Modeling Linear Combinations of Multiple Pollutants for Health Outcomes

Name and affiliations of lead investigators

Name: Professor Patrick Brown
Affiliations: Centre for Global Health Research, St. Michael's Hospital
Department of Statistical Sciences, University of Toronto

Name: Professor Fateh Chebana
Affiliation: Institut national de la recherche scientifique

Name: Professor Cindy Feng
Affiliation: School of Epidemiology and Public Health, University of Ottawa

List of proposed collaborators, titles, and affiliations

Name: Kamal Rai
Title: PhD
Affiliations: Centre for Global Health Research, St. Michael's Hospital
Department of Statistical Sciences, University of Toronto

Name: Hwashin Shin
Title: Adjunct Associate Professor (?)
Affiliations: Environmental Health Science and Research Bureau, Health Canada
Department of Mathematics and Statistics, Queen's University

Name: Céline Campagna
Title: PhD
Affiliation: Institut National de Santé Publique du Québec

Name: Pierre Masselot
Title: Research Fellow
Affiliation: London School of Hygiene & Tropical Medicine

List of potential partner organizations

Centre for Global Health Research, St. Michael's Hospital

Institut National de Santé Publique du Québec

Health Canada

Research Aims

There is growing interest in developing a simple, intuitive air quality index that simultaneously accounts for the health effects of multiple air pollutants (Bopp *et al.*, 2018; Dominici *et al.*, 2010; Stieb *et al.*, 2008). There is a growing understanding that the health effects of air pollution depends on the composition of the mixture of pollutants in the air, as opposed to high levels of a single pollutant (Dominici *et al.*, 2010). An air quality index that reflects this understanding should account for the various levels and relative contributions of each air pollutant in the ambient air. In this proposal, we will introduce a statistical model designed to conduct inference on the health effects of simultaneous exposure to multiple environmental pollutants, with a focus on applications related to the health effects of air pollution.

The *constrained groupwise additive index model* (cGAIM) is introduced by Masselot *et al.* (2020), who develops frequentist inference methods for it that use sequential quadratic programming. For a response distribution D, parameter $\theta = (\theta_1, \ldots, \theta_d)$, link function g, and constraints as below, the cGAIM is,

$$\begin{aligned} &Y_t|\lambda_t \overset{i.i.d.}{\sim} D(\lambda_t), \\ &g(\lambda_t) = X\beta + s(\alpha^T Z) + f_1(\eta_1) + \ldots + f_k(\eta_k). \end{aligned}$$

Here, Y_t is the outcome of interest, β is the vector of fixed effects, and $f_1, \ldots, f_K$ are smoothing functions that account for potential confounding variables. The novelty of the cGAIM lies in s, a smooth function fit to a linear combination of environmental covariates Z (such as air pollutants). The α are parameters of the cGAIM that we wish to estimate – α can be constrained to have non-negative components that sum to one, so that it is a vector of weights on the entries of Z. Once estimated, these weights give the relative contribution of each component of Z to $\alpha^T Z$. In addition to constraining α, the cGAIM can also constrain the shape of the smooth function s by, for example, requiring it to be monotonic or convex.

We will develop a Bayesian implementation of the cGAIM – the bcGAIM. Compared to the cGAIM, the bcGAIM provides two main statistical benefits. The first is credible intervals for α; the cGAIM does not provide confidence intervals for α. Quantifying the uncertainty of the estimates for α is important for a very important scientific question – can the bcGAIM determine how much specific pollutants contribute to polluted air? Narrow credible intervals indicate that it can attribute the health effects to specific pollutants. Wide credible intervals indicate that it can identify the health effects but not attribute them to specific pollutants. Either result would be a significant contribution to the development of multiple pollutant models, as it would provide evidence on if the health effects of air pollution can be associated to particular pollutants (through α) or only to a mixture of pollutants (through $s(\alpha^T Z)$). The second advantage is that the bcGAIM can identify and address multi-modality in the posterior of α. If α is multi-modal under an uninformative prior, one can use relevant scientific knowledge to place stronger priors on α. In comparison, the cGAIM only provides a point estimate for α and cannot detect multi-modality.

We will also explore using different response distributions. The standard single pollutant model is the log-linear Poisson model (Dominici *et al.*, 2002; Liu *et al.*, 2019).[1] The case crossover has seen increased attention in the air pollution literature (Stringer *et al.*, 2020; Wei *et al.*, 2019). It can be viewed as a conditional Poisson model, obtained by stratifying by subject and conditioning on the number of events

[1] Let the average rate the outcome occurs on day t be denoted λ_t, and the single pollutant in the model be denoted P_t. The one-pollutant model is a special case of the cGAIM with a Poisson response distribution – $g(\lambda_t) = \log(\lambda_t)$ is the link function, $Z = P_t$ is the covariate of interest, $s(\alpha P_t) = \alpha P_t$ is the identity function, and there are no constraints.

Y_t occurring in an observation period. Additional details on the differences between these two models is given in the Methods section.

Objectives

This proposal has three main research objectives. The first is in developing an air quality index for Canadian air pollution that accounts for the health effects of a mixture of air pollutants. This index will be shared with Health Canada and INSQ. The second will be applications in other epidemiological studies that involve exposures to mixtures of environmental pollutants. Other health applications where mixtures of exposures have garnered some interest include chemical mixtures, metal mixtures, and pesticides (Bobb *et al.*, 2015; Braun *et al.*, 2016; Lazarevic *et al.*, 2019; Sanders *et al.*, 2015), and these applications will be actively explored with the Centre of Global Health Research during the course of this project. The third application is a specific application of the air quality index, namely investigating how long-term exposure to different mixtures of pollutants affects COVID-19 mortality.

Inter-disciplinary research teams and stakeholders conduct air pollution research. It is therefore crucial that air pollution models have interpretable parameters, so that estimation results (and their associated uncertainty) can be easily communicated to non-specialists. There is evidence that the health effects the single pollutant model attributes to P_t may be caused by a correlated pollutant not in the model. For example, Franklin and Schwartz (2008) found that the effect of ozone on non-accidental mortality was "substantially reduced" after adjusting for sulfate and Liu *et al.* (2019) found significant differences in the percentage change of all-cause mortality attributable to $PM_{2.5}$ and PM_{10} after adjusting for NO_2 or SO_2.

Many modeling approaches have been applied to the multiple pollutant model. Supervised methods that consider mixtures of pollutants include Pachon *et al.* (2012), who specifies weights from data rather than estimating them, and Roberts and Martin (2006), who introduce a model that is equivalent to assuming that s is linear. However, there is evidence that some health outcomes are nonlinearly related to health outcomes levels (Feng *et al.*, 2016; Xia and Tong, 2006). Compared to these models, the bcGAIM is more flexible. It estimates its weights from data, and specifies a nonlinear relationship between the mixture of pollutants and the health outcome. Meanwhile, the results of unsupervised methods (such as principle components analysis and clustering) are difficult to interpret (Davalos *et al.*, 2017). A popular nonparameteric method is the Bayesian Kernel Machine Regression (BKMR), introduced in Bobb *et al.* (2015). The BKMR models an exposure-response surface via a kernel function. Using a hierarchical Bayesian variable selection method, it can select one pollutant from a group of correlated ones, and is interpreted by visualizing cross-sections of a potentially high-dimensional exposure-response surface. Unlike the bcGAIM, the BKMR does not have easily interpretable parameters. This makes the bcGAIM more suitable to the communication needs of an inter-disciplinary research team – it has interpretable parameters (α) that enable easily communicable results regarding the health effects of multiple air pollutants.

ADD: Global Health paragraph

The relationship between daily COVID-19 deaths and air pollution levels has become an active area of research. For instance, Wu *et al.* (2020) apply a zero-inflated negative binomial to model U.S. data, where the zero-inflation accounts for counties with no COVID-19 deaths. They find that a 1 μg increase in long-term exposure to ambient $PM_{2.5}$ increases the COVID-19 death rate by 15%. Additional studies that examine this relationship include Conticini *et al.* (2020), Sciomer *et al.* (2020), and Setti *et al.* (2020).

However, much work remains to be done. For example, we do not have an accurate measure of excess deaths attributable to COVID-19, due to potential under-reporting of COVID-19 case and death counts. Moreover, cumulative COVID-19 mortality will continue to rise for some time, making the question of excess deaths best suited to an ongoing inquiry. We will use the bcGAIM to examine the relationship between COVID-19 deaths and long-term exposure to air pollution. Compared to the log-linear negative binomial model in Wu *et al.* (2020), the GAIM is scalable, interpretable, and can capture non-linearities in the relationship between mixtures and the response. Applying the bcGAIM to air pollution and COVID-19 death data will allow us to ascertain which mixtures of pollutants increase COVID-19 mortality.

Therefore, this inquiry will help determine which mixtures of air pollutants most increase COVID-19 deaths, and help identify how the effect of air pollution differs among age groups. In particular, COVID-19 and air pollution both disproportionately affect vulnerable populations such as seniors. Determining how much of this is attributable to air pollution would be a significant contribution that will further our understanding of the relationship between COVID-19 deaths and air pollution.

Methods

The bcGAIM will make four major contributions to modeling the health effects of mixtures of pollutants.

1. Extending the cGAIM to higher dimensional problems.
2. Fully exploring the parameter space using posterior distributions, particularly for α.
3. Developing priors for shape-constrained Bayesian inference on the smooth function s.
4. Comparing estimation results under the log-linear Poisson and case crossover models.

In Masselot *et al.* (2020), the cGAIM uses an iterative two-step optimization scheme – α is updated using a quadratic program, then s is updated using the shape-constrained additive model methodology of Pya and Wood (2015). The bcGAIM will initially be implemented in Stan, a statistical modeling language that performs optimization using Hamiltonian Monte Carlo (Carpenter *et al.*, 2017). The Stan modeling language makes it straightforward to extend the bcGAIM to include additional pollutants, additional lags for pollutants in the model, and additional smooth functions s. Up to having suitable priors, all 3 extensions only require adding terms to a sum in Stan. Following its implementation in Stan, the bcGAIM will be implemented using non-MCMC inference methods, similar to Iterated Nested Laplace Approximation (INLA) (Rue *et al.*, 2009). These non-MCMC methods provide significant computational and ease-of-use benefits, and will expand the types of problems and number of users who can use the bcGAIM methodology. Finally, to facilitate use by other researchers, all bcGAIM software will be released in an R package.

Using a Bayesian approach to estimate the health effects of mixtures of pollutants has numerous benefits. Being Bayesian allows us to obtain posterior distributions and credible intervals for all model parameters. This is particularly beneficial with α, as different values of α could maximize the posterior. In other words, different combinations of pollutants could have the same effect on health outcomes. It is very important to be able to identify if these different values exist, and what those values are. The practical implication is that an air quality index must consider the value of multiple pollutants when providing guidelines. Furthermore, we could if desired address the multi-modality of α by constructing scientifically justified priors that encourage α towards one of its modes. For example, this could be done if there was evidence from the composition of each pollutant that one was more responsible for the health effects than the others.

A major task in developing the bcGAIM is in developing priors for shape-constrained Bayesian inference on s. In addition to being free of unwanted behavior, the chosen priors should be simple and interpretable so that non-statistical experts can use it. This is difficult to achieve for shape constraints. The first reason for this is that s_i may not have any parameters related to the desired constraint; for example, a 1st-order random walk has no parameters related to monotonicity. To overcome this, s could be re-parameterized, the functional form of s could be exploited, or data augmentation schemes that introduce derivative observations could be used (Riihimäki and Vehtari, 2010). The second reason is that substantial mathematical analysis is required to ensure priors do not introduce unwanted behavior. For example, a truncated multivariate normal (tMVN) prior can induce monotonicity when placed on the coefficients $\beta^{'}$ of a finite basis expansion of s (Maatouk and Bay, 2017). However, the tMVN prior places negligible mass on near-flat regions of s. While Zhou *et al.* (2020) remedy this by introducing a scale parameter to the coordinates of the tMVN distribution, the modified tMVN prior is placed on $\beta^{'}$, not s. In comparison to the tMVN prior, we seek a prior on s and not its basis expansion. Such a prior will be less encumbered by mathematical details and more easily understood by non-statistical experts. This ease of interpretation should encourage adoption of the bcGAIM among non-statistical experts, one of the goals of the bcGAIM projects.

ADD: Poisson vs. Case Crossover paragraph

References

Bobb, J. F., Valeri, L., et al. (2015) Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, **16**, 493–508.

Bopp, S., Richarz, A., et al. (2018) Something from nothing: Ensuring the safety of chemical mixtures. *Ensuring the safety of chemical mixtures, Publications Office of the European Union, EUR*, **29258**.

Braun, J. M., Gennings, C., et al. (2016) What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environmental Health Perspectives*, **124**, A6–A9.

Carpenter, B., Gelman, A., et al. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**.

Conticini, E., Frediani, B., et al. (2020) Can atmospheric pollution be considered a co-factor in extremely high level of sars-cov-2 lethality in northern italy? *Environmental Pollution*,, 114465.

Davalos, A. D., Luben, T. J., et al. (2017) Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Annals of Epidemiology*, **27**, 145–153.

Dominici, F., McDermott, A., et al. (2002) On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, **156**, 193–203.

Dominici, F., Peng, R. D., et al. (2010) Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology (Cambridge, Mass.)*, **21**, 187.

Feng, C., Li, J., et al. (2016) Impact of ambient fine particulate matter (pm 2.5) exposure on the risk of influenza-like-illness: A time-series analysis in beijing, china. *Environmental Health*, **15**, 17.

Franklin, M. and Schwartz, J. (2008) The impact of secondary particles on the association between ambient ozone and mortality. *Environmental Health Perspectives*, **116**, 453–458.

Lazarevic, N., Barnett, A. G., et al. (2019) Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: A review of existing approaches and new alternatives. *Environmental Health Perspectives*, **127**, 026001.

Liu, C., Chen, R., et al. (2019) Ambient particulate air pollution and daily mortality in 652 cities. *NEJM*, **381**, 705–715.

Maatouk, H. and Bay, X. (2017) Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, **49**, 557–582.

Masselot, P., Chebana, F., et al. (2020) Constrained groupwise additive index models. *Submitted*.

Pachon, J. E., Balachandran, S., et al. (2012) Development of outcome-based, multipollutant mobile source indicators. *Journal of the Air & Waste Management Association*, **62**, 431–442.

Pya, N. and Wood, S. N. (2015) Shape constrained additive models. *Statistics and Computing*, **25**, 543–559.

Riihimäki, J. and Vehtari, A. (2010) Gaussian processes with monotonicity information. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 645–652.

Roberts, S. and Martin, M. A. (2006) Investigating the mixture of air pollutants associated with adverse health outcomes. *Atmospheric Environment*, **40**, 984–991.

Rue, H., Martino, S., et al. (2009) Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *JRSS: Series B*, **71**, 319–392.

Sanders, A. P., Henn, B. C., et al. (2015) Perinatal and childhood exposure to cadmium, manganese, and metal mixtures and effects on cognition and behavior: A review of recent literature. *Current Environmental Health Reports*, **2**, 284–294.

Sciomer, S., Moscucci, F., et al. (2020) SARS-cov-2 spread in northern italy: What about the pollution role? *Environmental Monitoring and Assessment*, **192**, 1–3.

Setti, L., Passarini, F., et al. (2020) Searching for sars-cov-2 on particulate matter: A possible early indicator of covid-19 epidemic recurrence.

Stieb, D. M., Burnett, R. T., et al. (2008) A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. *Journal of the Air & Waste Management Association*, **58**, 435–450. Taylor & Francis.

Stringer, A., Brown, P., et al. (2020) Approximate bayesian inference for case crossover models. *Submitted*, **15**.

Wei, Y., Wang, Y., et al. (2019) Short term exposure to fine particulate matter and hospital admission risks and costs in the medicare population: Time stratified, case crossover study. *BMJ*, **367**.

Wu, X., Nethery, R. C., et al. (2020) Exposure to air pollution and covid-19 mortality in the united states. *medRxiv*.

Xia, Y. and Tong, H. (2006) Cumulative effects of air pollution on public health. *Statistics in Medicine*, **25**, 3548–3559.

Zhou, S., Ray, P., et al. (2020) On truncated multivariate normal priors in constrained parameter spaces. *arXiv preprint arXiv:2001.09391*.

Anticipated roles of trainees (students and post-doctoral fellows)

Kamal will develop the Bayesian implementation of the GAIM models in Stan. This includes exploring determining appropriate prior distributions for the weights α, developing visualizations that communicate modeling results, and assisting other project members in developing shape constraints. He will be responsible for producing paper(s) summarizing the results of this model when run on Canadian air pollution and mortality data. To facilitate team communication and cohesion, he will also split time between Toronto (at the Centre for Global Health Research) and Ottawa (at the University of Ottawa), and use the proximity of the University of Ottawa to Quebec to occasionally visit project collaborators located there.

The University of Toronto PhD student will develop non-MCMC methods to conduct inference on the GAIM, and compare its results from those obtained from the Stan implementation. The University of Laval/University of Ottawa PhD student will develop methods to conduct shape-constrained (Bayesian) inference, and examine the relationship between COVID-19 deaths and air pollution levels.

Plans for dissemination and communication

The lead investigators of this proposal have a track record of publishing research results in leading statistical and epidemiological journals, and aim to publish the results of this project in high-impact journals. The results and findings of this multiple pollutant inquiry will also be shared with Health Canada and the Institut National de Santé Publique du Québec.

Suggested reviewers

Will be contributed by Patrick.

Preliminary budget description

The CANSSI Collaborative Research Team (CRT) grant is for $180,000 over 3 years. We propose the budget:

1. $30,000/year to support a post-doctoral student.
2. $24,000/year to support two PhD students ($12,000 per student).
3. $6,000/year to support travel to/from the cities of the lead investigators – Toronto, Ottawa, and Quebec – and annual team meetings held around the Statistical Society of Canada conference.