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 callaborators, titles, and affiliations

Name: Kamal Rai
Title: Post-Doctoral Fellow (?)
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: 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 Bayesisan implementation of the *groupwise additive index model* (GAIM), a constrained frequentist 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 (Sanders, Henn, and Wright 2015; Braun et al. 2016; Lazarevic et al. 2019; Bobb 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 constrasts 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, and to characterize their performance under a wide variety of different applications.

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 (Dominici et al. 2010; Billionnet et al. 2012; Davalos et al. 2017). Bayesian approaches can be seen in Blangiardo et al. (2019), Bobb, Dominici, and Peng (2013), and Huang, Lee, and Scott (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 (Samet et al. 2000; Dominici et al. 2002; Liu et al. 2019). 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, Brown, and Shin 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 appraches 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) indentified 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 constrast, 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) finds that a 1 μg increase in long-term exposure to ambient $PM_{2.5}$ increases the coronavirus death rate by 15%. Additional studies that examine this relationship include Conticini, Frediani, and Caro (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 vulerable 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 improve subsequent public health responses.

References

Billionnet, Cécile, Duane Sherrill, Isabella Annesi-Maesano, and GERIE study. 2012. "Estimating the Health Effects of Exposure to Multi-Pollutant Mixture." *Annals of Epidemiology* 22 (2): 126–41.

Blangiardo, Marta, Monica Pirani, Lauren Kanapka, Anna Hansell, and Gary Fuller. 2019. "A Hierarchical Modelling Approach to Assess Multi Pollutant Effects in Time-Series Studies." *PloS One* 14 (3).

Bobb, Jennifer F, Francesca Dominici, and Roger D Peng. 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 (3): 451–72.

Bobb, Jennifer F, Birgit Claus Henn, Linda Valeri, and Brent A Coull. 2018. "Statistical Software for Analyzing the Health Effects of Multiple Concurrent Exposures via Bayesian Kernel Machine Regression." *Environmental Health* 17 (1): 1–10.

Bobb, Jennifer F, Linda Valeri, Birgit Claus Henn, David C Christiani, Robert O Wright, Maitreyi Mazumdar, John J Godleski, and Brent A Coull. 2015. "Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures." *Biostatistics* 16 (3): 493–508.

Braun, Joseph M, Chris Gennings, Russ Hauser, and Thomas F Webster. 2016. "What Can Epidemiological Studies Tell Us About the Impact of Chemical Mixtures on Human Health?" *Environmental Health Perspectives* 124 (1): A6–A9.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Conticini, Edoardo, Bruno Frediani, and Dario Caro. 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, Angel D, Thomas J Luben, Amy H Herring, and Jason D Sacks. 2017. "Current Approaches Used in Epidemiologic Studies to Examine Short-Term Multipollutant Air Pollution Exposures." *Annals of Epidemiology* 27 (2): 145–53.

Dominici, Francesca, Aidan McDermott, Scott L Zeger, and Jonathan M Samet. 2002. "On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health." *American Journal of Epidemiology* 156 (3): 193–203.

Dominici, Francesca, Roger D Peng, Christopher D Barr, and Michelle L Bell. 2010. "Protecting Human Health from Air Pollution: Shifting from a Single-Pollutant to a Multi-Pollutant Approach." *Epidemiology (Cambridge, Mass.)* 21 (2): 187.

Feng, Cindy, Jian Li, Wenjie Sun, Yi Zhang, and Quanyi Wang. 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 (1): 17.

Franklin, Meredith, and Joel Schwartz. 2008. "The Impact of Secondary Particles on the Association Between Ambient Ozone and Mortality." *Environmental Health Perspectives* 116 (4): 453–58.

Huang, Guowen, Brown Partrick, and Hwashin H. Shin. 2020. "Daily Mortality/Morbidity and Air Quality: Using Multivariate Time Series with Seasonally Varying Covariances." *Submitted*.

Huang, Guowen, Duncan Lee, and E Marian Scott. 2018. "Multivariate Space-Time Modelling of Multiple Air Pollutants and Their Health Effects Accounting for Exposure Uncertainty." *Statistics in Medicine* 37 (7): 1134–48.

Lazarevic, Nina, Adrian G Barnett, Peter D Sly, and Luke D Knibbs. 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 (2): 026001.

Liu, Cong, Renjie Chen, Francesco Sera, Ana M Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline SZS Coelho, et al. 2019. "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities." *New England Journal of Medicine* 381 (8): 705–15.

Masselot, Pierre, Fateh Chebana, Céline Campagna, Éric Lavigne, Taha B.M.J. Ouarda, and Pierre Gosselin. 2020. "Constrained Groupwise Additive Index Models." *Submitted*.

Pachon, Jorge E, Sivaraman Balachandran, Yongtao Hu, James A Mulholland, Lyndsey A Darrow, Jeremy A Sarnat, Paige E Tolbert, and Armistead G Russell. 2012. "Development of Outcome-Based, Multipollutant Mobile Source Indicators." *Journal of the Air & Waste Management Association* 62 (4): 431–42.

Roberts, Steven, and Michael A Martin. 2006. "Investigating the Mixture of Air Pollutants Associated with Adverse Health Outcomes." *Atmospheric Environment* 40 (5): 984–91.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 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 (2): 319–92.

Samet, Jonathan M, Francesca Dominici, Frank C Curriero, Ivan Coursac, and Scott L Zeger. 2000. "Fine Particulate Air Pollution and Mortality in 20 Us Cities, 1987–1994." *New England Journal of Medicine* 343 (24): 1742–9.

Sanders, Alison P, Birgit Claus Henn, and Robert O Wright. 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 (3): 284–94.

Sciomer, Susanna, Federica Moscucci, Damiano Magrı‘, Roberto Badagliacca, Gianfranco Piccirillo, and Piergiuseppe Agostoni. 2020. "SARS-Cov-2 Spread in Northern Italy: What About the Pollution Role?" *Environmental Monitoring and Assessment* 192: 1–3.

Setti, Leonardo, Fabrizio Passarini, Gianluigi De Gennaro, Pierluigi Barbieri, Alberto Pallavicini, Maurizio Ruscio, Prisco Piscitelli, Annamaria Colao, and Alessandro Miani. 2020. "Searching for Sars-Cov-2 on Particulate Matter: A Possible Early Indicator of Covid-19 Epidemic Recurrence."

Wu, Xiao, Rachel C Nethery, Benjamin M Sabath, Danielle Braun, and Francesca Dominici. 2020. "Exposure to Air Pollution and Covid-19 Mortality in the United States." *medRxiv*.

Xia, Yingcun, and Howell Tong. 2006. "Cumulative Effects of Air Pollution on Public Health." *Statistics in Medicine* 25 (20): 3548–59.

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

Kamal will develop the Bayesian implementation of the multiple pollutant models in Stan. This includes exploring determining appropriate prior distributions for the weights α and developing visualizations that communicate model results in a simple and interpretable way. 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 to Quebec to visit the collaborators there.

The University of Toronto PhD student will compare the results from the Bayesian random walk models to those obtained from non-MCMC methods. For example, these could include frequentist methods that fit (natural cubic) splines or Bayesian inference using R-INLA (Rue, Martino, and Chopin 2009). The University of Laval/University of Ottawa PhD student will compare the results from the Stan implementation to those obtained by a case-crossover model.

Plans for dissemination and communication

The results and findings of this multiple pollutant inquiry will be shared with Health Canada and the Institut National de Santé Publique du Québec. The lead investigators have a track record of publishing their research results in statistical and epidemiolgical journals, and aim to publish the results of this project in high-impact journals. They (or the trainees) will also attend appropriate conferences to present the work while it in progress.

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.