Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No 151
DOI: 10.11159/icsta22.151

Development of a Multi Pollutant Model to Assess Air Pollution Association with Human Health Effects

Shannon Jarvis, Wesley S. Burr
Trent University
1600 W Bank Dr., Peterborough, Canada
shannonjarvis@trentu.ca; wesleyburr@trentu.ca

Abstract **-**A number of methodologies have been developed for investigation of the association between human health and exposure to a single pollutant [e.g., 1]. However, as pollutants are correlated, and the joint effect of pollutants is of high interest, work continues on development for multiple pollutant models. In this work, we discuss a method using Thin Plate Splines (TPS) to simultaneously model both $PM_{2.5}$ (particulate matter less than 2.5 μm in aerodynamic diameter) and O_3 (ozone) in association with human mortality. The results are compared to effect estimates obtained from single pollutant models. We find similar temporal trends in the estimates, with large movements in both $PM_{2.5}$ and O_3 being captured in the TPS estimates. The estimated errors for the TPS method are larger than the individual models combined and produce risks that are comparable but slightly elevated.

Keywords**:** Environmental Epidemiology; Air Pollution; Human Health; Generalized Additive Models; Multivariate Models; Thin Plate Splines

1. Introduction

Particulate matter (PM) has become one of the primary pollutants of interest for global health studies due to its evidenced relationship with human health effects (e.g., [2, 3]). In particular, evidence of association due to both short (acute) and long-term exposures of ambient air pollution on cardiovascular and respiratory health impacts has been examined and found to be significant [1]. More recently, studies examining human health effects associated with multiple pollutants simultaneously has been investigated, as a more realistic framework for actual exposure. Several recent attempts (and successes) include [4-6], where the two project teams explored models, mostly in Bayesian frameworks, for assessing the statistical effect of multiple air pollution constituents and unknown numbers of major sources. In this work, we explore the development of a multiple pollutant method that includes multiple, possibly correlated, pollutants (in this case, particulate matter with a diameter less than 2.5 μm ($PM_{2.5}$) and ozone (O_3)) with the use of a response surface, and thin plate splines.

2. Data

In Canada, the National Air Pollution Surveillance (NAPS) database, organized and gathered by Environment and Climate Change Canada [7] contains, among many pollutant measures, airborne particulate matter from a large number of stations. Pollutants are measured hourly by the instrumentation. For these acute risk studies, their concentration is aggregated to 24 hour mean concentration by geographic region, called a census division (CD), to act as a ecological proxy for the population-level exposure. In this process, missing days are imputed. For this study, measurements of $PM_{2.5}$ and O_3 across 53 census divisions are obtained. In addition to pollutant concentrations, daily temperatures measured by the Meteorological Service of Canada climate database [8] were obtained. The daily temperatures were aggregated to a daily measurement by census division, across multiple observing stations, as available. Imputation of missing data was not performed.

The $PM_{2.5}$, O_3 and temperatures for Toronto, ON, a large urban area with a population of almost 3 million, is shown in Figure 1. There are seasonal ozone and temperature patterns, where in the warmer season (May – October) ozone concentrations are higher. Patterns can also be seen for the $PM_{2.5}$ concentrations. An increase of concentrations in warmer seasons is a consequence of atmospheric chemical formation, where O_3 forms from nitrogen oxides via solar radiation, a process that largely occurs above 17°C [1]. These patterns, and hence correlation, between pollutants compel our exploration into the development of a multi pollutant model that controls for confounding. They are seen across all census divisions in the study. Note that a naive Pearson correlation between the two gives mild correlation only, but careful examination of the time series structure makes it clear that there are deeper relationships – these pollutants cannot be considered independent.

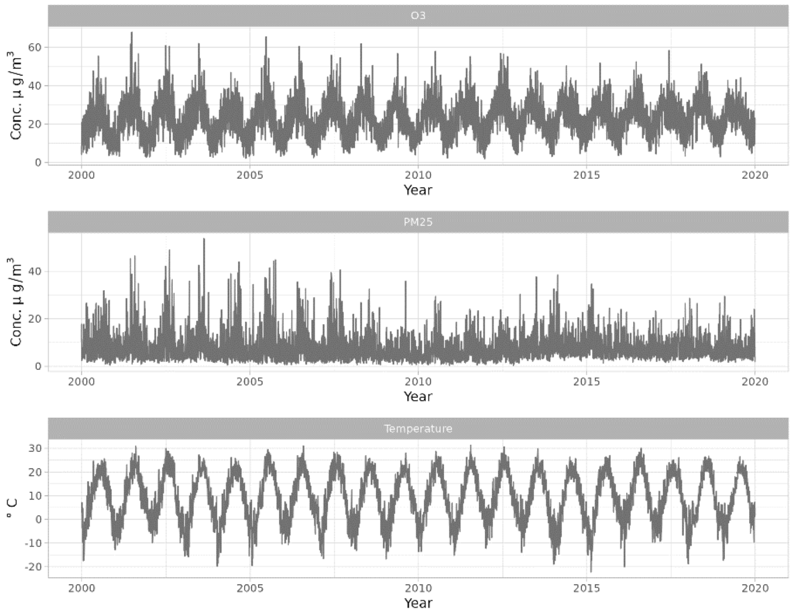

Figure 1: Daily pollutant and temperature measurements for Toronto, ON census division.

3. Single Pollutant Model

We will begin by exploring the single pollutant model, as developed by Dominici et al. originally [9]. In this model, Generalized Additive Models (GAMs) are used to estimate the annual risk. As counts, the natural family is Poisson, so the logarithm of the daily human mortality (health effect) are modelled by the pollutant of interest, a smooth function of time (Discrete Prolate Spheroidal (Slepian) Sequences with 6 or 12 degrees of freedom per year, [10]), temperature and the day of week. This can be formally written as:

$$\boldsymbol{log(\mu)} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} * \textbf{Pollutant} + \boldsymbol{s_1}(\textbf{Time, df} = \textbf{6 or 12/year}) + \boldsymbol{s_2}(\textbf{Temperature}) + \textbf{DOW} \tag{1}$$

where μ is the mean response, β_0 and β_1 are linear coefficients, s_1 and s_2 are smooth functions of time and temperature, DOW is the day of week and pollutant is the pollutant concentration of interest. A cut-off of 50% is used as the threshold for missing data (for both daily observations of pollutant concentrations and daily morbidity or mortality counts). For these models, we use an identity function for s_2, as in [11]. In both these models, and the multiple pollutant models that follow, there is a choice of seasonality possible: that is, are all days used in the estimate (annual estimate, so January – December), or are the days restricted to the "cold" or "warm" seasons prevalent in the Canadian climate (October – March or April – September).

4. Multiple Pollutant Model

4.1. Brief Introduction to Thin Plate Splines (TPS)

A method to model multiple correlated pollutants (in this case, $PM_{2.5}$ and O_3) was developed using response surface and thin plate spline bases. Thin plate splines [12] are a smooth function of one or more predictor variables, such that the response surface of the spline models the combined effect of multiple predictors. They do not require prior knowledge about the relationship between predictors or specification of knot locations. This flexibility motivates its use in modelling confounding pollutants.

4.2. TPS Model

To incorporate multiple pollutants, the single pollutant model is adjusted to incorporate a TPS and is written as:

$$\boldsymbol{log}(\boldsymbol{\mu}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} * \boldsymbol{s_{TPS}}(\textbf{Pollutant}_\mathbf{1}, \textbf{Pollutant}_\mathbf{2} + \boldsymbol{s_1}(\textbf{Time, df} = \textbf{6 or 12/year}) + \boldsymbol{s_2}(\textbf{Temperature}) + \textbf{DOW} \tag{2}$$

where s_{TPS} is the thin plate spline, and the smooth functions are the same as Eqn. (1). In this approach, the location on the response surface (after accounting for the temperature and a smooth function of time) that corresponds to the average of a particular pollutant concentration, and this location (average) plus one unit (e.g., ppb) can be determined. In traditional models, the result of interest for these models (e.g., a β_1) is geometrically interpreted as a slope, so we replicate this in the context of the bivariate surface as one of three quantities of interest: the slope of each pollutant at the location of the bivariate average, in the direction of a unit increase in that pollutant; and the slope of the pollutants jointly in the direction of their joint unit increase (taken to be equal contribution, so an increase of 1 unit in a bisecting angle to the two pollutant axes). This approach has the advantage that the interpretation stays the same as the univariate models, allowing easy comparison of the magnitudes and temporal trends.

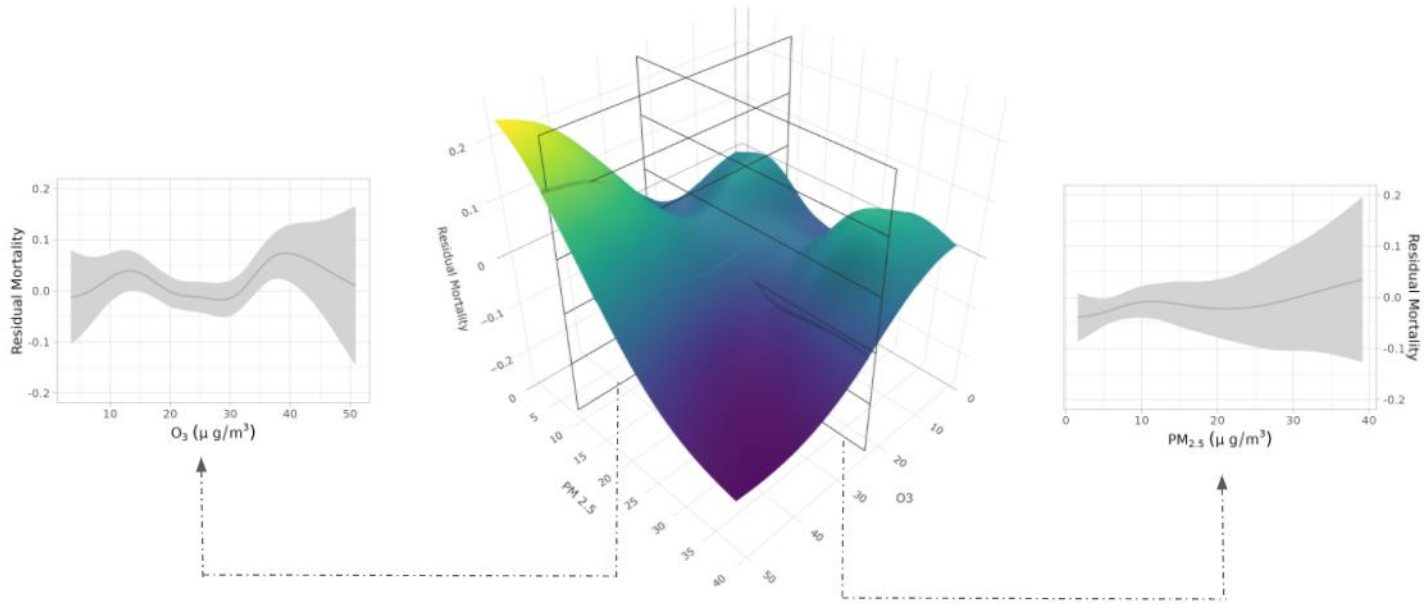

Figure 2: Thin Plate Spline residual mortality and cross-sectional plots at the average pollutant concentrations. Each of the single pollutant plots shows the surface and the prediction interval along that pollutant's axis, with the other pollutant fixed at its arithmetic overall average value. The response visualized here is the residual log mortality after accounting for temperature, DOW and the smooth function of time.

The thin plate spline surface of the residual mortality is visualized in Figure 2. Grids in the centre plot are used to show the cross sections (planes corresponding to the average $PM_{2.5}$ and O_3 concentrations). In Figure 2, left, the residual mortality TPS is plotted against the O_3 concentration while the $PM_{2.5}$ concentration is held at its mean value, 23.49 μg m-3, while the right shows the TPS residual response surface versus $PM_{2.5}$ concentration while the O_3 concentration is held at its mean value of 9.12 μg m-3.

4.3. Comparison of single and multiple pollutant model

The cardiovascular and respiratory mortality estimates from the thin plate spline (TPS) approach were compared to the single pollutant models (Figures 3-6) from 1997 to 2019 for Toronto, Vancouver, Edmonton and Ottawa, all million-plus population urban areas in Canada. For the univariate estimates, the smooth function of time was a DPSS (Slepian) spline with 12 degrees of freedom per year. In general, large increases and decreases in $PM_{2.5}$ and O_3 are captured in the TPS estimates. The TPS approach has significantly larger prediction error than either of the two individual models, or their additive combination. This is not entirely surprising, as the correlation structure of the two pollutants is known to vary over time due to component-level correlations between the respective elements of the series. In particular, the cross-sectional location of the prediction estimate varies year-to-year, as the annual averages of the $PM_{2.5}$ and O_3 vary. Control of this error in a more coherent way is a current open question.

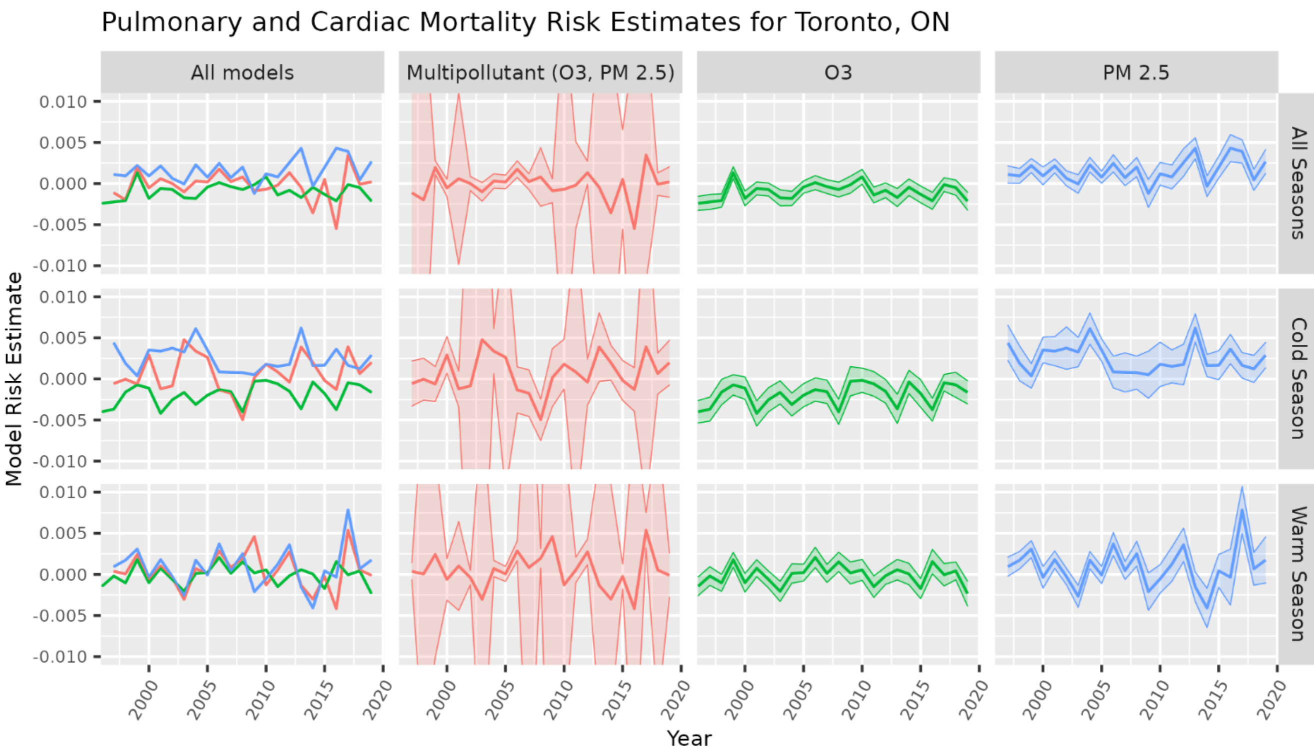

Figure 3: Mortality risk estimates from single and multiple pollutant model estimates Toronto, ON for: all seasons (January – December), cold season (October – March) and warm season (April – September).

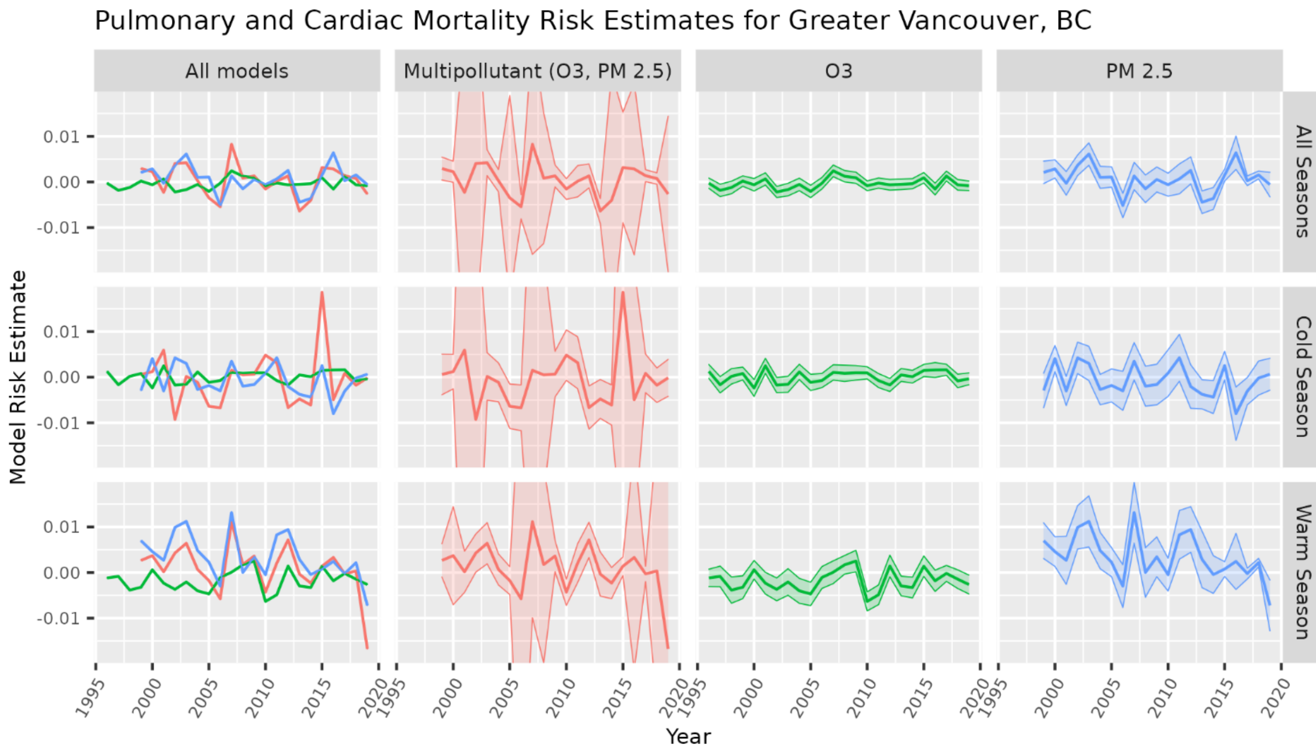

Figure 4: Mortality risk estimates from single and multiple pollutant model estimates Greater Vancouver, BC for: all seasons (January – December), cold season (October – March) and warm season (April – September).

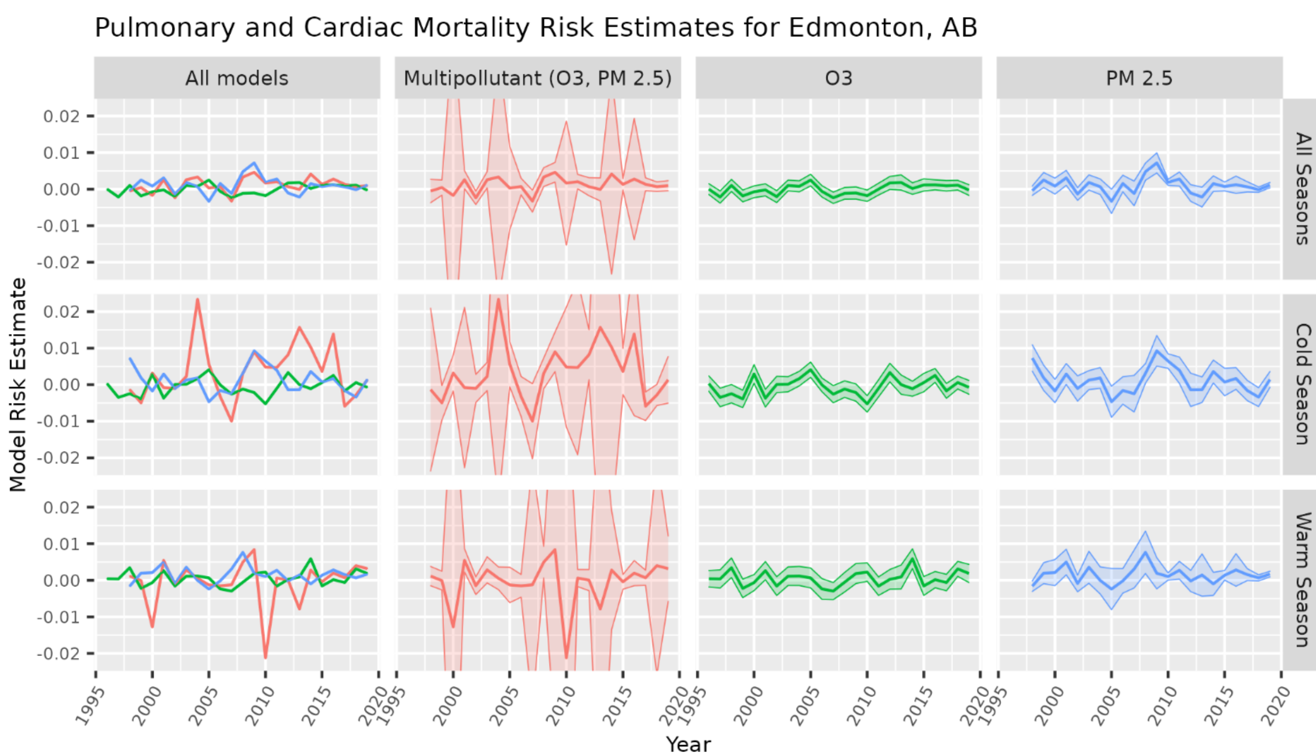

Figure 5: Mortality risk estimates from single and multiple pollutant model estimates Edmonton, AB for: all seasons (January – December), cold season (October – March) and warm season (April – September).

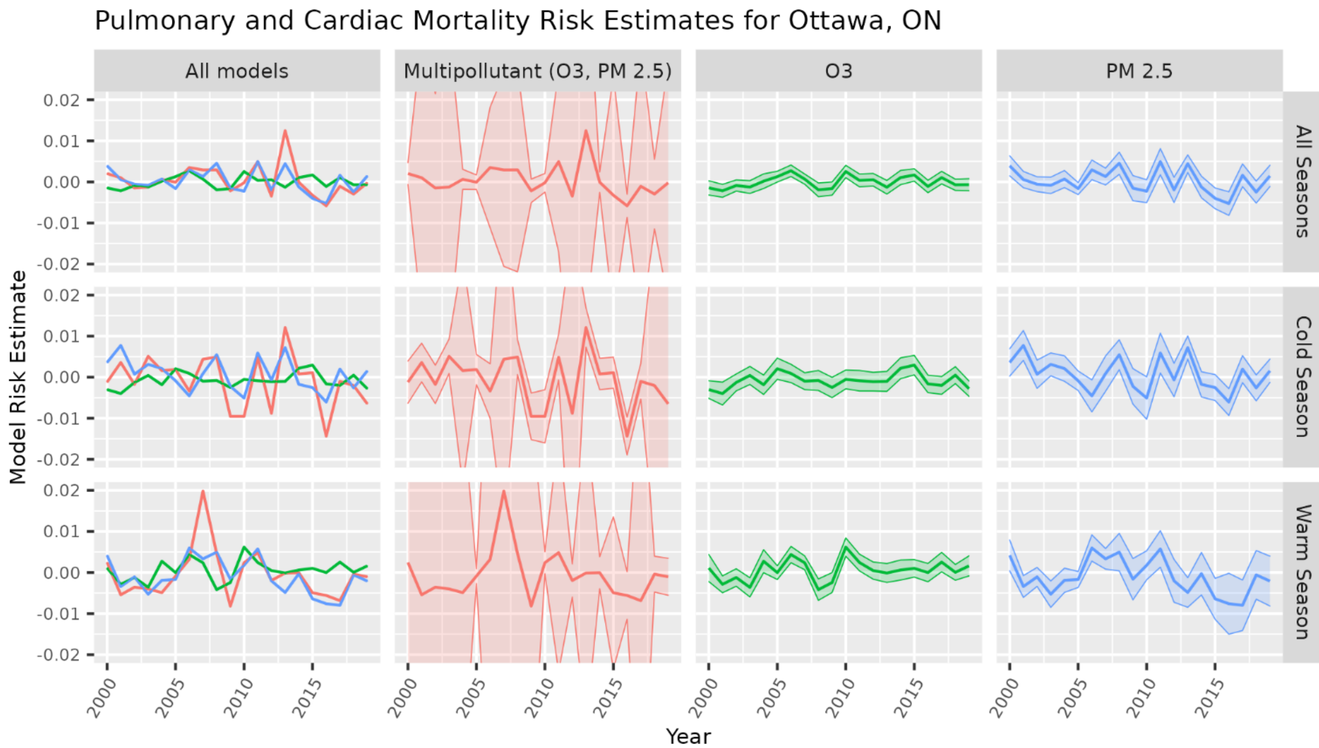

Figure 6: Mortality risk estimates from single and multiple pollutant model estimates Ottawa, ON for: all seasons (January – December), cold season (October – March) and warm season (April – September).

4. Conclusion

Particulate matter is one of the primary pollutants of interest for global health studies, and estimation of multiple pollutants simultaneously, with one of these being PM, is a problem of high research interest. In this work we have begun development of a multiple air pollutant model that builds a multiple dimension surface for the association between the pollutants simultaneously and a human health effect response. The mean association, jointly, between the two pollutants is quite similar to the average of two univariate models, but is not identical, demonstrating that modeling both simultaneously does have some value. The use of thin-plate splines is quite flexible, and could allow for further work to more carefully model the correlation and association structure of the pollutants, rather than treating them as simple bivariate contributors. The extraction of univariate 'slices' from the bivariate surface is also a technique that demonstrates promise, as these extracted slices are comparable in form to non-linear modeling outputs for the impact of air pollution on human health. In conclusion, this new method shows promise, and allows a new viewpoint on a modeling problem which is of high interest to a broad global community.

Acknowledgements

We would like to thank Hwashin H. Shin and Health Canada for access to the health effects data, and valued feedback on this work.

References

[1] World Health Organization. *Review of evidence on health aspects of air pollution: REVIHAAP project: technical report*. No. WHO/EURO: 2013-2663-42419-58845. World Health Organization. Regional Office for Europe, 2021.

[2] Ghio, A. J. and Huang, Y.-C. T. (2004) Exposure to concentrated ambient particles (CAPs): A Review. *Inhalation Toxicology*, vol. 16, no. 1, pp. 53–59.

[3] Franklin, B. A., Brook, R. and Pope III, C.A. (2015) Air pollution and cardiovascular disease. *Current Problems in Cardiology*, vol. 40, no. 5, pp. 207–238.
[4] Bobb, J. F., Valeri, L., Claus Henn, B., …, & Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3), 493-508.
[5] Park, E. S., & Oh, M. S. (2018). Accounting for uncertainty in source-specific exposures in the evaluation of health effects of pollution sources on daily cause-specific mortality. *Environmetrics*, 29(1), e2484.
[6] Hassan Bhat, T., Jiawen, G., & Farzaneh, H. (2021). Air pollution health risk assessment (AP-HRA), principles and applications. *International Journal of Environmental Research and Public Health*, 18(4), 1935.
[7] Environment and Climate Change Canada. (2022) *National Air Pollution Surveillance Program*. Open Government License, https://open.canada.ca/data/en/dataset/1b36a356-defd-4813-acea-47bc3abd859b.
[8] Government of Canada. (2022) *Historical Climate Data*. Open Government License, https://climate.weather.gc.ca/.
[9] Dominici, F., Samet, J. M., & Zeger, S. L. (2000). Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(3), 263-302.
[10] Burr, W. S., Takahara, G., & Shin, H. H. (2015). Bias correction in estimation of public health risk attributable to short-term air pollution exposure. *Environmetrics*, *26*(4), 298-311.
[11] Burr, W. S., & Shin, H. H. (2015). Accounting for Temperature when Modeling Population Health Risk Due to Air Pollution. In *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science* (pp. 105-112). Springer, Cham.
[12] Duchon, J. (1977) Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive Theory of Functions of Several Variables*, pp. 85–100.