_k}|\boldsymbol{\Theta}, \mathbf{y})$ is the pdf of a multivariate normal distribution with mean $\frac{1}{r}\sum_{i=1}^{r} \mathbf{F'}_{rk+i}\, \boldsymbol{\theta}_{rk+i} + \mathbf{X'}\boldsymbol{\beta}$, and covariance matrix $\frac{1}{r}\boldsymbol{\Sigma}$.

The inference procedure is performed under the Bayesian paradigm where model specification is complete after assigning a prior distribution $p(\boldsymbol{\Theta})$ for the parameter vector $\boldsymbol{\Theta}$. Here, prior independence of some of the components of $\boldsymbol{\Theta}$ is assumed, more specifically,

$$p(\boldsymbol{\Theta}) = \left[\prod_{t=1}^{T} p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \mathbf{W}\right)\right] p(\boldsymbol{\theta}_0) p(\boldsymbol{\beta})\, p(\sigma)\, p(\tau)\, p(\phi)\, p(\mathbf{W}). \tag{5}$$

For the standard deviations τ and σ, a half-Cauchy prior [151] is suggested; for the range parameter ϕ in the exponential correlation function, an exponential prior distribution is suggested with mean such that the practical range (distance at which the correlation is equal to 0.05) is reached at half of the maximum distance. For the parameter $\boldsymbol{\theta}_0$ assigning a non-informative prior is suggested, such as a normal distribution with mean $\mathbf{m}_0$ and variance $\mathbf{C}_0$. For the coefficients $\boldsymbol{\beta}$ associated with the land-use variables, a zero mean normal distribution with large variance is suggested. Finally, assigning a half-Cauchy or half-normal zero-mean distribution as a prior for the elements of $\mathbf{W}$ is suggested, since it is reasonable to allow for very small values of the elements of $\mathbf{W}$ as these allow the elements of the state vector to evolve smoothly with time.

5.5.3 Interpolation procedure

Given the nature of the problem, spatial interpolation can be thought of in different temporal scales. The following subsections discuss each of the cases that we envision for the problem at hand.

5.5.3.1 Fine-scale measurements at aggregated sites

One of the motivations for proposing this model is, similarly to Holan et al. 2010 [150] , to estimate the unobserved fine-scale measurements at sites where only the aggregated data is available. This is achieved by considering the fine-scale measurements at the coarse-scale sites as missing values such that there is borrow of strength from the sites with observed fine-scale measurements.

Let $\mathbf{Y}_t^{\boldsymbol{u}}$ be the set of unobserved fine-scale measurements at the set of sites $\boldsymbol{u}$ where only aggregated data were observed, let $\mathbf{Y}_t^{\boldsymbol{s}}$ be the observed fine-scale measurements at the set of sites $\boldsymbol{s}$, and let $\mathbf{Z}_k^{\boldsymbol{u}}$ be the observed aggregated measurements at the set of sites $\boldsymbol{u}$. In what follows we obtain the joint distribution of $\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}}, \mathbf{Y}_{rk+j}^{\boldsymbol{s}}, \mathbf{Z}_k^{\boldsymbol{u}}\right)'$. We start by computing the pairwise covariance matrix of the elements of $\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}}, \mathbf{Y}_{rk+j}^{\boldsymbol{s}}, \mathbf{Z}_k^{\boldsymbol{u}}\right)'$. Let $C(\cdot)$ be a valid covariance function and $d(\cdot,\cdot)$ be the Euclidean distance between two sets of sites, following the properties of the model, the elements of the covariance matrix for the pairwise covariances is given by,

$$\text{Cov}\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}}, \mathbf{Y}_{rk+j}^{\boldsymbol{s}}\right) = C\left(d(\boldsymbol{u}, \boldsymbol{s})\right) = \boldsymbol{\Psi}^{\boldsymbol{us}},$$

for the covariance between the unobserved fine-scale measurements at the coarse-scale sites and the observed fine-scale measurements,

$$\mathrm{Cov}\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}}, \mathbf{Z}_{k}^{\boldsymbol{u}}\right) = \mathrm{Cov}\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}}, \frac{1}{r}\sum_{i=1}^{r}\mathbf{Y}_{rk+i}^{\boldsymbol{u}}\right) = \frac{1}{r}\,\mathrm{Cov}\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}}, \mathbf{Y}_{rk+j}^{\boldsymbol{u}}\right) = \frac{1}{r}C\left(d(\boldsymbol{\mathcal{U}}, \boldsymbol{\mathcal{U}})\right) = \frac{1}{r}\boldsymbol{\Sigma}^{\boldsymbol{u}},$$

for the covariance between the unobserved fine-scale measurements and the coarse-scale measurements at the coarse-scale sites, and

$$\mathrm{Cov}\left(\mathbf{Y}_{rk+j}^{\boldsymbol{s}}, \mathbf{Z}_{k}^{\boldsymbol{u}}\right) = \mathrm{Cov}\left(\mathbf{Y}_{rk+j}^{\boldsymbol{s}}, \frac{1}{r}\sum_{i=1}^{r}\mathbf{Y}_{rk+i}^{\boldsymbol{u}}\right) = \frac{1}{r}\,\mathrm{Cov}\left(\mathbf{Y}_{rk+j}^{\boldsymbol{s}}, \mathbf{Y}_{rk+j}^{\boldsymbol{u}}\right) = \frac{1}{r}\boldsymbol{\Psi}^{\boldsymbol{su}},$$

for the covariance between the observed fine-scale measurements and the coarse-scale measurements.

Then, following equations (1) and (3), and the properties of the multivariate normal distribution, the joint distribution of $\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}}, \mathbf{Y}_{rk+j}^{\boldsymbol{s}}, \mathbf{Z}_{k}^{\boldsymbol{u}}\right)'$ is given by,

$$\begin{pmatrix} \mathbf{Y}_{rk+j}^{\boldsymbol{u}} \\ \mathbf{Y}_{rk+j}^{\boldsymbol{s}} \\ \mathbf{Z}_{k}^{\boldsymbol{u}} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_{rk+j}^{\boldsymbol{u}} \\ \boldsymbol{\mu}_{rk+j}^{\boldsymbol{s}} \\ \boldsymbol{\mu}_{k}^{\boldsymbol{u}} \end{pmatrix}; \begin{pmatrix} \boldsymbol{\Sigma}^{\boldsymbol{u}} & \boldsymbol{\Psi}^{\boldsymbol{us}} & \frac{1}{r}\boldsymbol{\Sigma}^{\boldsymbol{u}} \\ \boldsymbol{\Psi}^{\boldsymbol{su}} & \boldsymbol{\Sigma}^{\boldsymbol{s}} & \frac{1}{r}\boldsymbol{\Psi}^{\boldsymbol{su}} \\ \frac{1}{r}\boldsymbol{\Sigma}^{\boldsymbol{u}} & \frac{1}{r}\boldsymbol{\Psi}^{\boldsymbol{us}} & \frac{1}{r}\boldsymbol{\Sigma}^{\boldsymbol{u}} \end{pmatrix}\right],$$

where the matrix $\boldsymbol{\Sigma}^{\boldsymbol{s}}$ is the covariance matrix between the fine-scale sites, $\boldsymbol{\Sigma}^{\boldsymbol{u}}$ is the covariance matrix between the aggregated sites, and the matrix $\boldsymbol{\Psi}^{\boldsymbol{su}}$ is the covariance matrix between the fine-scale and the aggregated sites.

Following properties of the partition of the multivariate normal distribution, it follows that the distribution of $\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}} \mid \mathbf{Y}_{rk+j}^{\boldsymbol{s}}, \mathbf{Z}_{k}^{\boldsymbol{u}}, \boldsymbol{\Theta}\right)$, conditional on the model parameters and the observed fine-scale and coarse concentrations, is given by

$$\left(\mathbf{Y}_{rk+j}^{\boldsymbol{u}} \mid \mathbf{Y}_{rk+j}^{\boldsymbol{s}} = \mathbf{y}_{rk+j}^{\boldsymbol{s}}, \mathbf{Z}_{k}^{\boldsymbol{u}} = \mathbf{z}_{k}^{\boldsymbol{u}}, \boldsymbol{\Theta}\right) \sim N\left(\boldsymbol{\mu}_{rk+j}^{\boldsymbol{u}} + \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\left(\mathbf{a}_{rk+j}^{\boldsymbol{su}} - \boldsymbol{\mu}_{rk+j}^{\boldsymbol{su}}\right), \boldsymbol{\Sigma}^{\boldsymbol{u}} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\right),$$

where $\boldsymbol{\mu}_{rk+j}^{\mathcal{U}}$ is the mean vector for the fine-scale measurements at the aggregated sites; the vector $\mathbf{a}_{rk+j}^{\mathcal{SU}} = \left(\mathbf{y}_{rk+j}^{\mathcal{S}}, \mathbf{z}_{k}^{\mathcal{U}}\right)'$ contains data observed at the fine-scale and aggregated sites; and the vector $\boldsymbol{\mu}_{rk+j}^{\mathcal{SU}} = \left(\boldsymbol{\mu}_{rk+j}^{\mathcal{S}}, \boldsymbol{\mu}_{k}^{\mathcal{U}}\right)'$ contains the mean of the process at the fine-scale and aggregated sites. Finally, the covariance matrices are defined as,

$$\boldsymbol{\Omega}_{12} = [\boldsymbol{\Psi}^{\mathcal{US}} \quad \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{U}}],$$

$$\boldsymbol{\Omega}_{21} = \begin{bmatrix} \boldsymbol{\Psi}^{\mathcal{SU}} \\ \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{U}} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{22} = \begin{bmatrix} \boldsymbol{\Sigma}^{\mathcal{S}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{SU}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{US}} & \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{U}} \end{bmatrix}.$$

5.5.3.2 Fine-scale measurements at unobserved sites

Additionally, it may be of interest to interpolate the process at a set of completely unobserved sites, $\boldsymbol{\mathcal{H}} = (\mathbf{h}_1, \cdots, \mathbf{h}_l)$ such that $\boldsymbol{\mathcal{H}} \cap (\boldsymbol{\mathcal{S}} \cup \boldsymbol{\mathcal{U}}) = \varnothing$. Following a similar approach to the one described in the previous subsection, the joint distribution of the fine-scale measurements at unobserved sites $\mathbf{Y}_t^{\mathcal{H}}$, the observed fine-scale measurements $\mathbf{Y}_t^{\mathcal{S}}$, and the observed aggregated measurements $\boldsymbol{Z}_k^{\mathcal{U}}$ is given by,

$$\begin{pmatrix} \mathbf{Y}_{rk+j}^{\mathcal{H}} \\ \mathbf{Y}_{rk+j}^{\mathcal{S}} \\ \boldsymbol{Z}_k^{\mathcal{U}} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_{rk+j}^{\mathcal{H}} \\ \boldsymbol{\mu}_{rk+j}^{\mathcal{S}} \\ \boldsymbol{\mu}_k^{\mathcal{U}} \end{pmatrix}; \begin{pmatrix} \boldsymbol{\Sigma}^{\mathcal{H}} & \boldsymbol{\Psi}^{\mathcal{HS}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{HU}} \\ \boldsymbol{\Psi}^{\mathcal{SH}} & \boldsymbol{\Sigma}^{\mathcal{S}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{SU}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{UH}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{US}} & \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{U}} \end{pmatrix} \right],$$

where $\boldsymbol{\Sigma}^{\mathcal{H}}$ is the covariance matrix between the unobserved sites, $\boldsymbol{\Psi}^{\mathcal{SH}}$ is the covariance matrix between the fine-scale and the unobserved sites, and $\boldsymbol{\Psi}^{\mathcal{UH}}$ is the covariance matrix between the aggregated and the unobserved sites.

Conditional on the model parameters, and following properties of the partition of the multivariate normal distribution, it can be shown that the conditional distribution $\left(\mathbf{Y}_{rk+j}^{\mathcal{H}} \mid \mathbf{Y}_{rk+j}^{\mathcal{S}}, \mathbf{Z}_k^{\mathcal{U}}, \boldsymbol{\Theta}\right)$ is given by

$$\left(\mathbf{Y}_{rk+j}^{\mathcal{H}} \mid \mathbf{Y}_{rk+j}^{\mathcal{S}} = \mathbf{y}_{rk+j}^{\mathcal{S}}, \mathbf{Z}_k^{\mathcal{U}} = \mathbf{z}_k^{\mathcal{U}}, \boldsymbol{\Theta}\right) \sim N\left(\boldsymbol{\mu}_{rk+j}^{\mathcal{H}} + \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\left(\mathbf{a}_{rk+j}^{\mathcal{SU}} - \boldsymbol{\mu}_{rk+j}^{\mathcal{SU}}\right), \boldsymbol{\Sigma}^{\mathcal{H}} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\right),$$

where $\boldsymbol{\mu}_{rk+j}^{\mathcal{H}}$ is the mean vector of the fine-scale measurements at the unobserved sites.

The covariance matrices are defined as follows,

$$\boldsymbol{\Omega}_{12} = \left[\boldsymbol{\Psi}^{\mathcal{HS}} \quad \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{HU}}\right],$$

$$\boldsymbol{\Omega}_{21} = \begin{bmatrix} \boldsymbol{\Psi}^{\mathcal{SH}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{UH}} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{22} = \begin{bmatrix} \boldsymbol{\Sigma}^{\mathcal{S}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{SU}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{US}} & \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{U}} \end{bmatrix}.$$

5.5.3.3 Aggregated measurements at unobserved sites

Finally, let $\mathbf{Z}_k^{\mathcal{H}}$ be a vector containing the aggregated measurements at a set of unobserved sites $\boldsymbol{\mathcal{H}}$, and let $\mathbf{Z}_k^{\mathcal{S}}$ be the aggregated measurements at the set of sites $\boldsymbol{\mathcal{S}}$. The joint distribution for the aggregated measurements is given by,

$$\begin{pmatrix} \mathbf{Z}_k^{\mathcal{H}} \\ \mathbf{Z}_k^{\mathcal{S}} \\ \mathbf{Z}_k^{\mathcal{U}} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_k^{\mathcal{H}} \\ \boldsymbol{\mu}_k^{\mathcal{S}} \\ \boldsymbol{\mu}_k^{\mathcal{U}} \end{pmatrix}; \begin{pmatrix} \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{H}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{HS}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{HU}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{SH}} & \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{S}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{SU}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{UH}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{US}} & \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{U}} \end{pmatrix} \right],$$

where all the matrices are defined as above. Conditional on the model parameters the conditional distribution, $\left(\mathbf{Z}_k^{\mathcal{H}} \mid \mathbf{Z}_k^{\mathcal{S}}, \mathbf{Z}_k^{\mathcal{U}}, \boldsymbol{\Theta}\right)$ is defined as,

$$\left(\mathbf{Z}_k^{\mathcal{H}} \mid \mathbf{Z}_k^{\mathcal{S}} = \mathbf{z}_k^{\mathcal{S}}, \mathbf{Z}_k^{\mathcal{U}} = \mathbf{z}_k^{\mathcal{U}}, \boldsymbol{\Theta}\right) \sim N\left(\boldsymbol{\mu}_k^{\mathcal{H}} + \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\left(\mathbf{a}_k^{\mathcal{SU}} - \boldsymbol{\mu}_k^{\mathcal{SU}}\right), \boldsymbol{\Sigma}^{\mathcal{H}} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\right),$$

where $\boldsymbol{\mu}_k^{\mathcal{H}}$ is the mean vector of the aggregated measurements at the unobserved sites; the vector $\mathbf{a}_k^{\mathcal{SU}} = \left(\mathbf{z}_k^{\mathcal{S}}, \mathbf{z}_k^{\mathcal{U}}\right)$ contains aggregated observed data; the vector $\boldsymbol{\mu}_k^{\mathcal{SU}}$ contains the mean of the process at the aggregated sites such that, $\boldsymbol{\mu}_k^{\mathcal{SU}} = \left(\boldsymbol{\mu}_k^{\mathcal{S}}, \boldsymbol{\mu}_k^{\mathcal{U}}\right)$; and the covariance matrices are defined as follows,

$$\boldsymbol{\Omega}_{12} = \left[\frac{1}{r}\boldsymbol{\Psi}^{\mathcal{HS}} \quad \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{HU}}\right],$$

$$\boldsymbol{\Omega}_{21} = \begin{bmatrix} \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{SH}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{UH}} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{22} = \begin{bmatrix} \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{S}} & \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{SU}} \\ \frac{1}{r}\boldsymbol{\Psi}^{\mathcal{US}} & \frac{1}{r}\boldsymbol{\Sigma}^{\mathcal{U}} \end{bmatrix}.$$

Following Bayes' theorem, the resultant posterior distribution obtained by combining equations (4) and (5) does not have a closed form.

Markov Chain Monte Carlo (MCMC) methods are used to obtain samples from the resultant posterior distribution. In particular, the NIMBLE system v.0.11.1 [44] is used in R version 4.1.0 [152] to obtain samples from the resultant posterior distribution.

Furthermore, if there are missing observations for some sites at any instant in time, these are considered as parameters to be estimated and, similarly to the steps described above it is possible to use the properties of the multivariate normal distribution to obtain samples from their posterior full conditionals.

5.6 Data analysis

This section starts by analyzing a set of synthetic data generated from the model to check if we are able to recover the true values of the parameters. In particular, the goal is to check if we are able to estimate well daily measurements at sites that only provide weekly observations to the model. The results for the simulation studies are analyzed in section 5.6.1. In section 5.6.2 the analysis of the total pollen concentration in Toronto described in section 5.4 is shown. The main goals for this analysis are two-fold: a) to predict the unobserved fine-scale measurements at sites where only the aggregated measurements were available by borrowing strength from the fine-scale observed measurements, and b) investigate for possible associations with both the spatial and temporal covariates by considering all available observations.

5.6.1 Simulation studies

For each simulation study, the parameters in equations (1), and (2) are defined as follows,

M1 $\boldsymbol{\theta}_t = \left(\theta_{1,t}, \boldsymbol{\alpha}\right)'$, $\mathbf{F}_t(\mathbf{s}) = (1, \mathbf{U}_t)'$, $\mathbf{G}_t = 1$, and $\mathbf{W} = W_{11}$. The state vector comprises a time-varying level model.

M2 $\boldsymbol{\theta}_t = (\theta_{1,t}, \theta_{2,t}, \boldsymbol{\alpha})'$, $\mathbf{F}_t(\mathbf{s}) = (1,0,\mathbf{U}_t)'$, $\mathbf{G}_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, and $\mathbf{W} = \text{diag}(W_{11}, W_{22})$. The state vector comprises a time-varying trend model.

M3 $\boldsymbol{\theta}_t = (\theta_{1,t}, \theta_{2,t}, \theta_{3,t}, \theta_{4,t})'$, $\mathbf{F}_t(\mathbf{s}) = (1, \mathbf{U}_t)'$, $\mathbf{G}_t = \mathbf{I}_4$, and $\mathbf{W} = \text{diag}(W_{11}, W_{22}, W_{33}, W_{44})$. The state vector includes a time-varying level and time-varying coefficients of three predictors: $\mathbf{U}_t = (\text{ wind-speed, precipitation and humidity})_t'$.

The vector $\mathbf{X}(\mathbf{s})$ in equation (1) contains spatial variables such as easting, northing, all roads, buildings, commercial, industrial, and open space areas at location $\mathbf{s}$. The coefficients associated with the temporal covariates are fixed at $\boldsymbol{\alpha}$ = (-0.2, -0.5, 0.3) and the coefficients associated with the spatial variables are fixed at $\boldsymbol{\beta}$ = (0.2, -0.2, 0.25, 0.15, 0.3, -0.4, -0.1). Finally, Table 1 shows the true values for the spatial structure and the variance of the system equation for each model.

Table 1 True values for the spatial structure parameters and the diagonal elements of the covariance matrix W for the simulation studies.

Parameter	**M1**	**M2**	**M3**
σ	1.5	0.9	1.5
τ	1.1	0.5	1
ϕ	3	3	3
diag($\mathbf{W}$)	0.04	(0.2, 0.1)	(0.1, 0.2, 0.3, 0.15)

For all simulations the fine-scale measurements are generated following equation (1) for 210 days then, the aggregated measurements $\mathbf{Z}_k(\mathbf{s})$ are calculated as in equation (3), by averaging over all seven days in each week for a total of 30 weeks at each site. Finally, the daily

measurements from seven sites (sites 12 to 18 in Figure 1) are removed to be considered as unobserved fine-scale measurements and for which predictions will be obtained. In summary, the simulated data sets consist of eleven sites with daily measurements and seven sites with weekly measurements only, emulating the motivating example.

Under the Bayesian paradigm model specification is complete after assigning a prior distribution to the parameter vector. For the inference procedure, the prior distributions are assigned as follows: a zero-mean normal distribution for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with standard deviation of 10, an exponential prior for ϕ with mean equal to the practical range, a half-Cauchy prior distribution for τ and σ with location 0 and scale 1, and similarly for the components of $\mathbf{W}$.

For each set of simulations, 20 samples are generated and for each sample two parallel chains were run for 30,000 iterations with a burn in of 12,000, and storing every 5th iteration. To check convergence of the chains, diagnostic tools from the coda [153], and bayesplot [154] packages are used.

Figure 2 shows the posterior summaries of the parameters estimates for: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ, τ, $\mathbf{W}$, and the components of the state vector $\boldsymbol{\theta}$ at times t = 5, 10, 50, 150, 200, for Model M1. For the majority of the samples, the actual true values fall within the 95% credible interval of the posterior. Similar results are obtained for models M2 and M3 (see Supplementary Material section S.1).

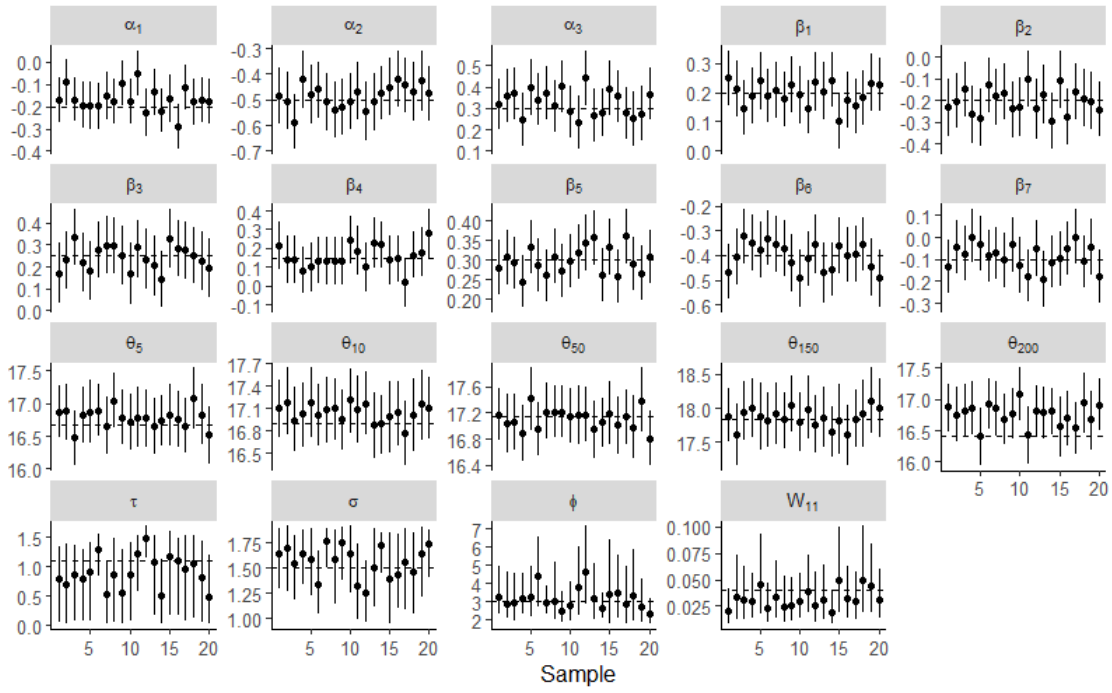


Figure 2 Posterior summary for 20 simulations for a local level model (M1). Mean (solid circles), and 95% posterior credible intervals (segments) of **α**, **β**, θ_5, θ_{10}, θ_{50}, θ_{150}, θ_{200}, τ, σ, ϕ, and W. Each dashed line within a panel represents the respective true value

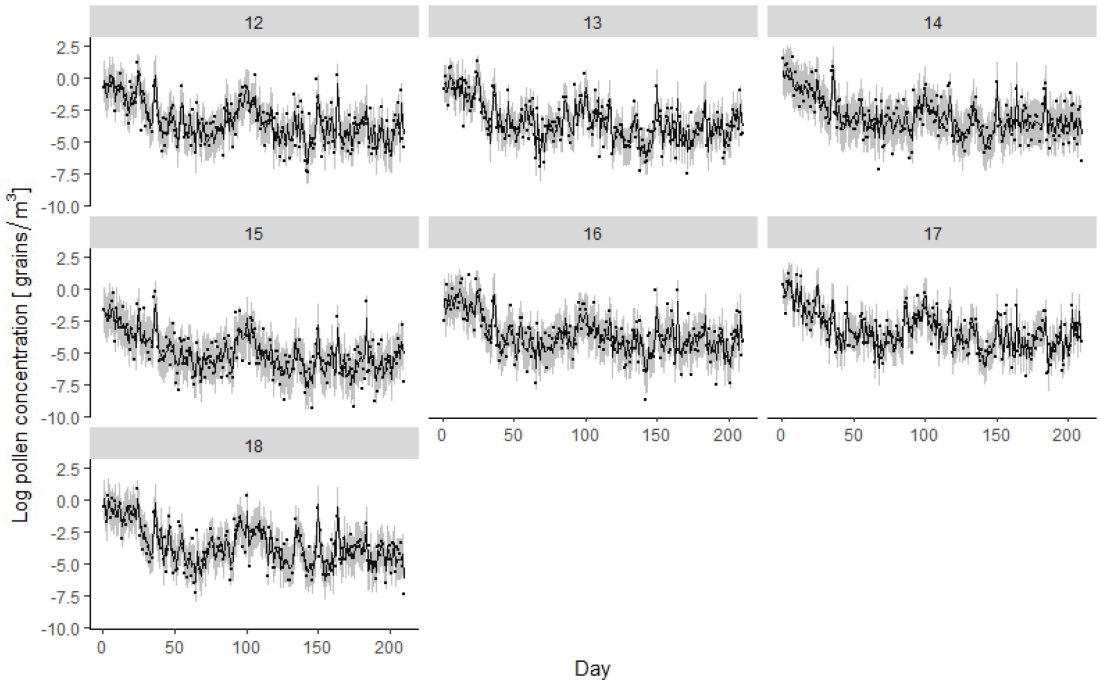


Figure 3 Posterior mean (solid line) and posterior 95% CI (shaded area) for the unobserved daily values at the weekly sites compared to the true values (solid circles) of a simulation of the model M1.

Additionally, the ability of the model to recover the removed daily measurements is also checked. Following M1, a sample with different parameters is generated to help visualization. Figure 3 shows the posterior mean of the unobserved fine-scale observations at the sites where only coarse scale measurements are provided to the model, along with the 95% credible intervals across days. This figure shows how the majority of the removed daily observations are included within the 95% credible interval. Similar results are obtained for models M2 and M3 (see Supplementary Material section S.1).

Additionally, analogous results are obtained when $\nu_t(\mathbf{s}) = 0$ where all the spatial structure has already been accounted for by the land-use covariates (results are not shown here).

5.6.2 The case of total pollen in Toronto

In this section, the total pollen concentration in the city of Toronto is analyzed. As described in section 5.4.1, there are available observations for 210 days at 11 sites and 30 weeks at 7 sites. Additionally, there are available land-use variables at various buffer sizes for each location and weather-related variables for the 210 days which change only across time. Land-use variables and buffer sizes are chosen so that they are not highly correlated with each other.

Models M1, M2, and M3 as described in section 5.6.1 are fitted to the log concentration of total pollen with and without a latent spatial structure. There are few observations equal to zero for some of the daily measurements so, in order to compute the logarithm of the observed values, a random positive jitter is added. The results show that the more complex structure from models M2 and M3 is not appropriate in this case therefore, the following analysis contains the results for the local level model (M1) with and without a latent spatial structure.

Model comparison for the two variations of M1 is performed using the Watanabe-Akaike Information Criterion (WAIC) which accounts for both the goodness of fit and the complexity of the model [108]. In total, there are 191 missing observations in the daily data and 23 in the weekly data.

For both models using the suggested priors in section 5.5.2 a WAIC of 6954.49 is obtained for the model without spatial structure and 22619.92 for the model with spatial structure. Therefore, since lower values of WAIC are preferred, the best model between the two is the one without a spatio-temporal latent component $\nu_t(\mathbf{s})$. The results shown next are based on model M1 with $\nu_t(\mathbf{s}) = 0$.

In order to compare how much additional information is obtained by using the model that accommodates for the temporal misalignment, model M1 is fitted considering two different temporal scales: only the daily data at the 11 observed sites, and the weekly data observed at the 18 sites. Clearly, the analysis that considers only the sites with daily measurements has fewer observations than when the weekly observations are considered.

Figure 4 provides the posterior summaries (mean and limits of 95% credible intervals) for the coefficients associated with each of the environmental variables and land-use factors for all three models that assume different temporal scales. Following the proposed misalignment model, the results show a negative relationship between the pollen concentration and the open space, industrial, all roads, northing, humidity, and precipitation variables. Additionally, there is a positive association of pollen concentration and commercial areas, buildings, easting and wind speed. The results show wider credible intervals for the coefficients when the model only considers the data aggregated at the weekly scale, especially for coefficients associated with weather-related variables. For most cases, the direction of the relationship between the land use variables and the pollen concentration is consistent between the daily model and the misalignment model. However, the credible intervals for the land-use variables for the misalignment model are also narrower than for the daily model. This is not surprising as the model under the misaligned data borrows strength from the weekly observations whereas the daily model only considers 11 observed sites *vis-à-vis* the 18 available in total.

Finally, for the misalignment model, the posterior summary (mean and limits of 95% posterior credible intervals) for the observation standard deviation and the system standard deviation were $\tau = 1.45\ (1.11, 1.18)$ and $W = 0.45\ (0.37, 0.53)$, respectively.

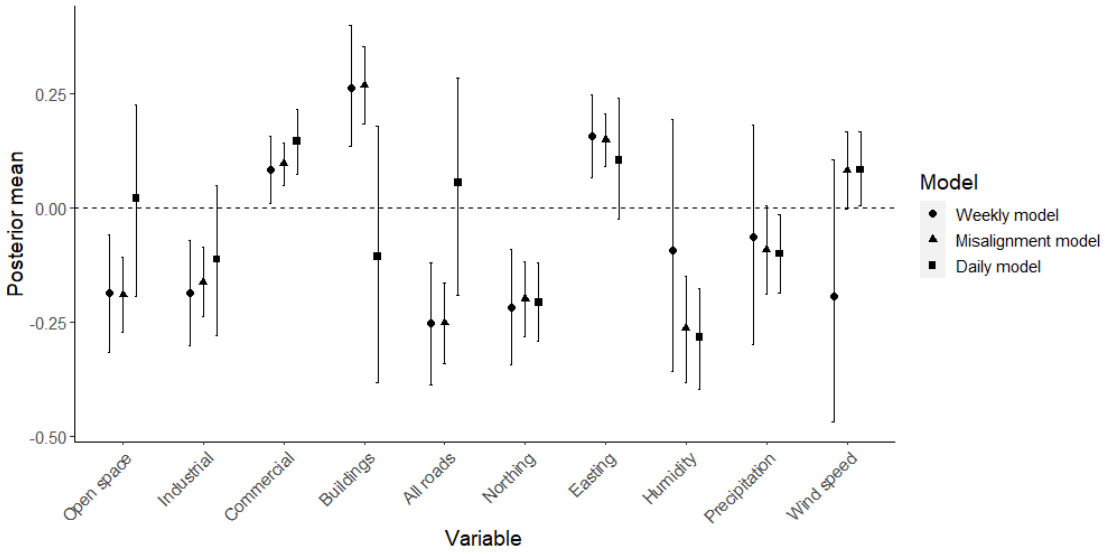


Figure 4 Posterior summaries for the environmental and land-use variables in a model with no latent spatial component for the total pollen data in Toronto. Posterior mean for the weekly data (solid circle), temporal misalignment (solid triangle), and daily data (solid square).

Figure 5 shows the predicted surfaces in a 1km by 1km grid for four different days and weeks using the proposed temporal misalignment model (upper panel), the daily data model (middle panel), and the weekly data model (lower panel).The surfaces at different stages show how the spatial variability obtained through the temporal misalignment model is somehow a combination of the daily model and the weekly model (see Supplementary Material subsection S.2).

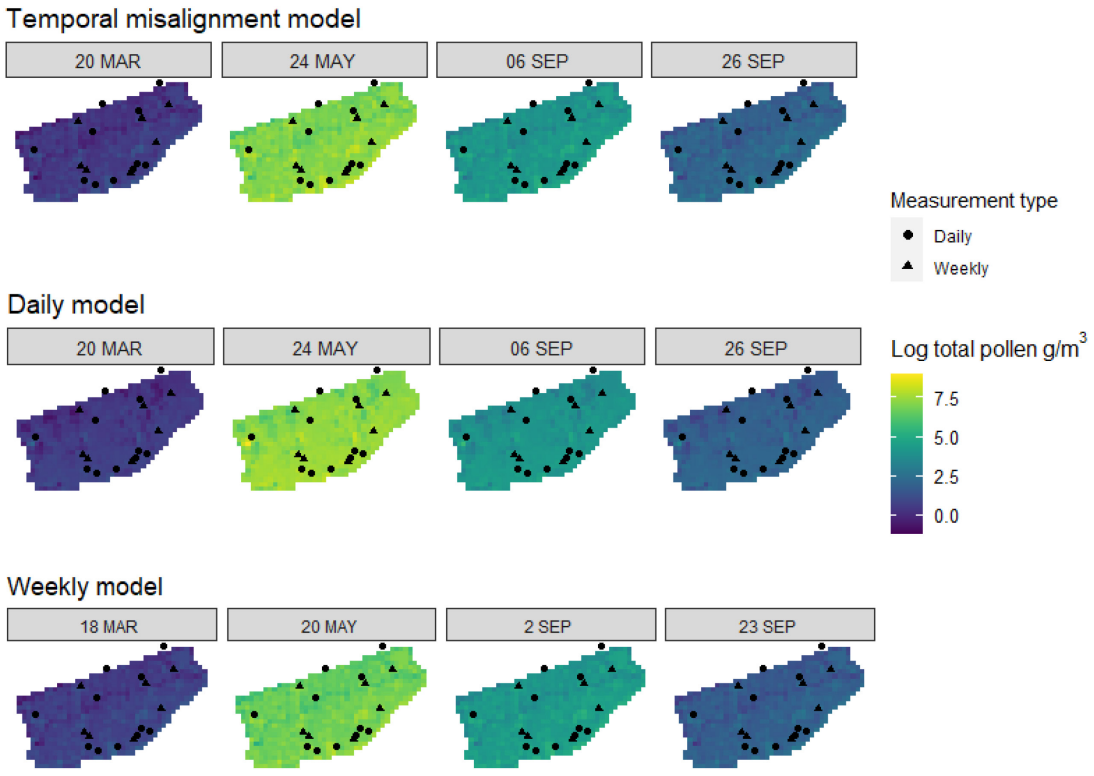


Figure 5 Mean of the posterior predictive distribution for the temporal misalignment model (top panel), a model using only fine-scale (mid panel), and a model using only aggregated data (bottom panel) for the city of Toronto.

Additionally, Figure 6 shows the posterior mean and 95% credible interval of the state vector $\boldsymbol{\theta}$ in the daily scale. This parameter captures the overall temporal trend of total pollen across the city of Toronto. Two major increases can be observed in the total pollen concentration, the first one around mid May and a smaller one around late August.

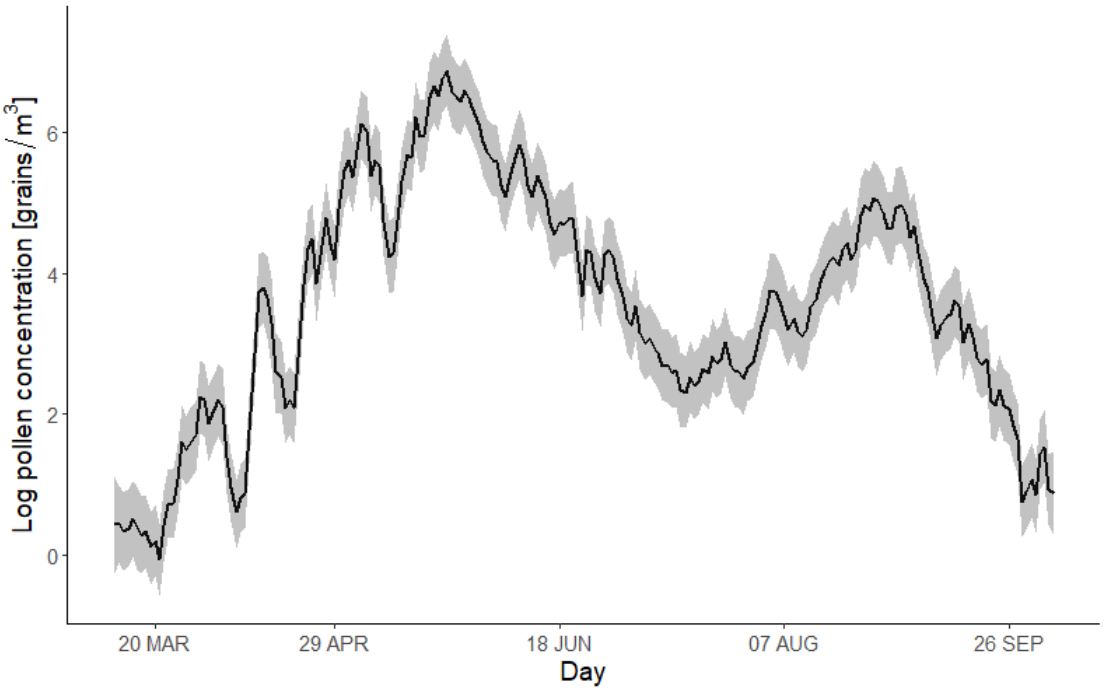


Figure 6 Posterior mean (solid line) and 95% posterior credible interval (grey shaded area) for the daily state vector θ of the time-varying mean level of total pollen in Toronto from day 1 to day 210.

Figure 7 shows the posterior mean and 95% CI for the estimated unobserved daily values at the 7 sites with weekly data only. Furthermore, after fitting the proposed temporal misalignment model, the aggregated posterior mean of the estimated daily measurements at the weekly sites is compared with the *observed weekly measurements* at these sites. Figure 8 shows how the fine-scale estimates, to some extent, recover the coarser scale measurements when aggregated.

The results for the fitted and missing values at the 11 sites with fine-scale measurements can be found in the Supplementary Material section S.2 along with the posterior mean and standard deviation for the predicted surface.

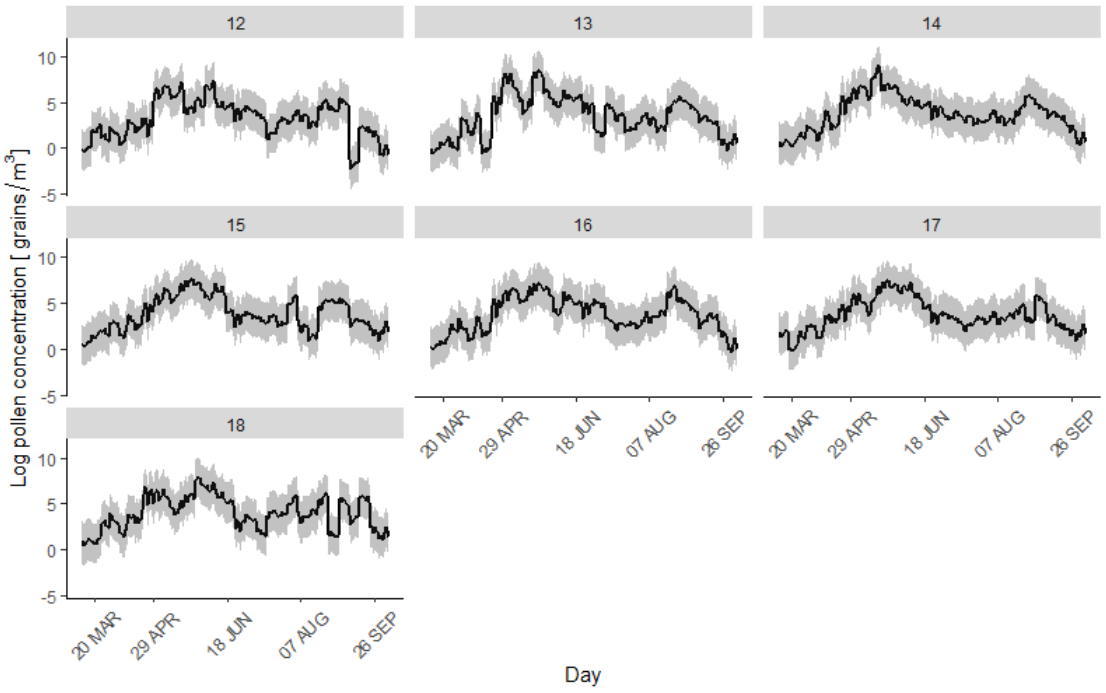


Figure 7 Estimated mean (solid black line), and 95% posterior credible intervals (shaded area) for the daily log of the total pollen concentration in grains/m^3 for 210 days at the seven weekly sites.

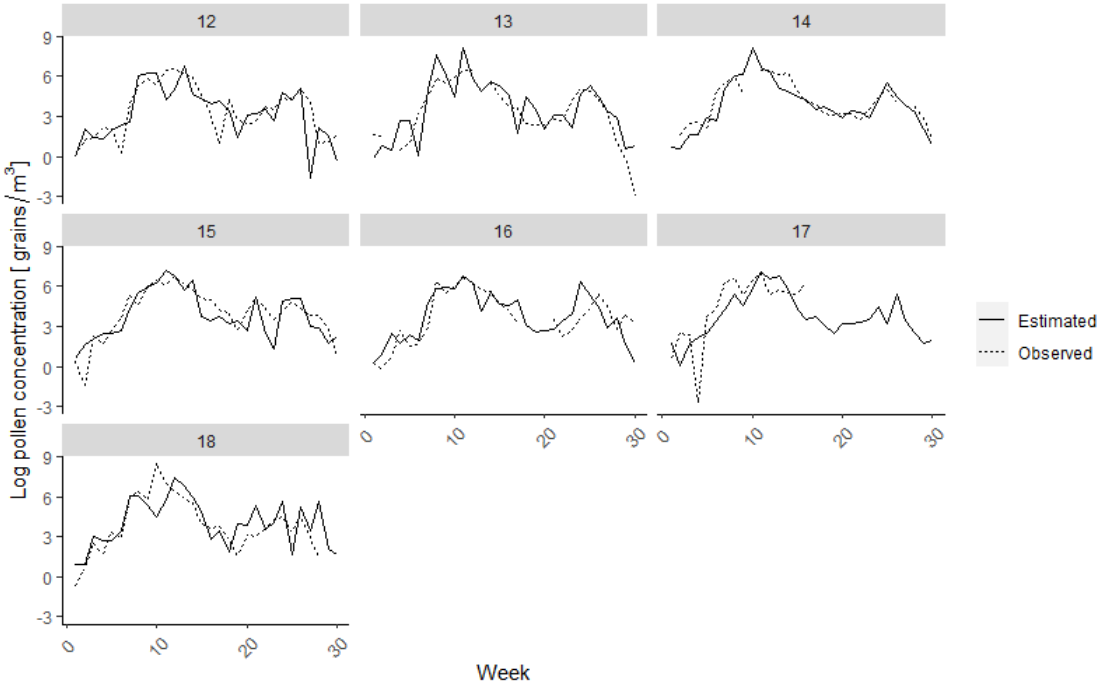


Figure 8 Aggregated estimated daily values (solid line) and observed weekly measurements (dashed line) at the weekly sites.

5.7 Discussion

Although spatial misalignment is a well-studied area in the spatial statistics literature, little attention has been paid to the temporal misalignment data as defined in the introduction. This study proposes a model to account for the temporal misalignment in spatial data using the properties of the multivariate dynamic linear model to estimate the unobserved process at sites where only the aggregated data is observed. The proposed model can be applied when the coarser temporal scale of the measurements is the sum or average of the lower temporal scale.

The misalignment model is fitted to three different types of DLMs, generating 20 samples for each type. The results in section 5.6.1 show that, for the majority of the samples the true values are recovered, suggesting that one is able to estimate the process at the finer level.

In section 5.6.2 the total pollen concentration in Toronto in the log scale is analyzed. Using the proposed framework, the daily trend of pollen concentration across the city is estimated (Figure 6), as well as the unobserved daily measurements at the weekly sites. The two increases in total pollen concentration correspond to an increase in tree pollen and grass pollen concentration in the first peak and weed pollen concentration in the second peak, as a previous analysis of the data has shown [155] .

Furthermore, a comparison is made to understand what additional information is obtained by using the data from all 18 sites while accounting for the temporal misalignment versus the same model when only using the observed daily data at the 11 sites or using the aggregated data on the weekly scale at all 18 sites. The results in Figure 4 show how accounting for a larger number of sites while correcting for the temporal misalignment across times results in narrower credible intervals, especially for the climatic variables. Furthermore, when only considering the aggregated data, the effect of the land-use variables is, in some cases, the opposite to the ones obtained in both the misalignment and the daily data models, that is the case of all roads, buildings, and open spaces land-use, however zero is within the 95% credible interval for all these cases. This shows how the temporal scale may impact the conclusions drawn from the data.

Following the misalignment model, there is a negative association of pollen concentration with humidity and precipitation in accordance with previous studies which show that, under certain conditions, rainfall can cleanse the air of pollen and spores [156] . Additionally, other studies also found a negative association with humidity due to the mechanisms of pollen release in plants [157] . The relationship with wind speed has proven to be more complex to explain due to the efficiency of the samplers under different climatic conditions [158] .

The framework proposed in this study shows a viable way in which it is possible to coherently combine information across temporal scales. In the future, this might allow lowering the costs of monitoring campaigns while still being able to draw meaningful conclusions at finer temporal scales. However, it is worth noting that the majority of the sites in this study were measured at the finer temporal scales. In the future, it would be helpful to analyze the proportion between fine-scale measurements and aggregated measurements at which the proposed framework would still give valuable results.

5.8 Acknowledgements

We thank Aerobiology Research Laboratories for collecting the data and Calcul Quebec for the computational resources. Zapata-Marin thank CONACYT, COMECYT, AMEXCID, and FRQNT for the financial support during her PhD studies. Schmidt acknowledges financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grant RGPIN-2017-04999).

References

44. de Valpine PaP, C. and Turek, D. and Michaud, N. and Anderson-Bergman, C. and Obermeyer, F. and Wehrhahn Cortes, C. and Rodriguez, A. and Temple Lang, D. and Paganin, S. NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. 0.11.1 ed2021.

46. Banerjee S, Gelfand AE, Carlin BP. Hierarchical modeling and analysis for spatial data. Boca Raton, Florida Chapman & Hall/CRC Press; 2015.

47. Petris G, Petrone S, Campagnoli P. Dynamic linear models. Dynamic Linear Models with R: Springer; 2009. p. 31-84.

49. West M, Harrison J. Bayesian forecasting and dynamic models. New York: Springer; 1997.

108. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing. 2014;24(6):997-1016.

143. Berrocal VJ, Gelfand AE, Holland DM. A spatio-temporal downscaler for output from numerical models. Journal of agricultural, biological, and environmental statistics. 2010;15(2):176-97.

144. Berrocal VJ, Gelfand AE, Holland DM. A bivariate space-time downscaler under space and time misalignment. The annals of applied statistics. 2010;4(4):1942.

145. Lawson AB, Choi J, Cai B, Hossain M, Kirby RS, Liu J. Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data. Journal of agricultural, biological, and environmental statistics. 2012;17(3):417-41.

146. Amemiya T, Wu RY. The effect of aggregation on prediction in the autoregressive model. Journal of the American Statistical Association. 1972;67(339):628-32.

147. Schmidt AM, Gamerman D. Temporal aggregation in dynamic linear models. Journal of Forecasting. 1997;16(5):293-310.

148. Ferreira MA, Higdon DM, Lee HK, West M. Multi-scale and hidden resolution time series models. Bayesian Analysis. 2006;1(4):947-67.

149. Ferreira MAR, Lee HK. Multiscale modeling: a Bayesian perspective: Springer; 2007.

150. Holan SH, Toth D, Ferreira MA, Karr AF. Bayesian multiscale multiple imputation with implications for data confidentiality. Journal of the American Statistical Association. 2010;105(490):564-77.

151. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian analysis. 2006;1(3):515-34.

152. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

153. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006;6(1):4.

154. Gabry J, Mahr T. bayesplot: Plotting for Bayesian Models. 2021.

155. Zapata-Marin S, Schmidt AM, Weichenthal S, Katz DS, Takaro T, Brook J, et al. Within city spatiotemporal variation of pollen concentration in the city of Toronto, Canada. Environmental research. 2022;206:112566.

156. McDonald JE. Collection and washout of airborne pollens and spores by raindrops. Science. 1962;135(3502):435-7.

157. Pehkonen E, Rantio-Lehtimäki A. Variations in airborne pollen antigenic particles caused by meteorologic factors. Allergy. 1994;49(6):472-7.

158. Weber RW. Meteorologic variables in aerobiology. Immunology and Allergy Clinics. 2003;23(3):411-22.

6 Conclusions

6.1 Summary

The work presented here contributes to the study of the spatial and spatio-temporal distribution of air pollution and aeroallergens in urban settings. The three manuscripts presented here (Chapters 3, 4, 5) show different extensions of classical land-use regression methods that explore spatial and spatio-temporal patterns in the data using Bayesian hierarchical models.

Chapter 3 presents the analysis of the spatial distribution of five VOCs in Montreal, Quebec, Canada. The proposed models in this work account for the variability due to the different monitoring campaigns, land-use variables, and any additional spatial variability.

The proposed models combine land-use regression with geostatistical methods that capture the potential spatial structure left after accounting for the land-use covariates. Additionally, different parameters are added to capture the variation across monitoring campaigns as part of the model. Adding these components on top of the land-use variables enables the researcher to learn more details about the spatial distribution of VOCs within a city. For example, it is shown how the spatial structure changes across monitoring campaigns and how not one unique model fits all types of VOCs. Furthermore, it is possible to identify areas in Montreal with a high concentration of VOCs.

Chapters 4 and 5 analyze the pollen distribution in Toronto, Ontario, Canada due to its strong association with adverse health effects. The data analyzed in these two projects presented some statistical challenges due to the difference in sampling frequency across campaigns and the high number of measurements equal to zero. Two different approaches are proposed to study the spatio-temporal patterns in the data.

Chapter 4 analyzes the weekly spatial distribution of different pollen types. A hurdle model is used to account for the high number of measurements equal to zero. This allows us to show the probability of observing any of the pollen types each week. Additionally, a time-varying mean component is used to capture the overall concentration of pollen across the city.

Using this model, it is possible to show the time periods with a higher concentration of each pollen type. Furthermore, some associations between land-use variables and pollen concentration are also reported.

In Chapter 5, the daily distribution of total pollen is analyzed. The proposed model explicitly accounts for the fact that the measurements were collected at different time scales across monitoring sites. With this model, it is possible to recover the fine-scale measurements at sites where only coarse-scale measurements are available. Additionally, it is shown how the associations with the different parameters depend on the temporal scale being considered.

Obtaining exposure estimates for aeroallergens and air pollutants faces many obstacles. Even though the classical land-use regression methods can provide good estimates, hierarchical Bayesian methods can easily accommodate more complex structures when analyzing the distribution of air pollutants and aeroallergens across space and time. Overall, this work shows how by considering a spatial or spatio-temporal structure, it is possible to learn about the spatial patterns and the temporal dynamics of aeroallergens and air pollutants dispersion.

6.2 Limitations and future work

In Chapter 3, it is important to consider that these models only account for outdoor exposure. However, people might be exposed to these air pollutants in other places (on the road or indoors

at home and work), so when these results are used as exposure estimates, it is worth considering the measurement error due to these other unmeasured sources.

In Chapters 4 and 5, the characterization of aeroallergens exposure might be improved with a denser monitoring network with a higher number of monitors, but we recognize the difficulties present when monitoring this kind of aeroallergens.

A natural extension of the work developed in Chapters 4 and 5 is to propose a model to account for the high number of zeros in the data and the temporal misalignment.

For the work developed in Chapter 5, it would also be possible to explore a multi-scale model in which the different temporal scales have different time-varying components but are coherently related by a link equation across temporal scales [148].

All the predicted surfaces obtained in this thesis will be used in future health studies.

References

148. Ferreira MA, Higdon DM, Lee HK, West M. Multi-scale and hidden resolution time series models. Bayesian Analysis. 2006;1(4):947-67.

7 Appendices

7.1 Appendix A: Supplementary Material for Chapter 3

Supplementary Material: Spatial modelling of ambient concentrations of volatile organic compounds in Montreal, Canada

Sara Zapata-Marin, Alexandra M. Schmidt, Dan Crouse, Vikki Ho, France Labrèche, Eric Lavigne, Marie-Élise Parent, Mark S. Goldberg

1. Detection limits

Table SM-1 Detection limits (from the August 2006 survey) in ng/µL

	Detection Limit[1] (ng/µL)
Hexane	0.024
n-Decane	0.024
Ethylbenzene	0.011
Benzene	0.014
1,2,4-Trimethylbenzene	0.010

[1] Detection limits for the study were based on the U.S. Federal Register CFR 40 method.

2. Predictors description

Table SM-2 Description and sources of land-use predictors

Variable	Description	Source
Buildings	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Commercial area	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Government and institutional	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Resource and industrial area	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Open area	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Parks	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Population		Joined dissemination area boundary file 2011 and population data from 2016
Residential	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Roads length	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)
Average NOx emissions		VISSIM + MOVES
Total NOx emissions		VISSIM + MOVES
Average daily traffic volume		MOVES
Total daily traffic volume		MOVES
Waterbody	Proportion of buildings land use within a defined buffer size	DMTI 2013 Spatial Inc. (Markham Ontario)

3. Estimated coefficients

This section contains the posterior means and 95% credible intervals for the coefficients and parameters of the spatial structure of the chosen model for each VOC. Recall that our modelling strategy means that these results should not be used to identify important predictor variables.

Table SM-3 Buffer sizes, point estimates and 95% credible intervals, in brackets, at each campaign for the coefficients of easting, northing and land-use variables for benzene under Model 1.

Benzene		
Variable	Buffer radius (m)	Posterior Mean and 95% CI
Intercept December 2005	-	0.18 (-0.64, 1.00)
Intercept April 2006	-	0.02 (-1.16, 1.18)
Intercept August 2006	-	-0.92 (-2.19, 0.36)
Easting	-	0.15 (-0.03, 0.33)
Northing	-	0.03 (-0.14, 0.19)
Average NO_x	500	0.10 (0.04, 0.16)
	1000	-0.06 (-0.13, 0.00)
Buildings land use	50	0.02 (-0.02, 0.07)
	200	-0.04 (-0.10, 0.01)
	1000	0.10 (0.02, 0.19)
Open area land use	100	0.03 (-0.01, 0.06)
Population	50	0.05 (0.00, 0.1)
	500	-0.01 (-0.09, 0.07)
	1000	0.02 (-0.08, 0.12)
Residential land use	50	-0.01 (-0.07, 0.05)
	200	0.01 (-0.06, 0.07)
Resource and Industrial	1000	0.02 (-0.02, 0.06)

Table SM-4 Point estimates and 95% credible intervals, in brackets, for τ^2 (nugget effect), σ^2 (spatial variance), and ϕ (practical range) for benzene under Model 1

Benzene			
Variable	Posterior Mean and 95% CI		
	December 2005	April 2006	August 2006
τ^2		0.05 (0.04, 0.06)	
σ^2	0.31 (0.14, 0.58)	0.26 (0.14, 0.45)	0.48 (0.17, 0.97)
ϕ	100.53 (47.64, 134.40)	107.19 (59.16, 134.85)	77.13 (21.60, 132.54)

Table SM-5 Buffer sizes, point estimates and 95% credible intervals, in brackets, for the coefficients of easting, northing and land-use variables for n-decane under Model 1

n-decane		
Variable	Buffer radius (m)	Posterior Mean and 95% CI
Intercept December 2005	-	0.70 (0.44, 0.99)
Intercept April 2006	-	-0.12 (-0.74, 0.47)
Intercept August 2006	-	-0.87 (-1.17, -0.61)
Easting	-	0.14 (0.03, 0.25)
Northing	-	-0.02 (-0.12, 0.06)
Average NO_x	50	0.03 (-0.03, 0.09)
	200	0.07 (-0.02, 0.18)
	500	-0.07 (-0.20, 0.04)
	1000	0.08 (-0.02, 0.18)
Buildings land use	50	0.09 (0.01, 0.17)
	100	0.01 (-0.1, 0.13)
	200	0.06 (-0.07, 0.19)
	500	-0.11 (-0.26, 0.04)
Commercial land use	1000	0.03 (-0.02, 0.08)
Government and Institutional land	200	0.01 (-0.05, 0.06)
use	500	-0.01 (-0.06, 0.05)
Open area land use	100	0.00 (-0.05, 0.06)
Population	500	-0.11 (-0.20, -0.01)
	1000	0.12 (-0.01, 0.25)
Residential land use	50	0.03 (-0.03, 0.08)
	1000	0.00 (-0.08, 0.09)
Resource and Industrial	500	0.01 (-0.07, 0.10)
	1000	0.00 (-0.11, 0.11)
Roads land use	50	-0.01 (-0.08, 0.05)
	100	-0.01 (-0.09, 0.08)
	200	0.00 (-0.07, 0.08)
Waterbody	1000	0.00 (-0.07, 0.07)

Table SM-6 Point estimates and 95% credible intervals, in brackets, for τ^2 (nugget effect), σ^2 (spatial variance), and ϕ (practical range) for n-decane under Model 1

n-decane			
Variable	Posterior Mean and 95% CI		
	December 2005	April 2006	August 2006
τ^2	0.06 (0.04, 0.08)		
σ^2	0.14 (0.08, 0.30)	0.17 (0.09, 0.31)	0.17 (0.12, 0.25)
ϕ	8.82 (1.32, 38.22)	57.9 (25.02, 104.19)	1.53 (0.06, 4.32)

Table SM-7 Buffer sizes, point estimates and 95% credible intervals, in brackets, at each campaign for the coefficients of easting, northing and land-use variables for ethylbenzene under Model 4.

Ethylbenzene				
Variable	Buffer radius (m)	Posterior Mean and 95% CI		
		December 2005	April 2006	August 2006
Intercept	-	1.19 (1.13, 1.24)	0.97 (0.91, 1.02)	0.60 (0.54, 0.65)
Easting	-	0.13 (0.04, 0.21)	-0.06 (-0.15, 0.03)	0.13 (0.04, 0.21)
Northing	-	0.11 (0.03, 0.18)	0.01 (-0.07, 0.08)	-0.08 (-0.16, 0.00)
Average NO_x	500	0.18 (0.06, 0.3)	0.05 (-0.07, 0.18)	0.06 (-0.06, 0.17)
	1000	-0.06 (-0.21, 0.07)	0.05 (-0.10, 0.19)	-0.09 (-0.22, 0.03)
Buildings land use	50	0.11 (0.02, 0.19)	0.02 (-0.06, 0.10)	0.07 (-0.02, 0.15)
	500	0.09 (-0.11, 0.29)	0.00 (-0.20, 0.19)	0.05 (-0.14, 0.24)
	200	0.04 (-0.09, 0.16)	0.05 (-0.07, 0.18)	0.05 (-0.07, 0.19)
	1000	-0.13 (-0.29, 0.02)	-0.03 (-0.20, 0.12)	-0.05 (-0.21, 0.10)
Commercial land use	1000	0.05 (-0.02, 0.13)	0.01 (-0.08, 0.07)	0.05 (-0.03, 0.12)
Government and Institutional	500	0.02 (-0.04, 0.08)	0.01 (-0.05, 0.07)	-0.05 (-0.12, 0.01)
Population	100	-0.03 (-0.1, 0.05)	-0.03 (-0.11, 0.05)	-0.08 (-0.16, 0.00)
	1000	0.09 (-0.02, 0.21)	0.12 (0.01, 0.23)	0.01 (-0.1, 0.13)
Roads land use	1000	-0.08 (-0.17, 0.01)	-0.02 (-0.11, 0.07)	0.07 (-0.02, 0.16)

Table SM-8 Buffer sizes, point estimates and 95% credible intervals at each campaign for the coefficients of easting, northing and land-use variables for hexane under Model 4.

Hexane				
Variable	Buffer radius (m)	Posterior Mean and 95% CI		
		December 2005	April 2006	August 2006
Intercept	-	2.57 (2.50, 2.65)	1.75 (1.68, 1.82)	0.34 (0.27, 0.42)
Easting	-	0.11 (0.00, 0.23)	0.01 (-0.11, 0.13)	0.10 (-0.02, 0.21)
Northing	-	-0.12 (-0.22, -0.02)	0.03 (-0.08, 0.13)	0.05 (-0.05, 0.16)
Average NO_x	50	-0.02 (-0.10, 0.06)	-0.06 (-0.14, 0.03)	0.00 (-0.08, 0.08)
	500	0.08 (-0.1, 0.27)	-0.03 (-0.21, 0.15)	0.07 (-0.12, 0.24)
	1000	-0.09 (-0.27, 0.07)	0.03 (-0.15, 0. 19)	-0.07 (-0.24, 0.11)
Buildings land use	1000	-0.01 (-0.14, 0.10)	0.04 (-0.07, 0.16)	0.02 (-0.1, 0.14)
Commercial land use	1000	0.04 (-0.06, 0.14)	0.10 (0.01, 0.21)	-0.01 (-0.11, 0.09)
Government and	200	-0.05 (-0.17, 0.06)	0.06 (-0.05, 0.17)	-0.07 (-0.19, 0.04)
Institutional	1000	0.03 (-0.06, 0.12)	-0.03 (-0.12, 0.06)	-0.01 (-0.1, 0.08)
Population	500	-0.01 (-0.14, 0.11)	-0.03 (-0.16, 0.1)	0.03 (-0.11, 0.15)
Residential land use	1000	-0.08 (-0.23, 0.06)	0.02 (-0.14, 0.16)	0.00 (-0.15, 0.16)
	100	0.02 (-0.07, 0.10)	-0.09 (-0.17, 0.00)	-0.06 (-0.15, 0.03)
	500	0.05 (-0.10, 0.19)	0.09 (-0.05, 0.24)	0.02 (-0.14, 0.16)
Resource and Industrial	1000	0.04 (-0.07, 0.16)	0.02 (-0.09, 0.14)	0.14 (0.03, 0.25)
Roads land use	100	0.07 (-0.04, 0.19)	0.11 (0.00, 0.22)	0.02 (-0.08, 0.14)
	200	-0.02 (-0.14, 0.10)	-0.07 (-0.19, 0.05)	0.04 (-0.08, 0.15)
Waterbody	1000	-0.03 (-0.14, 0.09)	0.07 (-0.04, 0.19)	0.02 (-0.10, 0.14)

Table SM-9 Buffer sizes, point estimates and 95% credible intervals, in brackets, at each campaign for the coefficients of easting, northing and land-use variables for 1,2,4-trimethylbenzene under Model 4.

Variable	Buffer radius (m)	1,2,4 - trimethylbenzene Posterior Mean and 95% CI		
		December 2005	April 2006	August 2006
Intercept	-	0.07 (0.03, 0.11)	-0.05 (-0.09, -0.01)	-0.03 (-0.07, 0.01)
Easting	-	0.09 (0.02, 0.15)	-0.07 (-0.13, -0.01)	0.10 (0.04, 0.16)
Northing	-	0.07 (0.01, 0.12)	-0.01 (-0.07, 0.04)	-0.06 (-0.11, -0.01)
Average NOx	200	0.00 (-0.09, 0.08)	0.01 (-0.08, 0.09)	0.06 (-0.02, 0.14)
	500	0.20 (0.08, 0.34)	0.04 (-0.08, 0.17)	-0.05 (-0.17, 0.09)
	1000	-0.04 (-0.14, 0.05)	0.02 (-0.07, 0.11)	0.03 (-0.07, 0.12)
Building	50	0.07 (0.01, 0.14)	0.02 (-0.04, 0.09)	0.07 (0.01, 0.14)
	200	0.01 (-0.06, 0.08)	0.02 (-0.04, 0.10)	-0.01 (-0.08, 0.07)
	1000	0.05 (-0.04, 0.16)	0.06 (-0.03, 0.16)	0.03 (-0.08, 0.13)
Government and Institutional	200	-0.03 (-0.09, 0.03)	0.00 (-0.07, 0.06)	-0.05 (-0.11, 0.01)
Residential	50	-0.01 (-0.05, 0.04)	0.01 (-0.04, 0.05)	0.01 (-0.03, 0.06)
	200	0.06 (-0.02, 0.14)	0.06 (-0.02, 0.14)	0.04 (-0.04, 0.11)
	500	-0.02 (-0.11, 0.08)	-0.03 (-0.11, 0.06)	-0.04 (-0.13, 0.05)
Roads	50	0.00 (-0.05, 0.05)	0.03 (-0.03, 0.08)	0.03 (-0.02, 0.09)
	100	-0.03 (-0.1, 0.04)	0.01 (-0.06, 0.08)	-0.06 (-0.13, 0.02)
	200	0.10 (0.01, 0.18)	0.02 (-0.06, 0.11)	0.04 (-0.05, 0.13)
	500	-0.15 (-0.25, -0.05)	-0.03 (-0.13, 0.06)	0.01 (-0.09, 0.10)
Open Area	100	-0.04 (-0.11, 0.03)	-0.01 (-0.08, 0.05)	0.03 (-0.04, 0.09)
Population	200	0.03 (-0.03, 0.10)	-0.01 (-0.07, 0.05)	-0.03 (-0.1, 0.03)
	1000	0.00 (-0.09, 0.09)	0.10 (0.00, 0.19)	0.02 (-0.08, 0.11)
Resource and Industrial	1000	-0.01 (-0.07, 0.05)	0.01 (-0.05, 0.08)	-0.01 (-0.07, 0.05)
Waterbody	1000	-0.02 (-0.08, 0.05)	0.01 (-0.06, 0.07)	0.03 (-0.03, 0.10)

4. Predicted surfaces

This section contains the posterior mean and posterior standard deviation for the predicted surfaces of each VOC. The standard deviation shows the uncertainty about our measurements.

Figure SM-1 Posterior mean and posterior standard deviation for the benzene predicted surface in the log scale across the December, April and August monitoring campaigns.

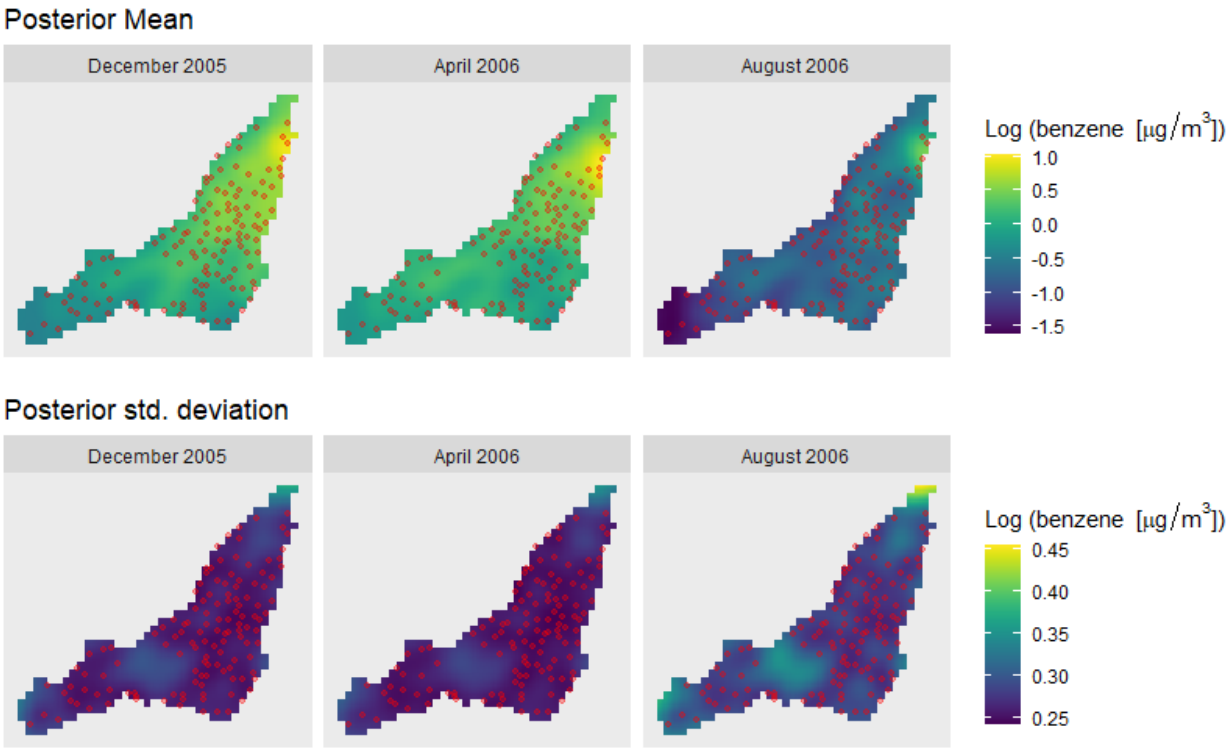


Figure SM-2 Posterior mean and posterior standard deviation for the n-decane predicted surface in the log scale across the December, April and August monitoring campaigns.

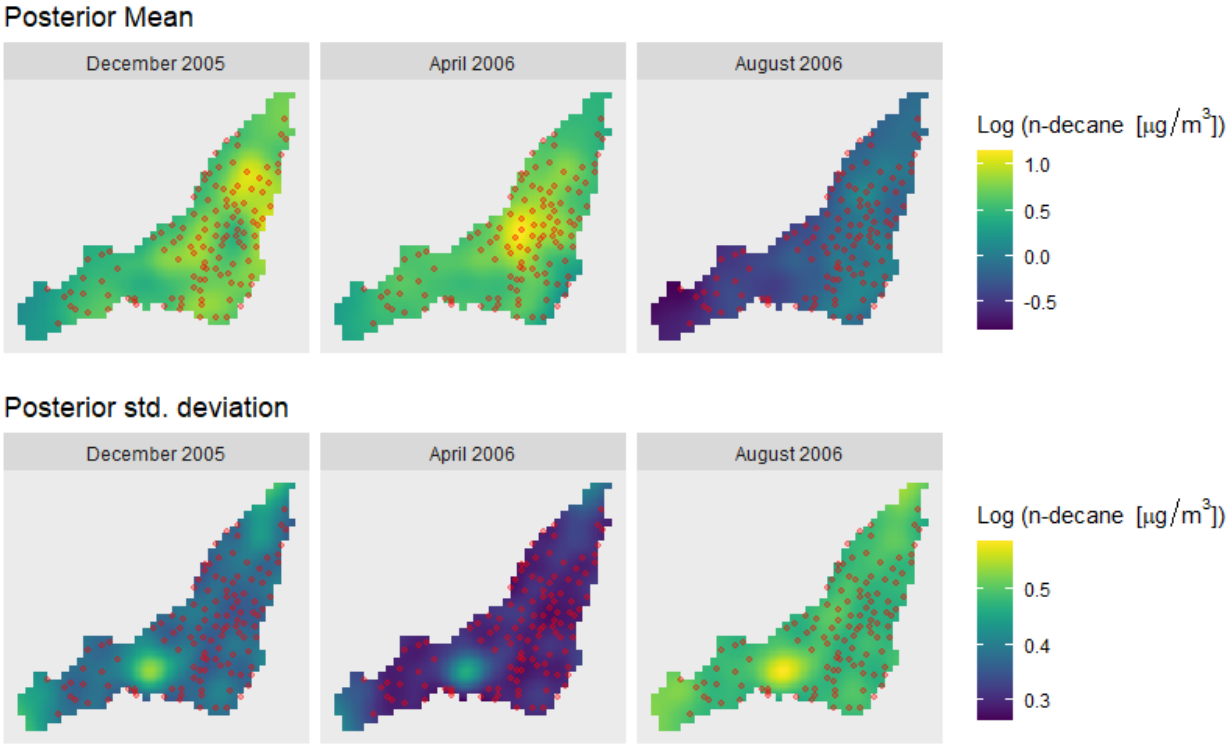


Figure SM-3 Posterior mean and posterior standard deviation for the ethylbenzene predicted surface in the log scale across the December, April and August monitoring campaigns.

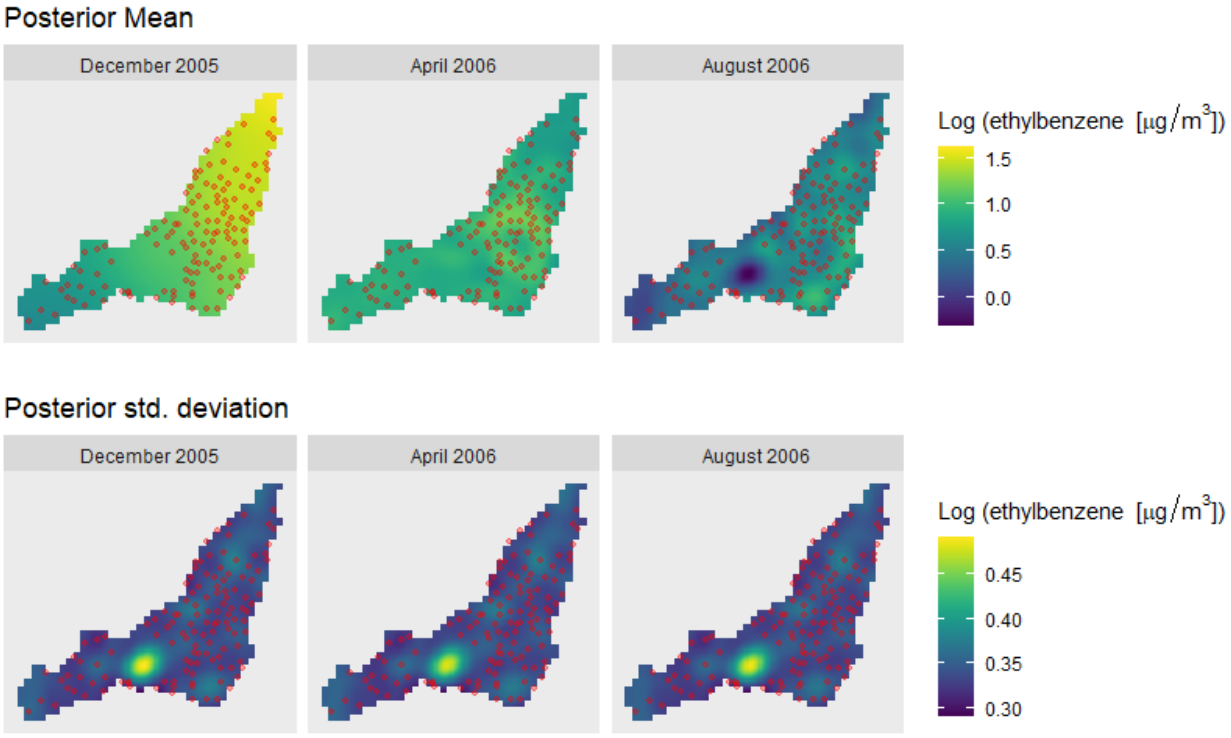


Figure SM-4 Posterior mean and posterior standard deviation for the hexane predicted surface in the log scale across the December, April and August monitoring campaigns.

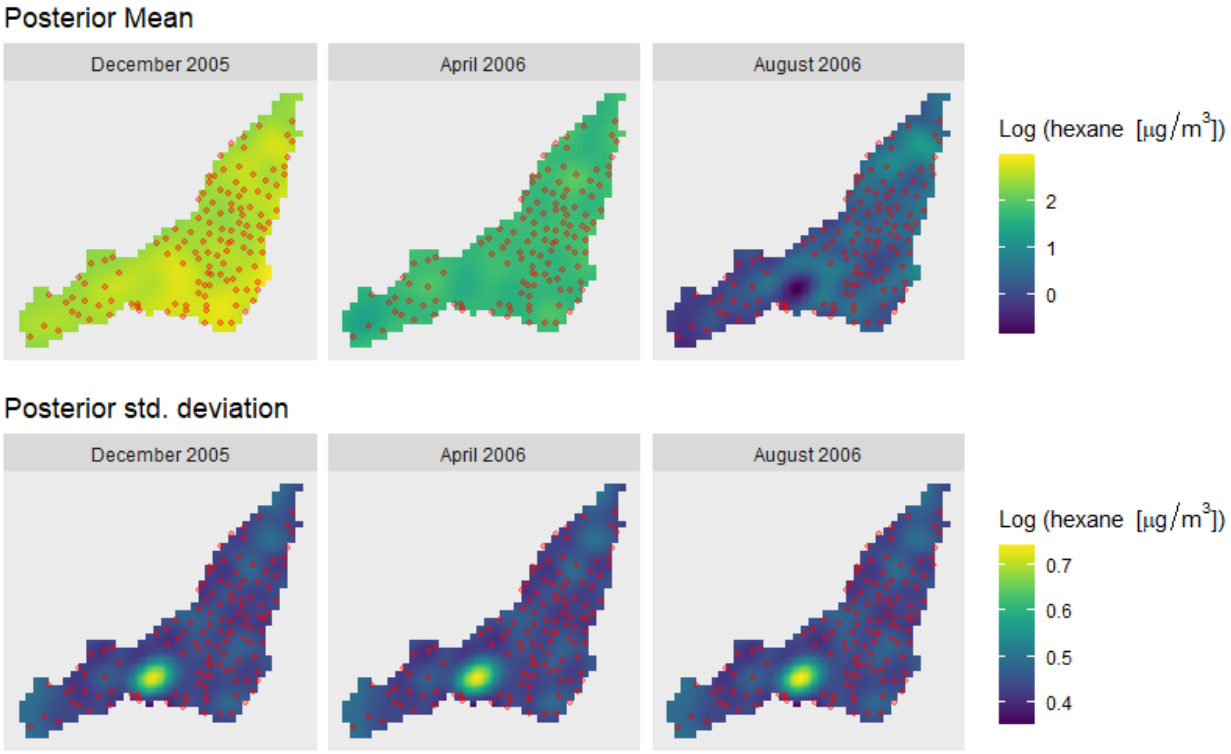
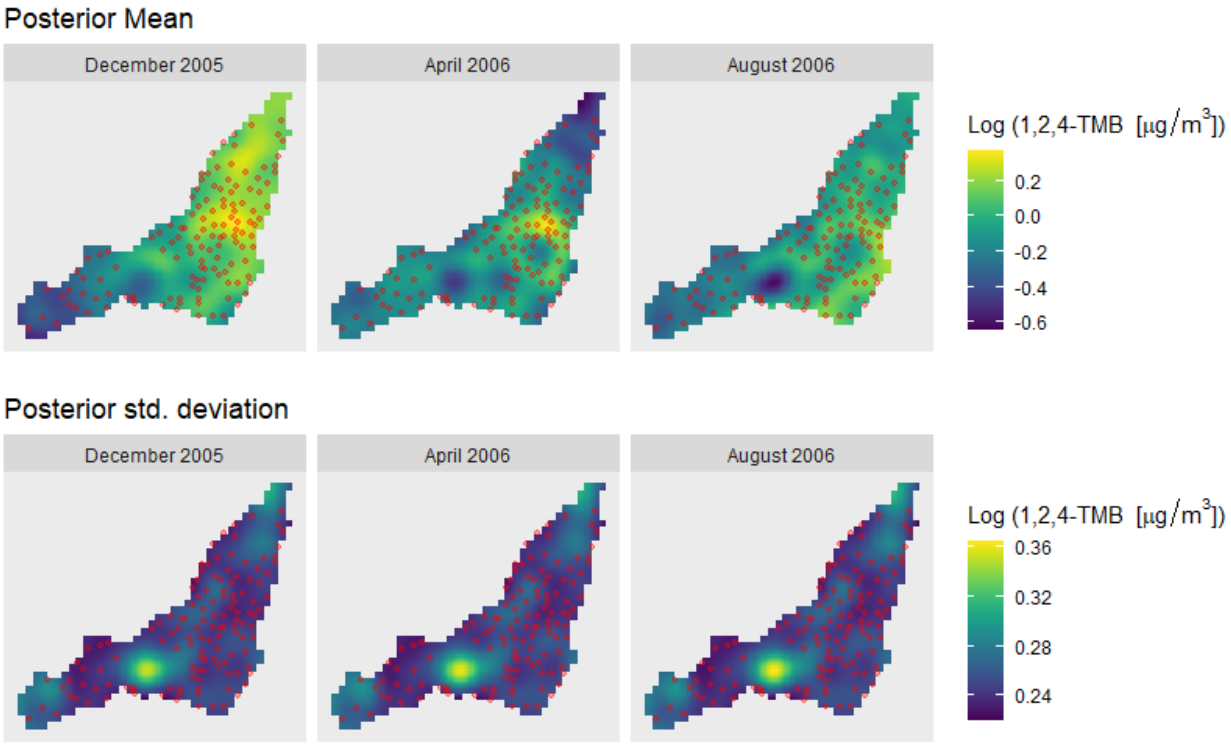


Figure SM-5 Posterior mean and posterior standard deviation for the 1,2,4-trimethylbenzene predicted surface in the log scale across the December, April and August monitoring campaigns



7.2 Appendix B: Supplementary Material for Chapter 4

Supplementary material: Within city spatiotemporal variation of pollen concentration in the city of Toronto, Canada

By Sara Zapata-Marin, Alexandra M. Schmidt, Scott Weichenthal, Daniel S.W. Katz, Tim Takaro, Jeffrey Brook and Eric Lavigne

The following document is organized as follows. Section 1 shows the land-use variables surfaces; Section 2 contains a table with a description of the predictors and their sources; Section 3 contains an explanation of the inference procedure, the spatial interpolation, and the forecasting procedure that could be used when trying to predict for future pollen seasons in the same area; Section 4 shows the values of the model comparison criteria WAIC and LOO for each pollen type and model; Section 5 shows the estimated missing and fitted values for the observed data, and Section 6 shows the predicted surface for total pollen in the original scale overlaid on a map of the city.

1. Predictors surface

Figures 1-5 show the surfaces of the land use predictors used for all the models. For the commercial, industrial, major roads land use and tree cover, a 1000m buffer size was used, while

for the grass cover we used a 500m buffer size. The red stars show the location of the monitoring stations.

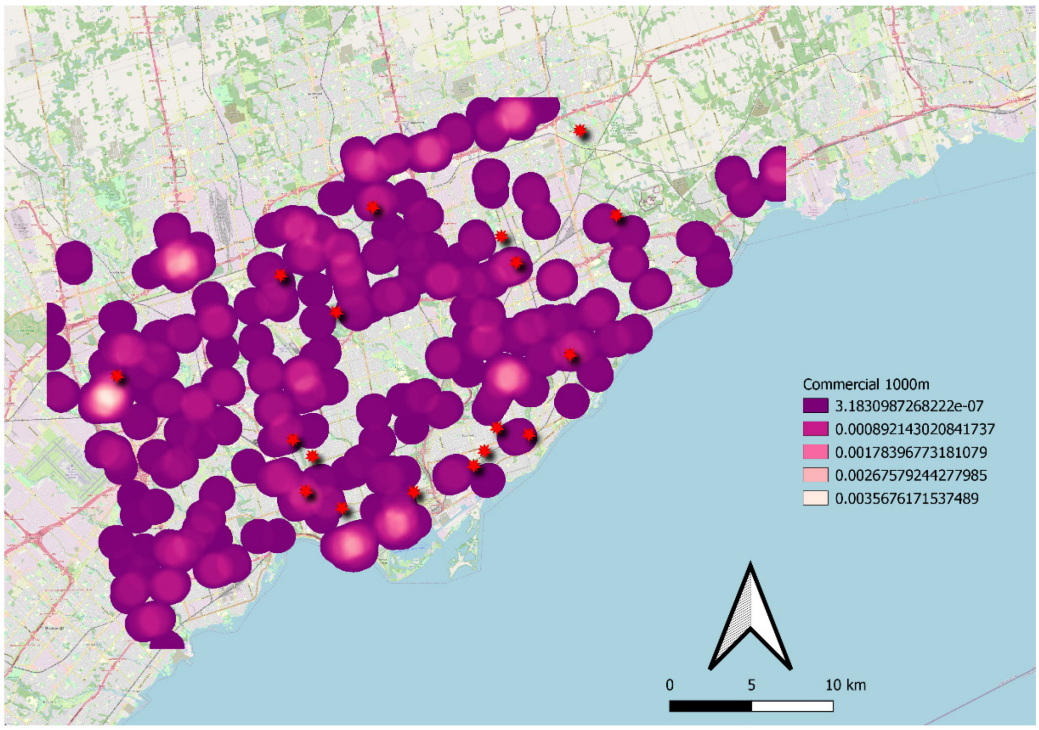


Figure 1 Commercial land use surface at a 1000 m buffer size together with the monitoring sites (red stars).

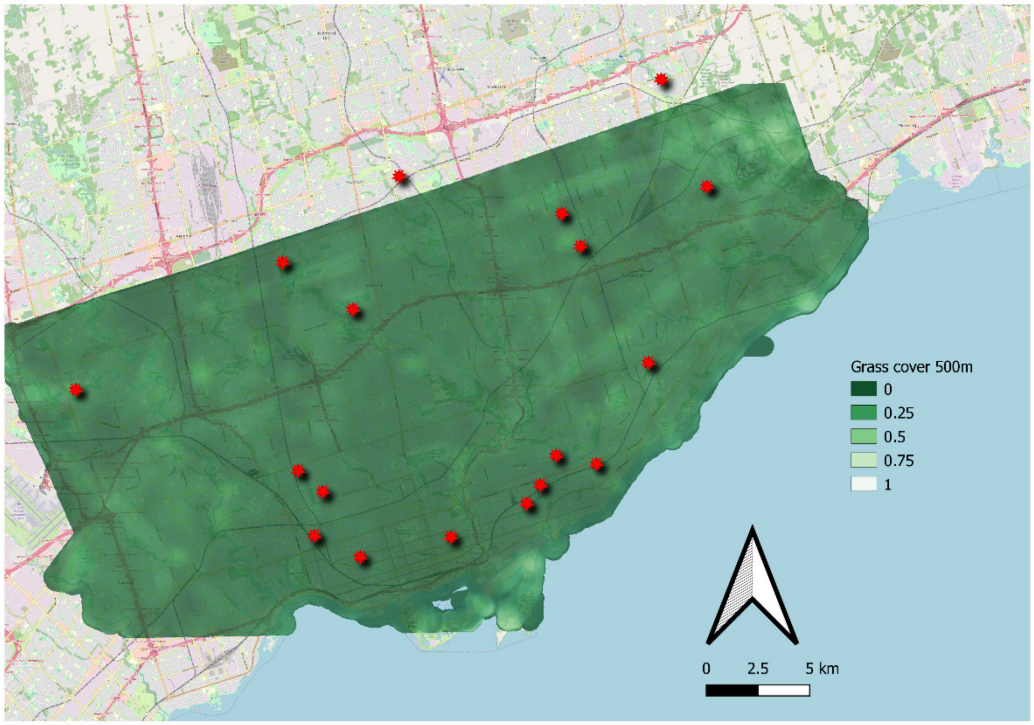


Figure 2 Grass cover surface at a 500 m buffer size together with the monitoring sites (red stars).

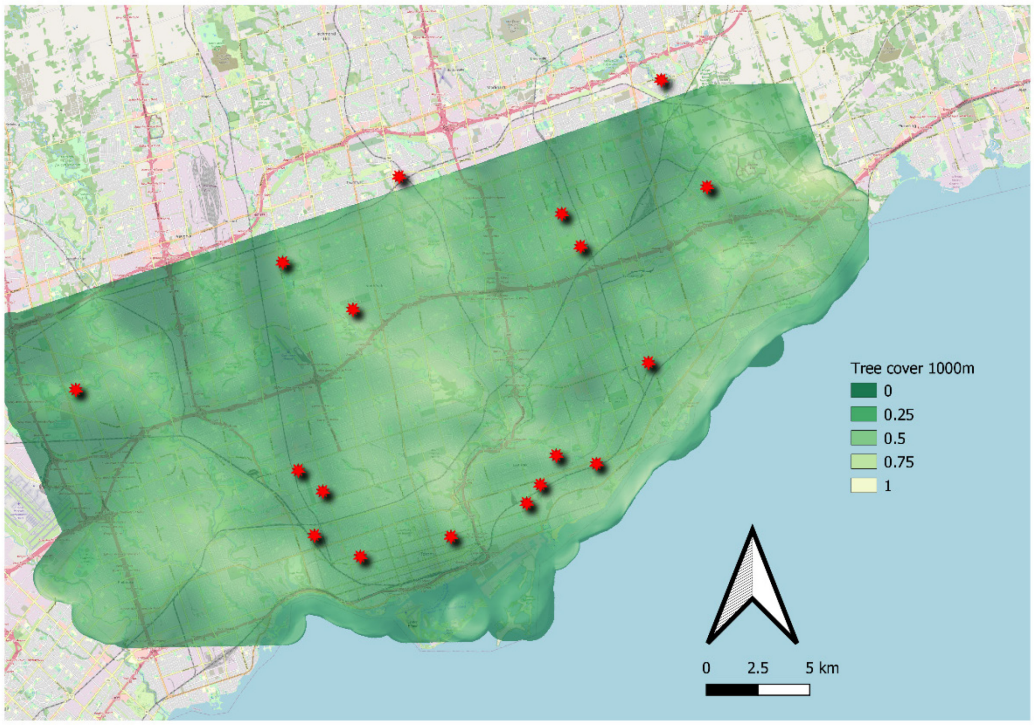


Figure 3 Tree cover surface at a 1000 m buffer size together with the monitoring sites (red stars).

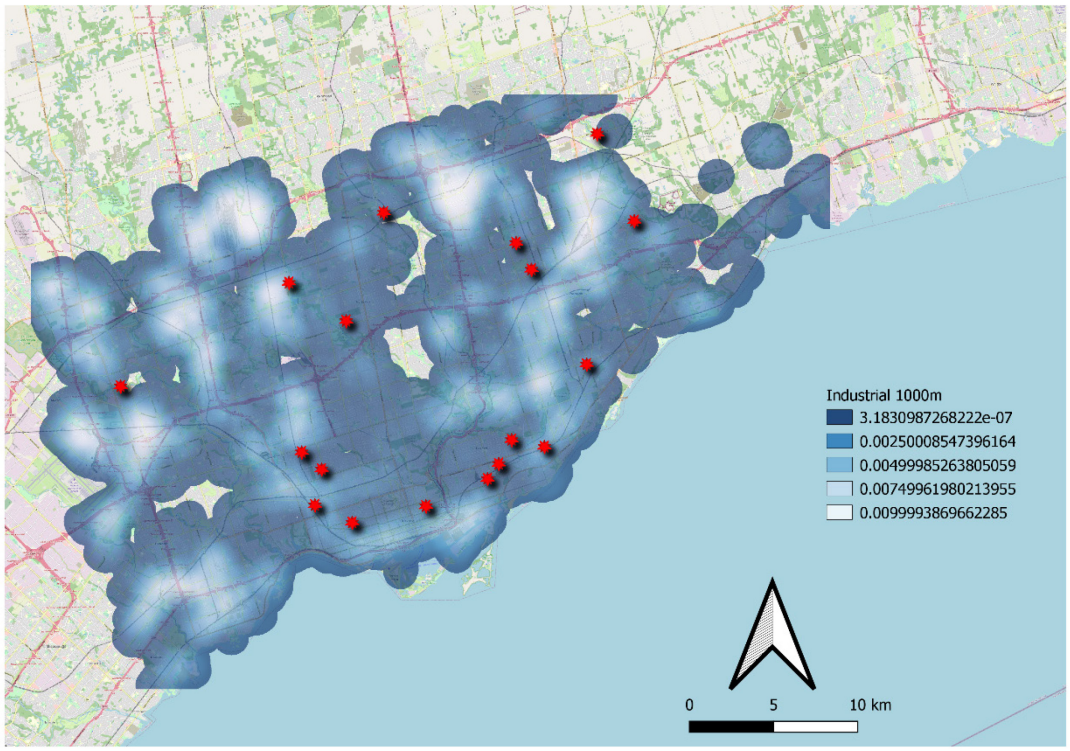


Figure 4 Industrial land use surface at a 1000 m buffer size together with the monitoring sites (red stars).

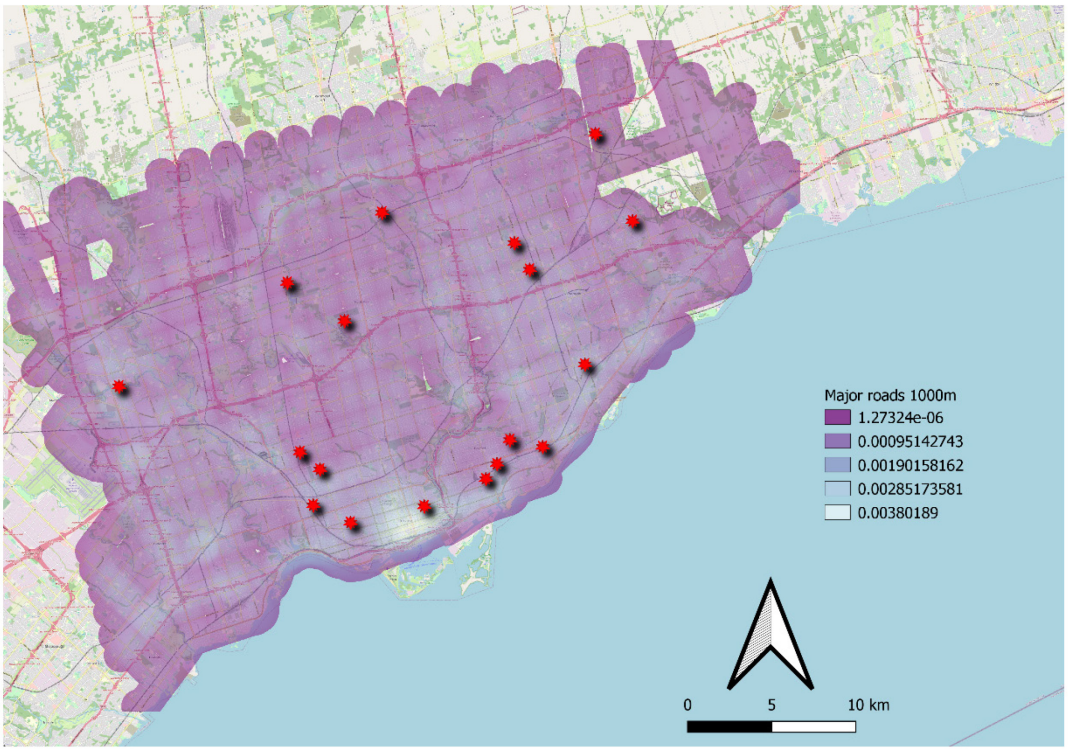


Figure 5 Major roads land use surface at a 1000 m buffer size together with the monitoring sites (red stars).

2. Description of the predictors

Table 1 shows the description of each of the variables used for all the models, as well as their units and their source. All variables were available for buffers 50m, 100m, 200m, 500m and 1000m around each monitoring site.

Variable	Unit	Description	Source
Commercial land use	Proportion (0 to 1)	Proportion of commercial land use within a defined buffer size.	DMTI Spatial Inc. (Markham Ontario)
Grass cover	Proportion (0 to 1)	Proportion of grass cover within a defined buffer size	City of Toronto 2018 Tree Canopy Study
Industrial land use	Proportion (0 to 1)	Proportion of industrial land use within a defined buffer size	DMTI Spatial Inc. (Markham Ontario)
Major roads	Proportion (0 to 1)	Proportion of buffer covered by a major road	DMTI Spatial Inc. (Markham Ontario)
Mean humidity	Percentage	Mean humidity within a day (or average humidity within a week)	Environment Canada
Mean temperature	Degrees Celsius	Mean temperature within a day (or average temperature within a week)	Environment Canada
Precipitation level	mm	Precipitation level within a day (or within a week)	Environment Canada
Tasseled cap transformation (TC) brightness	Index (-1 to 1)	Proportion of buffer covered by land use/cover with high albedo (overall brightness of the image)	Landsat 8 satellite images with high spatial resolution (10-50m)
Tasseled cap transformation (TC) greenness	Proportion (0 to 1)	Amount of phontosynthetically-active green vegetation within different buffers	Landsat 8 satellite images with high spatial resolution (10-50m)
Tree cover	Proportion (0 to 1)	Proportion of tree cover within a defined buffer size	City of Toronto 2018 Tree Canopy Study
Commercial land use	Proportion (0 to 1)	Proportion of commercial land use within a defined buffer size	DMTI Spatial Inc. (Markham Ontario)

Table 1 Description of land use predictors and climatic variables.

3. Statistical analysis

This section explains in more detail how the inference procedure and the spatial interpolation were performed. Additionally, subsection 3.3 explains how predictions can be performed for future pollen seasons in the same study area.

3.1 Inference procedure

Let the parameter vector for each model be defined as $\boldsymbol{\Omega}_1 = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \rho, \boldsymbol{\theta}, \tau^2, W_1\}$ for model 1, $\boldsymbol{\Omega}_{2,4} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \tau^2, W_1\}$ for models 2 and 4, and $\boldsymbol{\Omega}_3 = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \tau^2, W_1, W_2\}$ for model 3, where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_T)$. Assuming that observations $y_t(s)$ are available across n locations and T instants in time, the likelihood function for a vector parameter $\boldsymbol{\Omega}_j$ for model *j* is defined as,

$$p(\boldsymbol{\Omega}_j|\boldsymbol{y}_t) = \prod_{t=1}^{T}\prod_{i=1}^{n} \rho_t \, 1(y_t(s_i) = 0) + (1 - \rho_t)p(y_t(s_i)|\, \mu_t(s_i), \tau^2), \qquad (1)$$

where $p(y_t(s_i)|\, \mu_t(s), \tau^2)$ is defined as in Equation (1) in the main text, and for Model 1 $\rho_t = \rho$.

The inference procedure followed the Bayesian paradigm [46] where model specification is complete after assigning a prior distribution to the parameter vector. We assumed all the parameters to be independent a priori except for θ_t and γ_t in model 3, which depend on θ_{t-1} and γ_{t-1} respectively.

In general for all four models the prior was defined as,

$$p(\boldsymbol{\Omega}_j) = \prod_{t=1}^{T} p(\theta_t|\theta_{t-1}, W_1)\, p(\theta_0) \times p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(W_1)p(\tau)p(\boldsymbol{\zeta}). \qquad (2)$$

Additionally, for Model 1 the term $p(\rho)$ has to be added to the product, for models 2 and 4 we added $p(\boldsymbol{\gamma})$, and for model 3 the term $\prod_{t=1}^{T} p(\gamma_t|\gamma_{t-1}, W_2)\, p(\gamma_0)p(W_2)$ has to be added.

As explained in the main text, we assigned a zero mean normal prior distribution with relatively large variance for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}$; a half-Cauchy prior for τ, a half-standard normal distribution prior for W_1 and W_2, and a zero mean normal prior to θ_0 and γ_0 with a large variance. Finally, for model 1, a uniform prior distribution in the interval (0, 1) was assigned to ρ. Additionally, for all pollen types, there were 35 missing values spread across the weeks. Since we followed the Bayesian paradigm throughout our analysis, the missing values, for each of the pollen types, become parameters to be estimated.

Since the resultant posterior distribution does not have a closed form, we used Markov Chain Monte Carlo methods [43] to obtain samples from the resultant posterior distribution. In particular, the MCMC algorithm was implemented through the software.

3.2 Spatial interpolation

To obtain the predicted surfaces for each pollen type [46], we predicted the pollen concentration for the same period at a new set of sites $\boldsymbol{y}_t^0 = (y_t(s_{01}), \ldots, y_t(s_{0m}))'$ which corresponded to the centroids of a 1km by 1km cell size grid, and have associated vectors $\mathbf{x}(s_{0j})$, which contains the land-use variables and coordinates, and $\mathbf{u}_t(s_{0j})$, which contains the remote-sensing indices. The joint predictive density for this new set of sites is defined as,

$$
\begin{aligned}
&\mathrm{p}\left(\mathrm{y_t}(\mathrm{s_{0j}})\middle|\mathbf{y_t}, \mathbf{x}, \boldsymbol{u}_{\mathrm{t}}, \boldsymbol{x}(s_{0j}), \boldsymbol{u}_t(s_{0j})\right) \\
&\quad = \int \mathrm{p}\left(\mathrm{y_t}(\mathrm{s_{0j}})\middle|\, \mathbf{y_t}, \boldsymbol{x}(s_{0j}), \boldsymbol{u}_t(s_{0j}), \boldsymbol{\Omega}_{\mathrm{i}}\right)\mathrm{p}(\boldsymbol{\Omega}_{\mathrm{i}}|\mathbf{y_t}, \boldsymbol{x}, \boldsymbol{u}_t)\mathrm{d}\boldsymbol{\Omega}_{\mathrm{i}} \\
&\quad \approx \frac{1}{\mathrm{G}}\sum_{\mathrm{g}=1}^{\mathrm{G}} \mathrm{p}\left(\mathrm{y_t}(\mathrm{s_{0j}})\middle|\mathbf{y_t}, \boldsymbol{\Omega}_{\mathrm{i}}^{(\mathrm{g})}, \boldsymbol{x}(s_{0j}), \boldsymbol{u}_t(s_{0j})\right),
\end{aligned}
\tag{3}
$$

where *G* is the size of the sample that approximates the posterior distribution. Equation (3) shows you how to obtain an approximation for the posterior predictive distribution. Samples from this distribution are obtained using *composition sampling*: for each $\boldsymbol{\Omega}_i^{(g)}$ from the posterior sample, we obtain a sample from $\mathrm{y_t^{(g)}}(\mathrm{s_{0j}}) \sim p\left(\mathrm{y_t}(\mathrm{s_{0j}})\middle|\, \mathbf{y_t}, \boldsymbol{x}(s_{0j}), \boldsymbol{u}_t(s_{0j}), \boldsymbol{\Omega}_{\mathrm{i}}^{(\mathrm{g})}\right)$ for $\mathrm{g} = 1, \ldots, \mathrm{G}$. We obtain the set $\{\mathrm{y_t^{(1)}}(\mathrm{s_{0j}}), \mathrm{y_t^{(2)}}(\mathrm{s_{0j}}), \ldots, \mathrm{y_t^{(G)}}(\mathrm{s_{0j}})\}$ that form a sample from the posterior predictive density.

3.3 Forecasting

It might also be of interest to use this study to predict the pollen concentration for future pollen seasons in the same study area. In the Dynamic Linear Models literature, this is known as *forecasting* [49]. If we have enough and updated information on the temporal covariates (i.e. weather related variables), and using the recursive form of the time varying-mean component, we can compute the *k*-step-ahead forecast updating sequentially as new data becomes available. In this context, *k*-steps-ahead would mean *k* weeks into the next pollen season.

The *k*-steps-ahead forecast distribution of the pollen concentration at site *s* for models 1, 2 and 4 is defined as,

$$p(\boldsymbol{y}_{t+k}|\boldsymbol{y})$$

$$= \int \prod_{j=1}^{k} \left[p\left(\boldsymbol{y}_{t+j}\middle|\boldsymbol{X}, \boldsymbol{u}_{t+j}, \boldsymbol{z}_{t+j}, \rho_{t+j}, \tau, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}\right) p\left(\theta_{t+j}\middle|\theta_{t+j-1}, W_1\right)\right] p(\boldsymbol{\Omega}_i|\boldsymbol{y})\, d\boldsymbol{\Omega}_i^*, \tag{4}$$

where $\boldsymbol{\Omega}_i^* = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathrm{W}_1, \tau, \theta_{T+1}, \ldots, \theta_{T+k}, \gamma_{T+1}, \ldots, \gamma_{T+k})$. Note that for model 4, the component $p\left(\gamma_{t+j}\middle|\gamma_{t+j-1}, W_2\right)$ needs to be added to the product (for more information on forecasting see Chapter 4 subsection 4.4 in [49]). Composition sampling can be used to obtain samples from this posterior predictive distribution.

4. WAIC and LOO

Tables 2 shows the Widely Applicable Information Criterion (WAIC), and leave-one-out cross validation (LOO) for each of the fitted models and each type of pollen. This table contains the point estimate of the expected log pointwise predictive density (elpd_waic, elpd_loo), the effective number of parameters (p_waic, p_loo), the information criterion waic (-2* elpd_waic), and the leave-one-out cross-validation ($-2*$elpd_loo). We chose the model with the smallest values of WAIC and LOO, the best model among the fitted ones.

Model	elpd_waic	p_waic	waic	elpd_loo	p_loo	looic
	Grass Pollen					
Model 1	-1047.99	33.04	2095.97	-1048.46	33.51	2096.91
Model 2	-947.57	34.45	1895.14	-948.01	34.89	1896.02
Model 3	-934.44	41.70	1868.88	-935.00	42.26	1870.00
Model 4	-940.01	34.98	1880.03	-940.45	35.42	1880.90
	Tree Pollen					
Model 1	-1647.34	28.49	3294.67	-1647.66	28.82	3295.32
Model 2	-1545.10	32.08	3090.19	-1545.65	32.63	3091.30
Model 3	-1507.72	33.04	3015.44	-1508.28	33.60	3016.55
Model 4	-1564.79	30.74	3129.58	-1565.14	31.10	3130.28
	Weed Pollen					
Model 1	-1427.29	30.55	2854.58	-1427.68	30.93	2855.36
Model 2	-1282.25	31.77	2564.51	-1282.72	32.24	2565.44
Model 3	-1280.28	36.79	2560.55	-1280.93	37.45	2561.87
Model 4	-1278.99	30.94	2557.99	-1279.40	31.34	2558.80
	Total Pollen					
Model 1	-2369.47	37.88	4738.95	-2369.47	38.10	4739.38
Model 2	-2361.15	38.97	4722.29	-2361.44	39.27	4722.89
Model 3	-2362.18	41.73	4724.35	-2362.74	42.30	4725.48
Model 4	-2360.32	39.57	4720.63	-2360.61	39.87	4721.23

Table 2 Values of the model comparison criteria WAIC and LOO for each of the fitted models for the grass, tree, weed and total pollen. Smaller values indicates best fitted model.

5. Estimated missing and fitted values

Figure 6 shows the estimated missing (open circles), fitted (solid line), and observed values (solid circles) at some of the observed locations for each pollen types. The other locations not shown here exhibit a similar pattern.

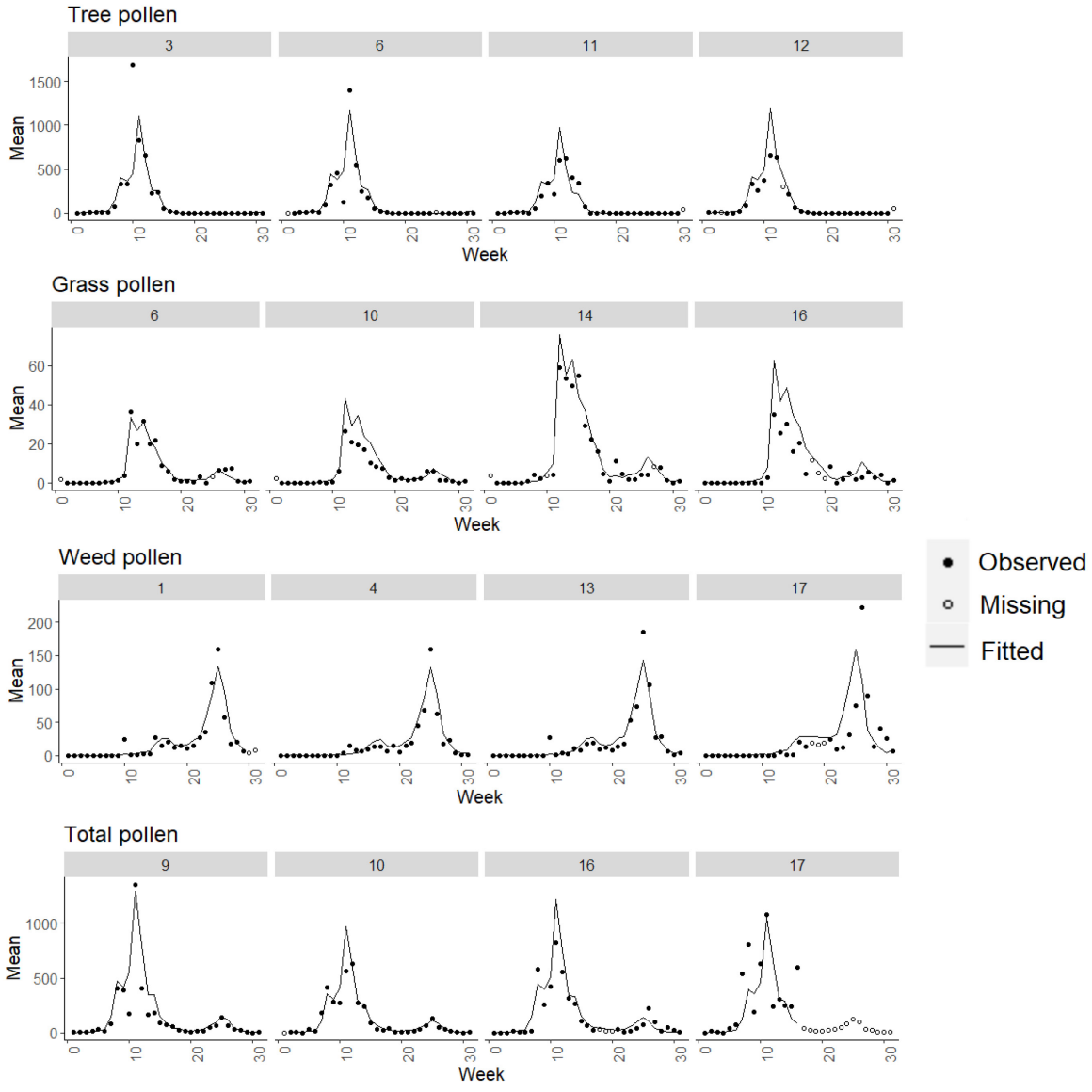


Figure 6 Observed values (solid circles) along with the estimated missing (open circles) and fitted (solid line) values at some sites for each pollen type.

6. Predicted surface

Figure 7 shows the predicted surface for the total pollen in the original scale overlaid on a map of Toronto for the week of May 13, 2018.

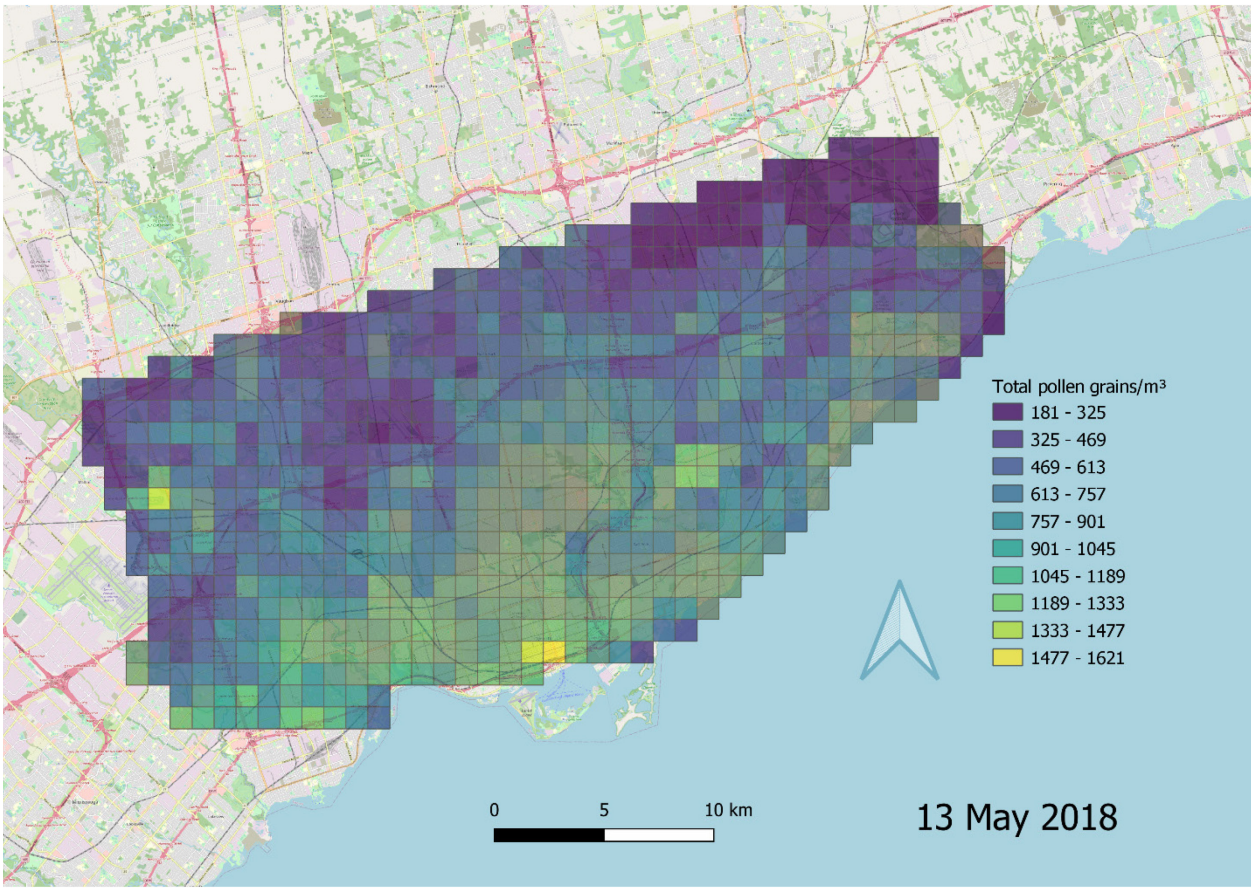


Figure 7 Predicted surface of the total pollen concentration in the original scale overlaid over a map of the city of Toronto.

References

43. Gamerman D, Lopes HF. Markov chain Monte Carlo : stochastic simulation for Bayesian inference. 2nd ed. / ed. Boca Raton: Taylor & Francis; 2006.

46. Banerjee S, Gelfand AE, Carlin BP. Hierarchical modeling and analysis for spatial data. Boca Raton, Florida Chapman & Hall/CRC Press; 2015.

49. West M, Harrison J. Bayesian forecasting and dynamic models. New York: Springer; 1997.

7.3 Appendix C: Supplementary Material for Chapter 5

Supplementary Material: Modelling temporally misaligned data across space: the case of total pollen concentration in Toronto

Sara Zapata-Marin, Alexandra M. Schmidt, Scott Weichenthal, Eric Lavigne

The present document is organized as follows. Section S.1 contains the posterior full conditional distributions obtained from the resultant posterior distribution described in subsection 3.1 of the main manuscript. Section S.2 shows the posterior summaries for the simulation studies of models M2 and M3 along with their estimated unobserved daily measurements at the weekly sites for these two models. Finally, section S.3 shows the results for the estimated fitted values for the daily sites and the predicted surface from section 4.2 in the main manuscript.

1. Simulation studies

Figures 1 and 2 show the posterior summaries (mean and 95% credible intervals) of the parameters estimates for: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ, τ, ϕ, $\mathbf{W}$, and the components of the state vector $\boldsymbol{\theta}$ at times t = 5, 10, 50, 150, 200, for models M2 and M3 described in Section 5.6. For the majority of the samples, the actual true values fall within the 95% credible interval of the posterior.

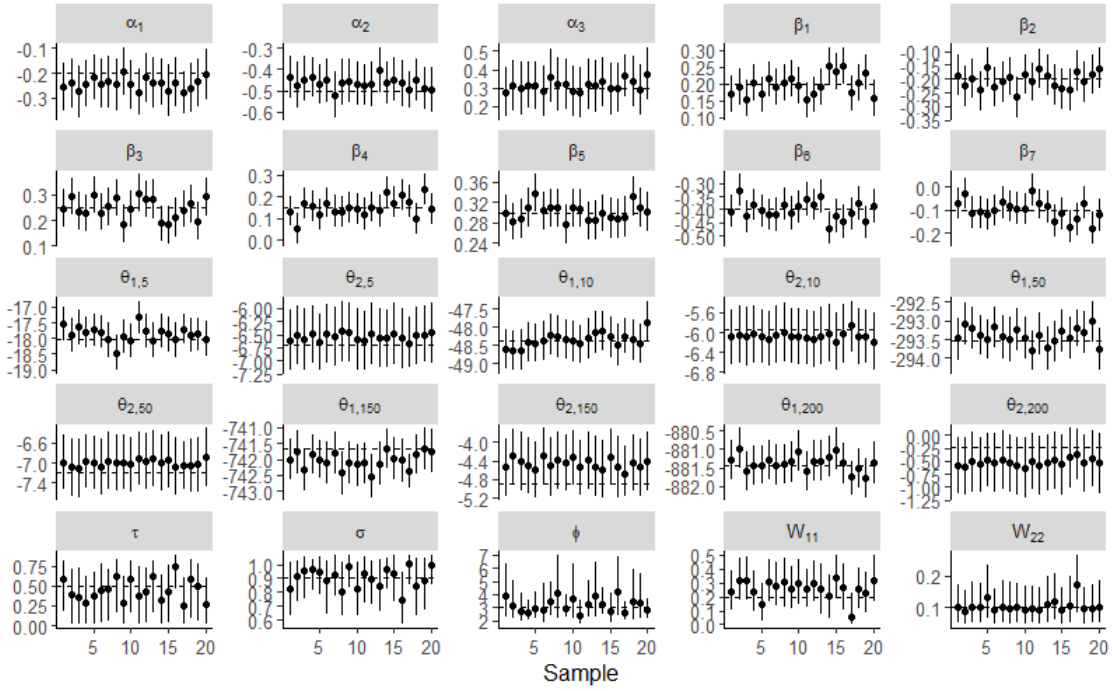


Figure S. 1 Posterior summary for 20 simulations for a level growth model (M2). Mean (solid circles), and 95% posterior credible intervals (vertical solid lines) of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\theta_{i,5}$, $\theta_{i,10}$, $\theta_{i,50}$,$\theta_{i,150}$, $\theta_{i,200}$, τ, σ, ϕ, and the elements of $\mathbf{W}$. Within each panel, the dashed line represents the true value for each parameter used to generate the data.

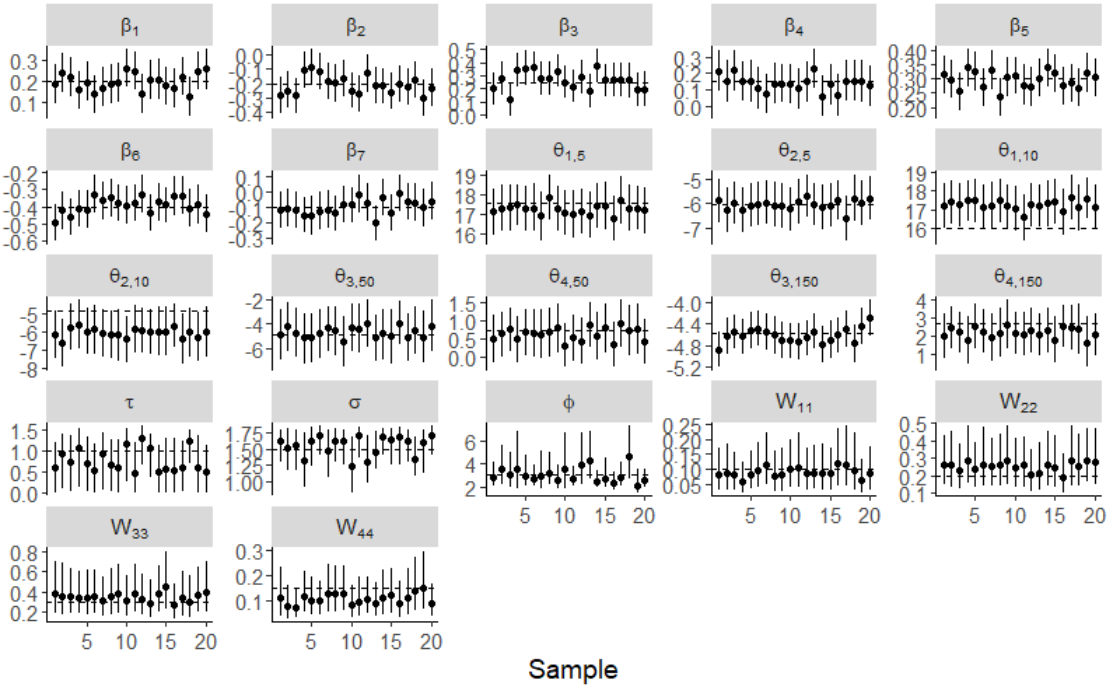


Figure S. 2 Posterior summary for 20 simulations for a dynamic standard regression model (M3). Mean (solid circles) and 95% posterior credible intervals (vertical solid lines) of **β**, $\theta_{i,5}$, $\theta_{i,10}$, $\theta_{i,50}$, $\theta_{i,150}$, τ, σ, ϕ, and the elements of **W**. Each dashed line represents the true value for each parameter used to generate the simulation studies.

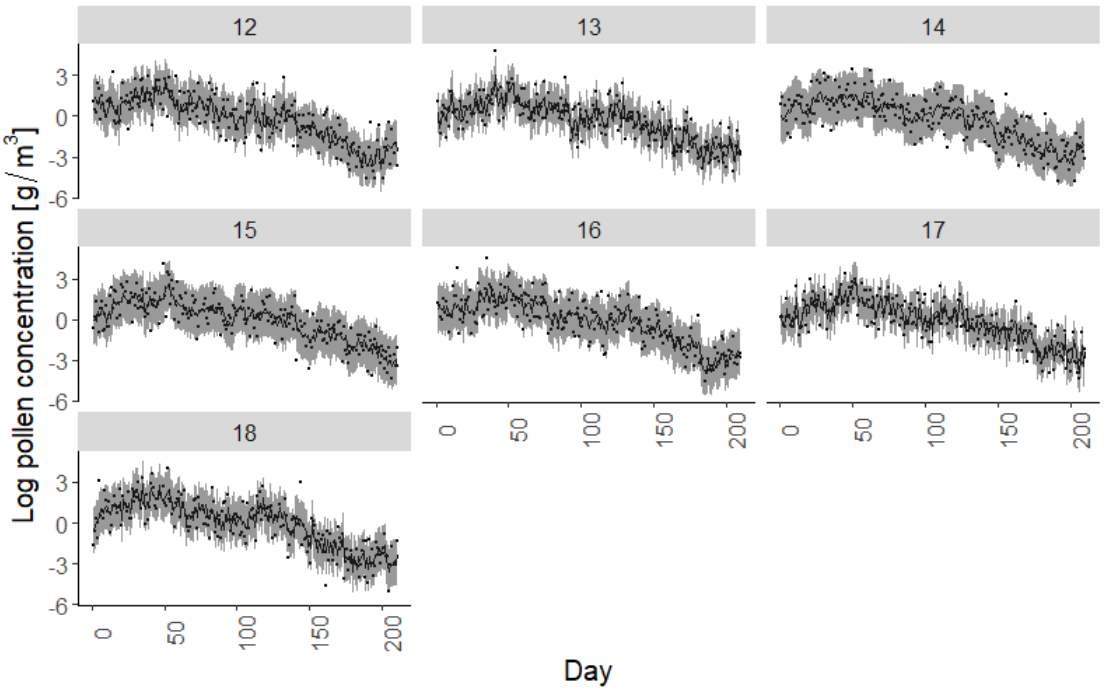


Figure S. 3 Posterior mean (solid line) and posterior 95% CI (shaded area) for the unobserved daily values at the weekly sites compared to the true values (solid dots) of a simulation of the model M2.

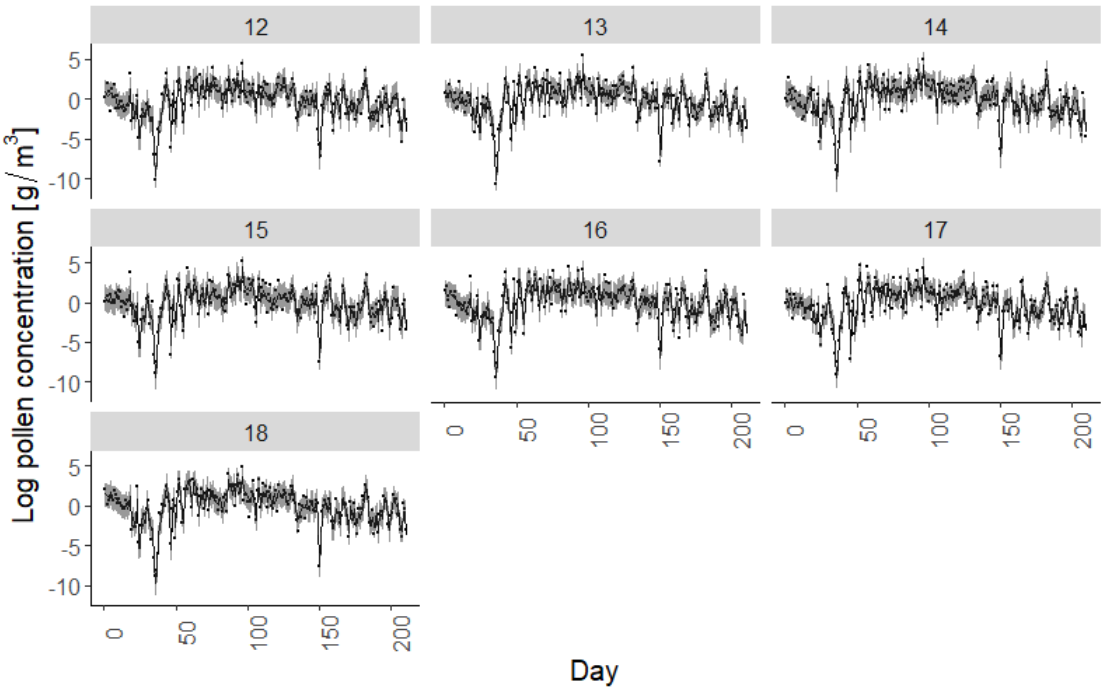


Figure S. 4 Posterior mean (solid line) and posterior 95% CI (shaded area) for the unobserved daily values at the weekly sites compared to the true values (solid dots) of a simulation of the model M3.

Additionally, we investigated the ability of the proposed model to recover the removed daily measurements at the weekly sites. Figures 3 and 4 show the posterior mean of the unobserved fine-scale observations at the sites where only coarse scale measurements were provided to the model, along with the 95% credible intervals across days for another simulation of models M2 and M3. For both models, the majority of the removed daily observations are included within the 95% credible interval.

2. The case of total pollen in Toronto

2.1. Fitted values

Panels of Figure 5 show the mean (solid line) and 95% credible interval (grey) for the fitted values along with the observations (solid dots) at the 11 sites where daily measurements of pollen were observed. For all sites, the majority of the observations are included within the 95% credible interval with the exception of some values that correspond to observations when the observed pollen concentration is zero.

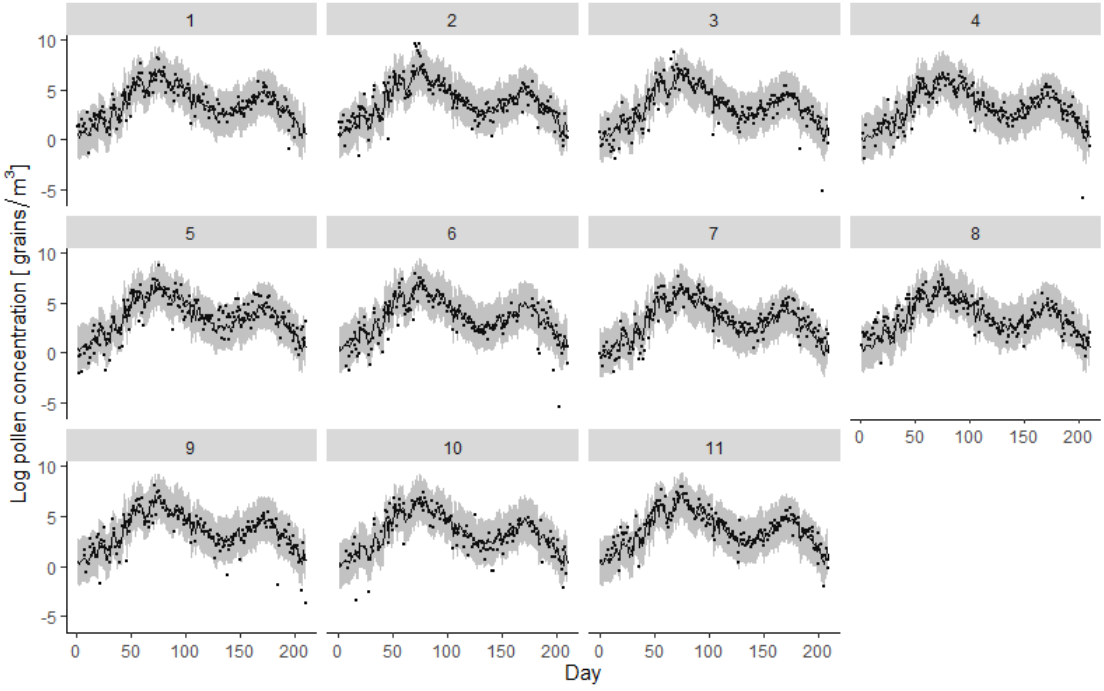


Figure S. 5 Fitted values for 11 sites with daily measurements of total pollen in the city of Toronto. Measured values (solid dots), estimated measurements (solid black line), and 95% posterior credible intervals (grey band) for the log of the total pollen concentration in grains/m^3 for 210 days.

2.2. Predicted surfaces

Figure 6 shows the posterior predictive mean (upper panel), and standard deviation (lower panel) for the predicted surface of the misalignment model. Since the best model among the fitted ones does not have a spatial structure the uncertainty across space varies a lot depending on the location of the cell grid.

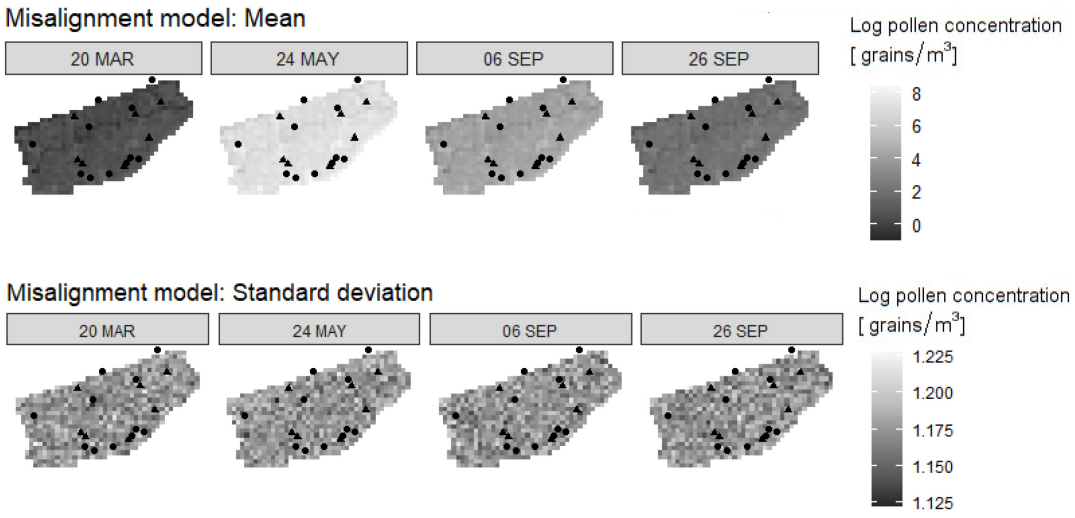


Figure S. 6 Posterior predictive mean (upper panel) and standard deviation (lower panel) for the log concentration of total pollen in Toronto for the best model among the fitted ones.

Figure 7 shows the posterior predictive mean for the predicted surface for March 20 and the week of March 18 for the temporal misalignment, daily and weekly model respectively. These maps show how the temporal misalignment predicted surface is a combination of the weekly and daily data. For example, the weekly model shows higher concentrations in the middle part of the city that are not shown in the daily model, but that are also being captured by the misalignment model. This can be explained by the presence of some monitoring sites that measured the weekly

concentration in the south and central part of the city (sites 12, 16, and 17 in Figure 1 of the main manuscript).

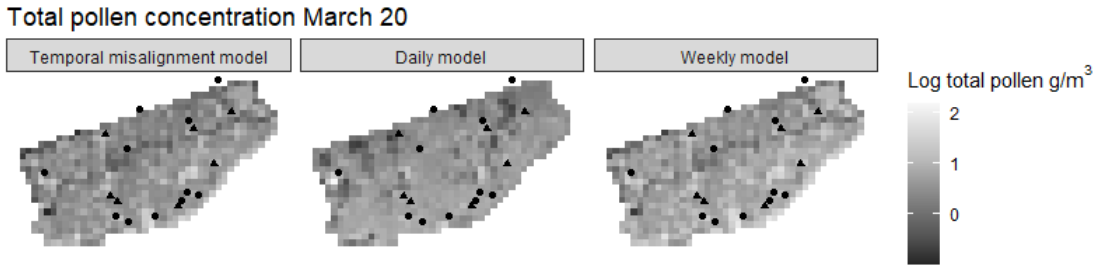


Figure S. 7 Posterior predictive mean for the log concentration of total pollen in Toronto for March 20 and the week of March 18.

8 General references

1. Prüss-Üstün A, Wolf J, Corvalán C, Bos R, Neira M. Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks: World Health Organization; 2016.

2. World Health Organization. Ambient air pollution: a global assessment of exposure and burden of disease. Geneva: World Health Organization; 2016.

3. World Health Organization. Air Pollution. Available from https://www.who.int/health-topics/air-pollution

4. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health. 2020;8.

5. Almetwally AA, Bin-Jumah M, Allam AA. Ambient air pollution and its influence on human health and welfare: an overview. Environmental Science and Pollution Research. 2020;27(20):24815-30.

6. World Health O. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: World Health Organization; 2021 2021.

7. Reimann S, Lewis AC. Anthropogenic VOCs. Volatile Organic Compounds in the Atmosphere2007. p. 33-81.

8. Koppmann R. Volatile organic compounds in the atmosphere: John Wiley & Sons; 2008.

9. Muralikrishna IV, Manickam V. Chapter Eight - Environmental Risk Assessment. In: Muralikrishna IV, Manickam V, editors. Environmental Management: Butterworth-Heinemann; 2017. p. 135-52.

10. Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment. 2008;42(33):7561-78.

11. Reid CE, Gamble JL. Aeroallergens, allergic disease, and climate change: impacts and adaptation. EcoHealth. 2009;6:458-70.

12. Peden D, Reed CE. Environmental and occupational allergies. Journal of Allergy and Clinical Immunology. 2010;125(2, Supplement 2):S150-S60.

13. Rider CF, Yamamoto M, Günther OP, Hirota JA, Singh A, Tebbutt SJ, et al. Controlled diesel exhaust and allergen coexposure modulates microRNA and gene expression in humans: effects on inflammatory lung markers. Journal of Allergy and Clinical Immunology. 2016;138(6):1690-700.

14. Singer BD, Ziska LH, Frenz DA, Gebhard DE, Straka JG. Increasing Amb a 1 content in common ragweed (Ambrosia artemisiifolia) pollen as a function of rising atmospheric CO2 concentration. Functional Plant Biology. 2005;32(7):667-70.

15. Zhang Q, Qiu Z, Chung KF, Huang S-K. Link between environmental air pollution and allergic asthma: East meets West. Journal of thoracic disease. 2015;7(1):14.

16. D'Amato G, Bergmann KC, Cecchi L, Annesi-Maesano I, Sanduzzi A, Liccardi G, et al. Climate change and air pollution. Allergo Journal International. 2014;23(1):17-23.

17. D'Amato G, Cecchi L. Effects of climate change on environmental factors in respiratory allergic diseases. Clinical & Experimental Allergy. 2008;38(8):1264-74.

18. Sierra-Heredia C, North M, Brook J, Daly C, Ellis AK, Henderson D, et al. Aeroallergens in Canada: Distribution, Public Health Impacts, and Opportunities for Prevention. Int J Environ Res Public Health. 2018;15(8).

19. Park HJ, Lee J-H, Park KH, Kim KR, Han MJ, Choe H, et al. A Six-Year Study on the Changes in Airborne Pollen Counts and Skin Positivity Rates in Korea: 2008–2013. Yonsei Med J. 2016;57(3):714-20.

20. Ariano R, Canonica GW, Passalacqua G. Possible role of climate changes in variations in pollen seasons and allergic sensitizations during 27 years. Annals of Allergy, Asthma & Immunology. 2010;104(3):215-22.

21. Ziello C, Sparks TH, Estrella N, Belmonte J, Bergmann KC, Bucher E, et al. Changes to airborne pollen counts across Europe. PloS one. 2012;7(4):e34076.

22. Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, et al. Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science. 1997;11(7):699-718.

23. Katz DSW, Carey TS. Heterogeneity in ragweed pollen exposure is determined by plant composition at small spatial scales. Sci Total Environ. 2014;485-486:435-40.

24. Amann M, Kiesewetter G, Schöpp W, Klimont Z, Winiwarter W, Cofala J, et al. Reducing global air pollution: the scope for further policy interventions. Philosophical Transactions of the Royal Society A. 2020;378(2183):20190331.

25. Katz DSW, Batterman SA. Urban-scale variation in pollen concentrations: a single station is insufficient to characterize daily exposure. Aerobiologia. 2020.

26. Oiamo TH, Johnson M, Tang K, Luginaah IN. Assessing traffic and industrial contributions to ambient nitrogen dioxide and volatile organic compounds in a low pollution urban environment. Science of The Total Environment. 2015;529:149-57.

27. Crouse DL, Goldberg MS, Ross NA. A prediction-based approach to modelling temporal and spatial variability of traffic-related air pollution in Montreal, Canada. Atmospheric Environment. 2009;43(32):5075-84.

28. Henderson SB, Beckerman B, Jerrett M, Brauer M. Application of Land Use Regression to Estimate Long-Term Concentrations of Traffic-Related Nitrogen Oxides and Fine Particulate Matter. Environmental Science & Technology. 2007;41(7):2422-8.

29. Mohammadi A, Ghassoun Y, Löwner M-O, Behmanesh M, Faraji M, Nemati S, et al. Spatial analysis and risk assessment of urban BTEX compounds in Urmia, Iran. Chemosphere. 2020;246:125769.

30. Atari DO, Luginaah IN. Assessing the distribution of volatile organic compounds using land use regression in Sarnia," Chemical Valley", Ontario, Canada. Environmental Health. 2009;8(1):1-14.

31. Wang J, Cohan DS, Xu H. Spatiotemporal ozone pollution LUR models: Suitable statistical algorithms and time scales for a megacity scale. Atmospheric Environment. 2020;237:117671.

32. Alvarez-Mendoza CI, Teodoro A, Ramirez-Cando L. Spatial estimation of surface ozone concentrations in Quito Ecuador with remote sensing data, air pollution measurements and meteorological variables. Environmental Monitoring and Assessment. 2019;191(3):155.

33. Masiol M, Squizzato S, Chalupa D, Rich DQ, Hopke PK. Spatial-temporal variations of summertime ozone concentrations across a metropolitan area using a network of low-cost monitors to develop 24 hourly land-use regression models. Science of The Total Environment. 2019;654:1167-78.

34. Mukerjee S, Smith LA, Johnson MM, Neas LM, Stallings CA. Spatial analysis and land use regression of VOCs and NO2 from school-based urban air monitoring in Detroit/Dearborn, USA. Science of The Total Environment. 2009;407(16):4642-51.

35. Hjort J, Hugg TT, Antikainen H, Rusanen J, Sofiev M, Kukkonen J, et al. Fine-Scale Exposure to Allergenic Pollen in the Urban Environment: Evaluation of Land Use Regression Approach. Environ Health Perspect. 2016;124(5):619-26.

36. Weinberger KR, Kinney PL, Robinson GS, Sheehan D, Kheirbek I, Matte TD, et al. Levels and determinants of tree pollen in New York City. Journal of Exposure Science & Environmental Epidemiology. 2018;28(2):119-24.

37. Amini H, Yunesian M, Hosseini V, Schindler C, Henderson SB, Künzli N. A systematic review of land use regression models for volatile organic compounds. Atmospheric Environment. 2017;171:1-16.

38. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Third ed. Boca Raton: CRC Press; 2014.

39. McElreath R, O'Reilly for Higher E, Safari aORMC. Statistical Rethinking, 2nd Edition: Chapman and Hall/CRC; 2020.

40. Clark JS, Gelfand AE. A future for models and data in environmental science. Trends in Ecology & evolution. 2006;21(7):375-80.

41. Berliner LM. Hierarchical Bayesian time series models. Maximum entropy and Bayesian methods: Springer; 1996. p. 15-22.

42. Wikle CK. Hierarchical models in environmental science. International Statistical Review. 2003;71(2):181-99.

43. Gamerman D, Lopes HF. Markov chain Monte Carlo : stochastic simulation for Bayesian inference. 2nd ed. / ed. Boca Raton: Taylor & Francis; 2006.

44. de Valpine PaP, C. and Turek, D. and Michaud, N. and Anderson-Bergman, C. and Obermeyer, F. and Wehrhahn Cortes, C. and Rodriguez, A. and Temple Lang, D. and Paganin, S. NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. 0.11.1 ed2021.

45. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. 2.29 ed2022.

46. Banerjee S, Gelfand AE, Carlin BP. Hierarchical modeling and analysis for spatial data. Boca Raton, Florida Chapman & Hall/CRC Press; 2015.

47. Petris G, Petrone S, Campagnoli P. Dynamic linear models. Dynamic Linear Models with R: Springer; 2009. p. 31-84.

48. Schmidt AM, Lopes HF. Dynamic models. Handbook of Environmental and Ecological Statistics: CRC Press; 2019. p. 57-80.

49. West M, Harrison J. Bayesian forecasting and dynamic models. New York: Springer; 1997.

50. Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986;33(3):22.

51. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34(1):1-14.

52. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol. 2005;15(2):185-204.

53. Su JG, Jerrett M, Beckerman B, Verma D, Arain MA, Kanaroglou P, et al. A land use regression model for predicting ambient volatile organic compound concentrations in Toronto, Canada. Atmospheric Environment. 2010;44(29):3529-37.

54. Woodruff TJ, Axelrad DA, Caldwell J, Morello-Frosch R, Rosenbaum A. Public health implications of 1990 air toxics concentrations across the United States. Environmental Health Perspectives. 1998;106(5):245-51.

55. Wheeler AJ, Smith-Doiron M, Xu X, Gilbert NL, Brook JR. Intra-urban variability of air pollution in Windsor, Ontario--measurement and modeling for human exposure assessment. Environ Res. 2008;106(1):7-16.

56. Marć M, Bielawska M, Wardencki W, Namieśnik J, Zabiegała B. The influence of meteorological conditions and anthropogenic activities on the seasonal fluctuations of BTEX in the urban air of the Hanseatic city of Gdansk, Poland. Environmental Science and Pollution Research. 2015;22(15):11940-54.

57. Deville Cavellin L, Weichenthal S, Tack R, Ragettli MS, Smargiassi A, Hatzopoulou M. Investigating the Use Of Portable Air Pollution Sensors to Capture the Spatial Variability Of Traffic-Related Air Pollution. Environmental Science & Technology. 2016;50(1):313-20.

58. Gaeta A, Cattani G, Di Menno di Bucchianico A, De Santis A, Cesaroni G, Badaloni C, et al. Development of nitrogen dioxide and volatile organic compounds land use regression models to estimate air pollution exposure near an Italian airport. Atmospheric Environment. 2016;131:254-62.

59. Aguilera I, Sunyer J, Fernández-Patier R, Hoek G, Aguirre-Alfaro A, Meliefste K, et al. Estimation of Outdoor NOx, NO2, and BTEX Exposure in a Cohort of Pregnant Women Using Land Use Regression Modeling. Environmental Science & Technology. 2008;42(3):815-21.

60. Amini H, Hosseini V, Schindler C, Hassankhany H, Yunesian M, Henderson SB, et al. Spatiotemporal description of BTEX volatile organic compounds in a Middle Eastern megacity: Tehran Study of Exposure Prediction for Environmental Health Research (Tehran SEPEHR). Environmental Pollution. 2017;226:219-29.

61. Liu K, Zhang C, Cheng Y, Liu C, Zhang H, Zhang G, et al. Serious BTEX pollution in rural area of the North China Plain during winter season. Journal of Environmental Sciences. 2015;30:186-90.

62. Weber RW. Floristic zones and aeroallergen diversity. Immunology and Allergy Clinics of North America. 2003;23(3):357-69.

63. Ellis AK, North ML, Walker T, Steacy LM. Environmental exposure unit: a sensitive, specific, and reproducible methodology for allergen challenge. Annals of Allergy, Asthma & Immunology. 2013;111(5):323-8.

64. Ellis AK, Soliman M, Steacy L, Boulay M-È, Boulet L-P, Keith PK, et al. The Allergic Rhinitis – Clinical Investigator Collaborative (AR-CIC): nasal allergen challenge protocol optimization for studying AR pathophysiology and evaluating novel therapies. Allergy, Asthma & Clinical Immunology. 2015;11(1):16.

65. White JF, Bernstein DI. Key pollen allergens in North America. Annals of Allergy, Asthma & Immunology. 2003;91(5):425-35.

66. Portnoy J, Barnes C. Clinical relevance of spore and pollen counts. Immunology and Allergy Clinics of North America. 2003;23(3):389-410.

67. Martorano L, Erwin EA. Aeroallergen Exposure and Spread in the Modern Era. The Journal of Allergy and Clinical Immunology: In Practice. 2018;6(6):1835-42.

68. Werchan B, Werchan M, Mucke HG, Gauger U, Simoleit A, Zuberbier T, et al. Spatial distribution of allergenic pollen through a large metropolitan area. Environ Monit Assess. 2017;189(4):169.

69. Werchan B, Werchan M, Mücke H-G, Bergmann K-C. Spatial distribution of pollen-induced symptoms within a large metropolitan area—Berlin, Germany. Aerobiologia. 2018;34(4):539-56.

70. Peel RG, Hertel O, Smith M, Kennedy R. Personal exposure to grass pollen: relating inhaled dose to background concentration. Annals of Allergy, Asthma & Immunology. 2013;111(6):548-54.

71. Weinberger KR, Kinney PL, Lovasi GS. A review of spatial variation of allergenic tree pollen within cities. Arboriculture & Urban Forestry. 2015;41(2):57-68.

72. Rojo J, Oteros J, Pérez-Badia R, Cervigón P, Ferencova Z, Gutiérrez-Bustillo AM, et al. Near-ground effect of height on pollen exposure. Environmental Research. 2019;174:160-9.

73. Levetin E. Methods for aeroallergen sampling. Current allergy and asthma reports. 2004;4(5):376-83.

74. Weber RW. Outdoor aeroallergen sampling: not all that simple. Annals of Allergy, Asthma & Immunology. 2007;98(6):505-6.

75. Katz DSW, Dzul A, Kendel A, Batterman SA. Effect of intra-urban temperature variation on tree flowering phenology, airborne pollen, and measurement error in epidemiological studies of allergenic pollen. Sci Total Environ. 2019;653:1213-22.

76. Maya-Manzano JM, Sadyś M, Tormo-Molina R, Fernández-Rodríguez S, Oteros J, Silva-Palacios I, et al. Relationships between airborne pollen grains, wind direction and land cover using GIS and circular statistics. Science of The Total Environment. 2017;584-585:603-13.

77. Katz DSW, Batterman SA. Allergenic pollen production across a large city for common ragweed (Ambrosia artemisiifolia). Landsc Urban Plan. 2019;190.

78. Charalampopoulos A, Damialis A, Lazarina M, Halley JM, Vokou D. Spatiotemporal assessment of airborne pollen in the urban environment: The pollenscape of Thessaloniki as a case study. Atmospheric Environment. 2021;247:118185.

79. Rojo J, Rapp A, Lara B, Fernandez-Gonzalez F, Perez-Badia R. Effect of land uses and wind direction on the contribution of local sources to airborne pollen. Sci Total Environ. 2015;538:672-82.

80. Hugg TT, Hjort J, Antikainen H, Rusanen J, Tuokila M, Korkonen S, et al. Urbanity as a determinant of exposure to grass pollen in Helsinki Metropolitan area, Finland. PLOS ONE. 2017;12(10):e0186348.

81. Ríos B, Torres-Jardón R, Ramírez-Arriaga E, Martínez-Bernal A, Rosas I. Diurnal variations of airborne pollen concentration and the effect of ambient temperature in three sites of Mexico City. International Journal of Biometeorology. 2016;60(5):771-87.

82. Williams J, Koppmann R. Volatile Organic Compounds in the Atmosphere: An Overview. Volatile Organic Compounds in the Atmosphere2007. p. 1-32.

83. United States Environmental Protection Agency. Volatile organic compounds' impact on indoor air quality 2014. Available from https://www.epa.gov/indoor-air-quality-iaq/volatile-organic-compounds-impact-indoor-air-quality.

84. Centre international de recherche sur le cancer, IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Benzene. Lyon France; 2019.

85. Canada's air pollutant emissions inventory report. Environment and Climate Change Canada; 2021.

86. Gurney KR, Liang J, Roest G, Song Y, Mueller K, Lauvaux T. Under-reporting of greenhouse gas emissions in U.S. cities. Nature Communications. 2021;12(1):553.

87. Ren X, Salmon OE, Hansford JR, Ahn D, Hall D, Benish SE, et al. Methane Emissions From the Baltimore-Washington Area Based on Airborne Observations: Comparison to Emissions Inventories. Journal of Geophysical Research: Atmospheres. 2018;123(16):8869-82.

88. Turner AJ, Jacob DJ, Benmergui J, Wofsy SC, Maasakkers JD, Butz A, et al. A large increase in U.S. methane emissions over the past decade inferred from satellite data and surface observations. Geophysical Research Letters. 2016;43(5):2218-24.

89. Yifang Zhu WCH, Seongheon Kim, Si Shen, Constantinos Sioutas. Study of ultrafine particles near a major highway with heavy-duty diesel traffic. Atmospheric Environment. 2002;36(27):12.

90. Bozkurt Z, Üzmez ÖÖ, Döğeroğlu T, Artun G, Gaga EO. Atmospheric concentrations of SO2, NO2, ozone and VOCs in Düzce, Turkey using passive air samplers: Sources, spatial and seasonal variations and health risk estimation. Atmospheric Pollution Research. 2018;9(6):1146-56.

91. Kumar A, Singh D, Kumar K, Singh BB, Jain VK. Distribution of VOCs in urban and rural atmospheres of subtropical India: Temporal variation, source attribution, ratios, OFP and risk assessment. Sci Total Environ. 2018;613-614:492-501.

92. Li B, Ho SSH, Qu L, Gong S, Ho KF, Zhao D, et al. Temporal and spatial discrepancies of VOCs in an industrial-dominant city in China during summertime. Chemosphere. 2021;264(Pt 2):128536.

93. Basagaña X, Rivera M, Aguilera I, Agis D, Bouso L, Elosua R, et al. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmospheric Environment. 2012;54:634-42.

94. Crouse DL, Goldberg MS, Ross NA, Chen H, Labrèche F. Postmenopausal Breast Cancer Is Associated with Exposure to Traffic-Related Air Pollution in Montreal, Canada: A Case Control Study. Environmental Health Perspectives. 2010;118(11):1578-83.

95. Goldberg MS, Labrèche F, Weichenthal S, Lavigne E, Valois M-F, Hatzopoulou M, et al. The association between the incidence of postmenopausal breast cancer and concentrations at street-level of nitrogen dioxide and ultrafine particles. Environmental Research. 2017;158:7-15.

96. Parent M-É, Goldberg MS, Crouse DL, Ross NA, Chen H, Valois M-F, et al. Traffic-related air pollution and prostate cancer risk: a case–control study in Montreal, Canada. Occupational and Environmental Medicine. 2013;70(7):511-8.

97. Statistics Canada. Census Profile, 2016 Census 2019. Available from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CMACA&Code1=462&Geo2=PR&Code2=01&Data=Count&SearchText=Montreal&SearchType=Begins&SearchPR=01&TABID=1&B1=All.

98. Government of Canada. Canadian Climate Normals 1981-2010 Station Data 2021. Available from https://climate.weather.gc.ca/climate_normals/results_1981_2010_e.html?searchType=stnProv&lstProvince=QC&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&stnID=5415&dispBack=0.

99. Environmental Assessment Report Air Quality Montreal. Ville de Montréal, Service de l'environnement, Division de la planification et du suivi environnemental, Réseau de surveillance de la qualité de l'air (RSQA); 2018.

100. Boulet D, Melançon S. Environmental Assessment Report Air Quality Montreal. Ville de Montréal, Service des infrastructures, du transport et de l'environnement, Direction de l'environnement, Division de la planification et du suivi environnemental, RSQA; 2012.

101. Airzone. Available from https://www.airzoneone.com/lab-analysis/

102. DMTI Spatial. CanMap® GIS Data for GIS Mapping Software. Available from https://www.dmtispatial.com/canmap/

103. Xie X, Semanjski I, Gautama S, Tsiligianni E, Deligiannis N, Rajan R, et al. A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods. ISPRS International Journal of Geo-Information. 2017;6(12).

104. Ramos Y, Requia WJ, St-Onge B, Blanchet JP, Kestens Y, Smargiassi A. Spatial modeling of daily concentrations of ground-level ozone in Montreal, Canada: A comparison of geostatistical approaches. Environ Res. 2018;166:487-96.

105. PTV Vissim. Traffic Simulation Software: PTV Group. Available from https://www.ptvgroup.com/en/solutions/products/ptv-vissim/

106. United States Environmental Protection Agency. MOVES and Other Mobile Source Emissions Models. Available from http://www.epa.gov/moves.

107. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning : with applications in R: New York : Springer, [2013] ©2013; 2013.

108. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing. 2014;24(6):997-1016.

109. Gu P, Dallmann TR, Li HZ, Tan Y, Presto AA. Quantifying Urban Spatial Variations of Anthropogenic VOC Concentrations and Source Contributions with a Mobile Sampling Platform. Int J Environ Res Public Health. 2019;16(9).

110. Yang Y, Liu X, Zheng J, Tan Q, Feng M, Qu Y, et al. Characteristics of one-year observation of VOCs, NOx, and O3 at an urban site in Wuhan, China. J Environ Sci (China). 2019;79:297-310.

111. Shields HC, Weschler CJ. Analysis of Ambient Concentrations of Organic Vapors with a Passive Sampler. JAPCA. 1987;37(9):1039-45.

112. Stock TH, Morandi MT, Afshar M, Chung KC. Evaluation of the Use of Diffusive Air Samplers for Determining Temporal and Spatial Variation of Volatile Organic Compounds in the Ambient Air of Urban Communities. Journal of the Air & Waste Management Association. 2008;58(10):1303-10.

113. Senechal H, Visez N, Charpin D, Shahali Y, Peltre G, Biolley JP, et al. A Review of the Effects of Major Atmospheric Pollutants on Pollen Grains, Pollen Content, and Allergenicity. ScientificWorldJournal. 2015;2015:940243.

114. La Rosa M, Lionetti E, Reibaldi M, Russo A, Longo A, Leonardi S, et al. Allergic conjunctivitis: a comprehensive review of the literature. Italian Journal of Pediatrics. 2013;39(18).

115. Linneberg A, Henrik Nielsen N, Frølund L, Madsen F, Dirksen A, Jørgensen T. The link between allergic rhinitis and allergic asthma: A prospective population-based study. The Copenhagen Allergy Study. Allergy. 2002;57:4.

116. von Mutius E, Weiland SK, Fritzsch C, Duhme H, Keil U. Increasing prevalence of hay fever and atopy among children in Leipzig, East Germany. The Lancet. 1998;351(9106):862-6.

117. Sly M. Changing prevalence of allergic rhinitis and asthma. Ann Allergy Asthma Immunol 1999;82:19.

118. Report from the Canadian Chronic Disease Surveillance System: Asthma and Chronic Obstructive Pulmonary Disease (COPD) in Canada. Public Health Agency of Canada; 2018.

119. Dales RE, Cakmak S, Judek S, Dann T, Coates F, Brook JR, et al. Influence of outdoor aeroallergens on hospitalization for asthma in Canada. J Allergy Clin Immunol. 2004;113(2):303-6.

120. Cakmak S, Dales RE, Coates F. Does air pollution increase the effect of aeroallergens on hospitalization for asthma? J Allergy Clin Immunol. 2012;129(1):228-31.

121. Sun X, Waller A, Yeatts KB, Thie L. Pollen concentration and asthma exacerbations in Wake County, North Carolina, 2006-2012. Sci Total Environ. 2016;544:185-91.

122. Darrow LA, Hess J, Rogers CA, Tolbert PE, Klein M, Sarnat SE. Ambient pollen concentrations and emergency department visits for asthma and wheeze. J Allergy Clin Immunol. 2012;130(3):630-8 e4.

123. Osborne NJ, Alcock I, Wheeler BW, Hajat S, Sarran C, Clewlow Y, et al. Pollen exposure and hospitalization due to asthma exacerbations: daily time series in a European city. Int J Biometeorol. 2017;61(10):1837-48.

124. Erbas B, Jazayeri M, Lambert KA, Katelaris CH, Prendergast LA, Tham R, et al. Outdoor pollen is a trigger of child and adolescent asthma emergency department presentations: A systematic review and meta-analysis. Allergy. 2018;73(8):1632-41.

125. Ziska LH, Beggs PJ. Anthropogenic climate change and allergen exposure: The role of plant biology. J Allergy Clin Immunol. 2012;129(1):27-32.

126. D'Amato G, Holgate ST, Pawankar R, Ledford DK, Cecchi L, Al-Ahmad M, et al. Meteorological conditions, climate change, new emerging factors, and asthma and related allergic disorders. A statement of the World Allergy Organization. World Allergy Organ J. 2015;8(1):25.

127. Zhang Y, Bielory L, Mi Z, Cai T, Robock A, Georgopoulos P. Allergenic pollen season variations in the past two decades under changing climate in the United States. Glob Chang Biol. 2015;21(4):1581-9.

128. Ziska L, Knowlton K, Rogers C, Dalan D, Tierney N, Elder MA, et al. Recent warming by latitude associated with increased length of ragweed pollen season in central North America. Proc Natl Acad Sci U S A. 2011;108(10):4248-51.

129. Ziska LH, Gebhard DE, Frenz DA, Faulkner S, Singer BD, Straka JG. Cities as harbingers of climate change: common ragweed, urbanization, and public health. J Allergy Clin Immunol. 2003;111(2):290-5.

130. Schmidt CW. Pollen Overload: Seasonal Allergies in a Changing Climate. Environ Health Perspect. 2016;124(4):A70-5.

131. Skjøth CA, Ørby PV, Becker T, Geels C, Schlünssen V, Sigsgaard T, et al. Identifying urban sources as cause of elevated grass pollen concentrations using GIS and remote sensing. Biogeosciences. 2013;10(1):541-54.

132. Devadas R, Huete AR, Vicendese D, Erbas B, Beggs PJ, Medek D, et al. Dynamic ecological observations from satellites inform aerobiology of allergenic grass pollen. Sci Total Environ. 2018;633:441-51.

133. Wopfner N, Gadermaier G, Egger M, Asero R, Ebner C, Jahn-Schmid B, et al. The spectrum of allergens in ragweed and mugwort pollen. Int Arch Allergy Immunol. 2005;138(4):337-46.

134. Aerobiology Research Laboratories. GRIPST 2009 Rotation Impaction Sampler Manual

135. Reich BJ, Ghosh SK. Bayesian Statistical Methods. New York Chapman and Hall/CRC 2019.

136. Team RC. R: A Language and Environment for Statistical Computing. 3.6.3 ed. Vienna, Austria: R Foundation for Statistical Computing; 2020.

137. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of Statistical Software. 2017;76(1).

138. Stan Development Team. RStan: the R interface to Stan. R package version 2.17.3. 2018.

139. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016.

140. Liu Q, Huang C, Li H. Mapping plant communities within quasi-circular vegetation patches using tasseled cap brightness, greenness, and topsoil grain size index derived from GF-1 imagery. Earth Science Informatics. 2021;14(2):975-84.

141. Dymond CC, Mladenoff DJ, Radeloff VC. Phenological differences in Tasseled Cap indices improve deciduous forest classification. Remote Sensing of Environment. 2002;80(3):460-72.

142. Eisenman TS, Churkina G, Jariwala SP, Kumar P, Lovasi GS, Pataki DE, et al. Urban trees, air quality, and asthma: An interdisciplinary review. Landscape and Urban Planning. 2019;187:47-59.

143. Berrocal VJ, Gelfand AE, Holland DM. A spatio-temporal downscaler for output from numerical models. Journal of agricultural, biological, and environmental statistics. 2010;15(2):176-97.

144. Berrocal VJ, Gelfand AE, Holland DM. A bivariate space-time downscaler under space and time misalignment. The annals of applied statistics. 2010;4(4):1942.

145. Lawson AB, Choi J, Cai B, Hossain M, Kirby RS, Liu J. Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data. Journal of agricultural, biological, and environmental statistics. 2012;17(3):417-41.

146. Amemiya T, Wu RY. The effect of aggregation on prediction in the autoregressive model. Journal of the American Statistical Association. 1972;67(339):628-32.

147. Schmidt AM, Gamerman D. Temporal aggregation in dynamic linear models. Journal of Forecasting. 1997;16(5):293-310.

148. Ferreira MA, Higdon DM, Lee HK, West M. Multi-scale and hidden resolution time series models. Bayesian Analysis. 2006;1(4):947-67.

149. Ferreira MAR, Lee HK. Multiscale modeling: a Bayesian perspective: Springer; 2007.

150. Holan SH, Toth D, Ferreira MA, Karr AF. Bayesian multiscale multiple imputation with implications for data confidentiality. Journal of the American Statistical Association. 2010;105(490):564-77.

151. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian analysis. 2006;1(3):515-34.

152. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

153. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006;6(1):4.

154. Gabry J, Mahr T. bayesplot: Plotting for Bayesian Models. 2021.

155. Zapata-Marin S, Schmidt AM, Weichenthal S, Katz DS, Takaro T, Brook J, et al. Within city spatiotemporal variation of pollen concentration in the city of Toronto, Canada. Environmental research. 2022;206:112566.

156. McDonald JE. Collection and washout of airborne pollens and spores by raindrops. Science. 1962;135(3502):435-7.

157. Pehkonen E, Rantio-Lehtimäki A. Variations in airborne pollen antigenic particles caused by meteorologic factors. Allergy. 1994;49(6):472-7.

158. Weber RW. Meteorologic variables in aerobiology. Immunology and Allergy Clinics. 2003;23(3):411-22.