Title of the Project: Statistical Methods for Daily Mortality and Multiple Environmental Risk Factors

Name and affiliations of lead investigators

- Patrick E. Brown
  - Centre for Global Health Research, St. Michael's Hospital
  - Department of Statistical Sciences, University of Toronto
- Fateh Chebana
  - Centre Eau Terre Environnement, Institut national de la recherche scientifique, Québec.
- Cindy Feng
  - School of Epidemiology and Public Health, University of Ottawa
- Meredith Franklin
  - Keck School of Medicine, University of Southern California (until July 2021)
  - School for the Environment and Department of Statistical Sciences, University of Toronto (from July 2021)

List of proposed collaborators, titles, and affiliations

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

Name: Hwashin Shin
Title: Scientist
Affiliation: Environmental Health Science and Research Bureau, Health Canada

Name: Céline Campagna
Title: Responsable scientifique
Affiliation: Équipe Changements climatiques et santé, Institut National de Santé Publique du Québec

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

List of potential partner organizations

- The **Centre for Global Health Research, St. Michael's Hospital** will lead the health sciences research component of the project, providing data from the US and India and time of research staff to work on manuscripts. The Toronto-based component of the team will be located at CGHR and will be integrated into the Geospatial Mortality research group Dr. Brown leads, and a portion of the Toronto PhD student's salary will be funded though one of CGHR's research grants.
- **L'Institut National de Santé Publique du Québec** has the mandate to analyse, monitor and evaluate determinants of health, including environmental pollutants, and will give access to the Quebec provincial health databases. In-kind support will

also include health expertise for the statistical design and interpretation of research, including co-supervision of students.

- Hwashin Shin at **Health Canada** is central to this project. Her needs for an improved air quality indicator instigated the discussions which lead to this application. Dr. Shin has been funding Dr. Brown's team through Health Canada research contracts, and intends to provide at least $15,000 per year to the project.

Research Aims

Overview

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). Health effects of air pollution depend on the composition of pollutants in the air, not simply the 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 improve statistical methods for conducting inference on the health effects of simultaneous exposure to multiple environmental pollutants, with a focus on quantifying short-term effects of poor air quality on health outcomes at the population-level.

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 $\theta = (\theta_1, ..., \theta_d)$, link function g, and constraints as below, the cGAIM is,

$$
\begin{aligned}
&Y_t|\lambda_t \overset{i.i.d.}{\sim} D(\lambda_t, \tau) \\
&g(\lambda_t) = X_t\beta + s(\alpha^T Z_t) + f_1(W_{1t}) + ... + f_K(W_{Kt}).
\end{aligned}
$$

Here Y_t is the outcome of interest, β is the vector of regression coefficients for linear covariates X_t, and $f_1, ..., f_K$ are smoothing functions that account for potential confounding variables W_{kt}. The distinguishing feature of cGAIM lies in s, a smooth function fit to a linear combination of environmental covariates Z_t. The α is a vector of weights on the entries of Z_t, giving the relative contribution of each component of Z_t. The smooth function s is modelled with a flexible parametric family such as cubic splines or a random walk. Estimating α, or more specifically finding the set of plausible values for α, is the main challenge of working with cGAIM.

We will develop a Bayesian methodology for inference with the cGAIM – the bcGAIM. Compared to the cGAIM, the bcGAIM will provide two main statistical benefits. The first is quantification of the uncertainty for α; the cGAIM does not provide confidence intervals for α. Given that pollutants tend to be positively correlated, it is possible that α is not well identified and all reasonable measures of air quality are equally predictive of health outcomes. It is also possible that the posterior distribution of α is narrow and the importance of, say, nitrous oxide relative to ozone can be estimated with some certainty. Either result would be a significant contribution research on the health effects of multiple pollutant. The second advantage is that the bcGAIM will be able to accommodate a higher-dimensional α than previously used, allowing Z_t to contain pollution levels at different time lags. The joint posterior distributions of this high-dimensional α for different health outcomes will provide information on how quickly poor air quality affects various types of mortality and morbidity (or quite possibly show that multiple combinations of recent air quality measures are equally predictive).

We will also develop method for fitting a case/crossover models using bcGAIM, which offers advantages over the more standard log-linear Poisson model. 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 proportional hazards survival model, where each individual is in a separate strata and 'control days' are chosen to have the same baseline hazard as the event day. Additional details are given in the Methods section below.

Objectives

This proposal has three main research objectives. The first is to develop an air quality index for Canadian cities that accounts for the combined effects of multiple air pollutants. This index will be developed in collaboration with Health Canada and INSQ with the intention of it being used in a public warning system. The second aim is undertake epidemiological studies involving exposures to multiple environmental pollutants in the areas of the world where the Centre for Global Health Research has suitable data (Canada, India, USA). The third application is a specific application of the air quality index, namely investigating how exposure to different mixtures of pollutants affects daily COVID-19 mortality.

There is evidence that the health effects estimated from single pollutant models may be in fact be caused by a correlated pollutant omitted from the model. For example, Franklin and Schwartz (2008) found that the effect of ozone on non-accidental mortality was "substantially reduced" after adjusting for particle 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.

Further, there is evidence that some health outcomes are nonlinearly related to health outcomes levels (Feng *et al.*, 2016). Achieving the three objectives requires a non-linear (or semi-parametric) dose-response curve and a combined-effect multi-pollutant exposure model. Our bcGAIM model will meet the requirements necessary to fulfil these objectives due to its being able to estimate weights using data, allowing for nonlinear relationships between pollutants and health outcomes, and applicable to both Poisson time series and case-crossover models.

Interpretability of model outputs is another requirement of this project, and many 'unsupervised' methods (i.e. principle components analysis and clustering) are difficult to interpret (Davalos *et al.*, 2017). A popular nonparametric method is the Bayesian Kernel Machine Regression (BKMR), introduced in Bobb *et al.* (2015), which 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. The bcGAIM will provide some of the flexibility of BKMR and clustering while being sufficiently interpretable to meet the communication needs of an inter-disciplinary research team.

The relationship between daily COVID-19 deaths and air pollution levels has recently 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%. We will use variations on bcGAIM to examine the relationship between COVID-19 deaths and combinations of air pollutants. As well as exploring daily variations in the case fatality rate (and its relation to air quality), the model will be

adapted to consider long-term exposures and COVID-19 incidence rates.

Methods

The bcGAIM will make four methodological advancements for modeling the health effects of mixtures of pollutants. These are:

1. extending the cGAIM to higher dimensional problems;
2. fully exploring the parameter space to identify all plausible values for α;
3. developing priors for shape-constrained Bayesian inference on the smooth function s; and
4. using case crossover models in place of the Poisson response variable.

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 (in theory) to extend the bcGAIM to include additional pollutants, additional lags for pollutants in the model, and additional smooth functions s. It is expected that the α parameters will not always be well identified, and results will be sensitive to model assumptions and prior distributions. One task in this component of the research will be to find reparametrizations and multivariable prior distributions which enable prior elicitation from subject-area specialists.

Following its implementation in Stan, the bcGAIM will be implemented using non-MCMC inference methods, similar to Integrated Nested Laplace Approximation (INLA) (Rue *et al.*, 2009). One such algorithm for single-pollutant case-crossover models was recently developed by Stringer *et al.* (2020), and this approach will be extended to allow for inference on the α parameters in bcGAIM. 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. To facilitate use by other researchers, all bcGAIM software will be released in an R package.

After bcGAIM is implemented for a three-dimensional α, more specifically ozone, fine particulates, and nitrous oxide at two day lags, additional time lags will be added with the resulting α being 9- or 12 dimensional. The computational and methodological challenges at this stage are expected to be significant, and parallelizing the algorithm on cloud platforms will be used to dramatically increase the number of candidate values of α considered.

Another major task in developing the bcGAIM is in developing random effect distributions for shape-constrained Bayesian inference on s. In addition to having desirable statistical properties, the chosen models should be simple and interpretable so that it can be elicited from non-statistical experts. This is difficult to achieve for shape constraints. The first reason for this is that s may not have any parameters related to the desired constraint; for example, a 1^{st}-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, which is one of the goals of this project.

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**.

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

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., 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.

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*.

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.

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.

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. E., et al. (2020) Approximate Bayesian inference for case-crossover models.

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*.

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 Rai will complete his PhD in 2021 and will work on this project as a postdoc. He 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 the INLA-like Bayesian computations to conduct inference on the GAIM, and compare its results from those obtained from the Stan implementation. A University of Laval or 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. Drs Shin and Campagna will use the methodologies developed in their ongoing research and programatic work, and facilitate the adoption of the methods more widely in their organizations.

Suggested reviewers

- Jim Zidek, University of British Columbia. www.stat.ubc.ca/users/james-v-zidek-frsc-oc. Prof. Zidek is one of the world's foremost researchers on statistical method for environmental health.
- Rhonda Rosychuk, University of Alberta sites.ualberta.ca/~rhondar. Dr. Rosychuk has published extensively in both the health sciences literature and statistical journals, including on spatial and longitudinal models for environmental health problems.
- Samir Bhatt, Imperial College London www.imperial.ac.uk/people/s.bhatt, develops Bayesian inferential methods for complex models in public health research.

Preliminary budget description

The CANSSI Collaborative Research Team grant is for $60,000/year for 3 years. We propose an annual budget of:

1. $30,000/year to support a post-doctoral fellow.
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.

In addition to the CANSSI funding, Hwashin Shin will provide $15k/year towards the post-doc salary. This will likely be in the form of research contracts.