Modeling Linear Combinations of Multiple Pollutants

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 (INRS)

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: Post-Doctoral Fellow
Affiliation: London School of Hygiene & Tropical Medicine

List of potential partner organizations (optional)

Centre for Global Health Research, St. Michael's Hospital
Institut National de Santé Publique du Québec
Health Canada

Research Aims

We will develop a fully Bayesian implementation of the *groupwise additive index model* (GAIM), a constrained frequentest version of which is introduced in Masselot *et al.* (2020). For a response distribution D, parameter $\theta = (\theta_1, \dots, \theta_d)$, and link function g, the GAIM is,

$$Y_t|\theta_t = D(\theta),$$
$$g(\theta) = X\beta + \sum_{i=1}^{I} s_i(\alpha_1\gamma_{1,t} + \dots + \alpha_{n_i}\gamma_{n_i,t}) + \sum_{k=1}^{K} f_k(\eta_k)$$

Y_t is the outcome of interest, X is a design matrix, and β is a vector of fixed effects. The second sum consists of K smoothing functions $f_1, \dots, f_K$ applied to potential confounders η_k. The novelty lies in its first sum, which fits smooth functions to a weighted sum (*mixture*) of potentially time-varying covariates.

The GAIM is a flexible and extensible statistical model with a wide variety of potential applications, as the researcher can choose the number I of smooth functions s_i and the number of covariates n_i in each s_i. Possible epidemiological applications include examining the health effects of environmental exposures, such as chemical mixtures, metal mixtures, pesticides, and mixtures of air pollutants (Bobb *et al.*, 2015; Braun *et al.*, 2016; Lazarevic *et al.*, 2019; Sanders *et al.*, 2015). For example, a possible air pollution model could have one smooth function s_1 and two covariates $\gamma_{1,t}$ and $\gamma_{2,t}$, where $\gamma_{1,t}$ is the (time-varying) concentration of ozone (O_3) and $\gamma_{2,t}$ is the (time-varying) concentration of particulate matter less than 2.5 μg ($PM_{2.5}$). In this case, the weights $\alpha = (\alpha_1, \alpha_2)$ give the relative contribution of O_3 and $PM_{2.5}$ to the health outcome of interest. Another possible model could have 3 additional pollutants – nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter less than 10 μg (PM_{10}) – such that the 5 weights α give the relative contribution of each pollutant. The models could also include interaction terms, or additional smooth functions s_i that group the pollutants in a scientifically justified way.

To explore the benefits of the GAIM, we next compare it to a standard regression model and one where the s_i have a more general form. To ease notation, let us assume there are no confounders, one smooth function s_1, and two covariates, $\gamma_{1,t}$ and $\gamma_{2,t}$, in s_1. Then, these 3 models have link functions,

$$g(\theta) = X\beta + \alpha_1\gamma_1 + \alpha_2\gamma_2 \tag{1}$$
$$g(\theta) = X\beta + s_1(\alpha_1\gamma_1 + \alpha_2\gamma_2) \tag{2}$$
$$g(\theta) = X\beta + s_1^*(\gamma_1, \gamma_2) \tag{3}$$

(1) specifies a regression model with 2 fixed effects, α_1 and α_2, (2) specifies a GAIM with 2 weights α_1 and α_2, and (3) specifies s_1^* as a 2-dimensional smoothing function. Compared to the linear regression in (1), the GAIM is more flexible. The inclusion of s_1 allows it to capture nonlinearities in the relationship between $\gamma_{1,t}$, $\gamma_{2,t}$ and the outcome Y_t. Compared to (3), the GAIM is less computationally demanding – $\alpha_1\gamma_{1,t} + \alpha_2\gamma_{2,t}$ is always 1-dimensional, and remains so regardless of how many terms are the sum. This contrasts with s_1^*, whose dimensionality scales with its number of arguments. The GAIM is also more interpretable than (3). The weights α_1 and α_2 reflect the relative contribution of $\gamma_{1,t}$ and $\gamma_{2,t}$, respectively. In comparison, $s_1^*(\gamma_1, \gamma_2)$ has no such weights, and often has no interpretable parameters.

We will develop a fully Bayesian implementation of the GAIM in Stan, a widely used statistical modeling language that performs inference using Hamiltonian Monte Carlo (Carpenter *et al.*, 2017). No Bayesian implementation of the GAIM currently exists; developing a robust implementation and releasing an R package would be a significant contribution towards accessible and interpretable dimension-reduction models with epidemiological applications. Developing a robust Bayesian implementation includes identifying suitable priors for the weights α and methods for imposing shape constraints, such as monotonicity and convexity, on the s_i. We will also develop non-MCMC inference methods for these models, similar to those seen in Rue *et al.* (2009), to facilitate wider use of the GAIM.

Research Questions

In the course of this project, we will use the GAIM to explore the health effects of multiple air pollutants. Recent years have seen increased interest in modeling the joint effect of two or more pollutants in health outcomes (Billionnet *et al.*, 2012; Davalos *et al.*, 2017; Dominici *et al.*, 2010). Bayesian approaches can be seen in Blangiardo *et al.* (2019), Bobb *et al.* (2013), and Huang *et al.* (2018). We will consider two research questions in the course of our inquiry, namely,

1. What is the combined effect of multiple pollutants on various daily mortality outcomes?, and
2. What is the relationship between daily COVID-19 mortality and air pollution?

Question 1

The workhorse of the (predominantly frequentist) air pollution literature is the one-pollutant log-linear Poisson regression model. This model accounts for confounders using fixed effects and smooth functions, such as the natural cubic spline (Dominici *et al.*, 2002; Liu *et al.*, 2019; Samet *et al.*, 2000). Let the average rate an outcome occurs on day t be denoted by λ_t. Then, a typical one-pollutant model is,

$$
\begin{aligned}
Y_t|\lambda_t &= \text{Poisson}(\lambda_t),\\
\log(\lambda_t) &= X\beta + \gamma_1 P_{1,t} + \sum_{k=1}^{K} f_k(\eta_k).
\end{aligned}
$$

Here, Y_t is the health outcome of interest, such as respiratory mortality or morbidity. The design matrix X contains day-of-the-week effects and seasonal terms, and the $f_i(\eta_{k,*})$ are smooth functions of potential confounders such as time and temperature. One extension of this model would fit $P_{1,t}$ to a smooth function s_1. More generally, we could fit N pollutants to N smooth functions, $s_1, \ldots, s_N$. A third alternative, which we propose here, is to model N pollutants using the GAIM.

Air pollution is an excellent application area for the GAIM. People are exposed to the mixture of pollutants in their environment, many of which are highly correlated (Huang *et al.*, 2020). Thus, health effects the single pollutant model attributes to P_1 may very well be caused by a correlated pollutant, or only be present within certain combinations of a mixture of pollutants. Indeed, Franklin and Schwartz (2008) found that the effect of ozone on non-accidental mortality was “substantially reduced” after adjusting for particle sulfate. In Liu *et al.* (2019), the authors found significant differences in the percentage change in all-cause mortality attributable to $PM_{2.5}$ and PM_{10} when adjusting for NO_2 or SO_2.

It is therefore crucial to extend the one-pollutant model in a way that attributes health effects to the correct mixture of pollutants. This suggests using a model that conducts inference on mixtures. As for extending the s_i to be non-linear, there is evidence that some health outcomes are nonlinearly related to pollution levels, and that synergistic effects occur when multiple pollutants are present at higher levels

(Feng *et al.*, 2016; Xia and Tong, 2006). These extensions make the GAIM especially well-suited to estimating nonlinear effects of (pollutant) mixtures: the smooth function(s) s_i estimate the effects of their respective weighted sums, where the weights give the relative contribution of the components.

The multiple pollutant problem has received increased intention in recent years. . Five approaches are detailed in Davalos *et al.* (2017). Of these 5 methods, they note that adding additive main effects can lead to biased estimates in the presence of highly correlated variables, and that nonparameteric methods are often not very interpretable. The research teams and stakeholders involved in air pollution research are often diverse and inter-disciplinary. This makes it crucial that models have interpretable parameters, so that estimation results can be easily communicated to non-specialists. While the unsupervised dimension reduction methods (such as principle components analysis and clustering) identified in Davalos *et al.* (2017) are difficult to interpret, they note some supervised methods that consider weighted sums of pollutant concentrations. For example, Pachon *et al.* (2012) specify weights from data rather than estimating them, while Roberts and Martin (2006) introduces a model that is equivalent to assuming that s_1 is linear. While these are viable statistical methods, they are not as flexible or extensible as the GAIM.

Davalos *et al.* (2017) also discusses non-parametric methods, including Bayesian Kernel Machine Regression (BKMR). BKMR allows for estimation and variable selection. It was introduced in Bobb *et al.* (2015), and an R package was released with Bobb *et al.* (2018). BKMR models an exposure-response surface – the exposures can be pollutants and the response nonaccidental mortality – via a kernel function. Using a hierarchical Bayesian variable selection method, BKMR 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 GAIM, it does not have easily interpretable parameters. This makes the GAIM more suitable to the communication demands of inter-disciplinary research areas such as air pollution.

While there has been significant research interest in the multiple pollutant problem, the models proposed to date have either computational limitations or limited interpretability. In contrast, the GAIM has interpretable parameters and its computational burden does not scale with the dimensionality of its inputs. Therefore, using the GAIM to examine the health effect of a mixture of pollutants will provide interpretable and communicable results on this research question. Finally, note that while air pollution is the main example in this proposal, the GAIM has applications wherever the target of inference is a mixture of covariates that relate nonlinearly to an outcome of interest.

Question 2

The relationship between daily COVID-19 deaths and air pollution levels has become an active area of research in recent months. For instance, Wu *et al.* (2020) apply a zero-inflated negative bionomial model U.S. data, where the zero-inflation accounts for U.S. counties with no COVID-19 deaths. They use a log-linear link function with a state-level random effect, and 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 instance, non-COVID-19 daily mortality data is generally not yet available, such that we do not have an accurate measure of excess deaths attributable to COVID-19, especially among vulnerable populations such as seniors. These excess deaths could be attributable to under-reported COVID-19 case and death counts (due to limited testing), restricted access to care for patients with other health conditions, or potential reporting delays. Moreover, cumulative COVID-19 mortality will likely continue to rise for some time, making the question of excess deaths due to COVID-19 best suited to an ongoing inquiry that may help inform subsequent public health responses.

We will use the GAIM 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 scaleable, interpretable, and can capture non-linearities in the relationship between mixtures and the response. To apply the GAIM to this research question, we must choose the outcome of interest Y_t as COVID-19 deaths. To specifically investigate seniors, we may (for example) take Y_t to be those aged 65+. Applying the GAIM to air pollution and COVID-19 death data will allow us to ascertain which mixtures of pollutants increase COVID-19 mortality, as well as how their effects differ among age groups.

This inquiry will make two contributions to the growing COVID-19 literature. The first is that it will help determine how different mixtures of air pollutants amplify the impact of COVID-19. The second is that since people in different regions are exposed to different mixtures of pollutants, it will help identify which mixtures of pollutants have the largest impact on COVID-19 deaths. Both contributions will also further our understanding of the health impacts of air pollution.

References

Billionnet, C., Sherrill, D., Annesi-Maesano, I., et al. (2012) Estimating the health effects of exposure to multi-pollutant mixture. *Annals of Epidemiology*, **22**, 126–141.

Blangiardo, M., Pirani, M., Kanapka, L., et al. (2019) A hierarchical modelling approach to assess multi pollutant effects in time-series studies. *PloS one*, **14**.

Bobb, J. F., Dominici, F. and Peng, R. D. (2013) Reduced hierarchical models with application to estimating health effects of simultaneous exposure to multiple pollutants. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 451–472.

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

Bobb, J. F., Henn, B. C., Valeri, L., et al. (2018) Statistical software for analyzing the health effects of multiple concurrent exposures via bayesian kernel machine regression. *Environmental Health*, **17**, 1–10.

Braun, J. M., Gennings, C., Hauser, R., 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., Hoffman, M. D., et al. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**.

Conticini, E., Frediani, B. and Caro, D. (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., Herring, A. H., 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., Zeger, S. L., 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., Barr, C. 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., Sun, W., 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.

Huang, G., Lee, D. and Scott, E. M. (2018) Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. *Statistics in medicine*, **37**, 1134–1148.

Huang, G., Brown, Partrick and Shin, H. H. (2020) Daily mortality/morbidity and air quality: Using multivariate time series with seasonally varying covariances. *Submitted*.

Lazarevic, N., Barnett, A. G., Sly, P. D., 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., Sera, F., et al. (2019) Ambient particulate air pollution and daily mortality in 652 cities. *New England Journal of Medicine*, **381**, 705–715.

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

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

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. and Chopin, N. (2009) Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392.

Samet, J. M., Dominici, F., Curriero, F. C., et al. (2000) Fine particulate air pollution and mortality in 20 us cities, 1987–1994. *New England Journal of Medicine*, **343**, 1742–1749.

Sanders, A. P., Henn, B. C. and Wright, R. O. (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., Magrı‘, D., 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., De Gennaro, G., et al. (2020) Searching for sars-cov-2 on particulate matter: A possible early indicator of covid-19 epidemic recurrence.

Wu, X., Nethery, R. C., Sabath, B. M., 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.

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

Any suggestions?

Five CVs

- Patrick, Fateh, Hwashin, Meredith (?), Cindy

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 post-doctoral funding.
2. \$12,000/year to support a Laval University or University of Ottawa PhD student.
3. \$12,000/year to support a University of Toronto PhD student.
4. \$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.