Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 116 (The number assigned by the OpenConf System)
DOI: TBA

Mutual Information Based Correlation Analysis of Health-Related Multiple Air Pollutants

Huawei Han, Wesley S. Burr
Trent University
1600 West Bank Drive, Peterborough, ON Canada
rebeccahan@trentu.ca; wesleyburr@trentu.ca

Abstract **–** Air pollutants and their adverse impacts on public health have been extensively considered in environmental epidemiology studies. Human health risk models with short-term air pollutant exposure have been of research interest for years, and the statistical methodologies have been developed, both univariate models and multivariate models which consider the joint-effects of air pollutants. However, the challenge of including multiple pollutants simultaneously for health risk assessment still exists, especially considering the problems of statistical interaction of pollutants, risk factor selection, and variable dimension reduction. This paper presents a mutual information based correlation analysis of air pollutants in Toronto, ON, Canada with the purpose of facilitating risk factor selection in multi-pollutant health risk models. In this work, the air pollutants datasets from the Canadian National Air Pollution Surveillance (NAPS) Program are prepared, and mutual information based correlation of different air pollutants are analysed. Results show that mutual information can provide effective dependency evaluation of air pollutants, allowing for more targeted and strategic variable selection for health risk models.

Keywords**:** air pollutants, public health, mutual information, correlation analysis, environmental epidemiology

1. Introduction

Short- and long-term exposure to air pollutants has been associated with adverse public health effects in numerous epidemiology studies. Exposure to the major air pollutants like ground-level ozone (O_3), particulate matter (PM), and oxides of sulphur (SO_x) and nitrogen (NO_x) have been shown to have a relationship with increasing cardiovascular diseases [1, 2], respiratory diseases [3], and mortality and morbidity [4]. Among all air pollutant-related epidemiological studies, one widely used method is using time-series models, specifically generalized additive models (GAMs) to assess the short-term risk of ambient air pollutants on public health [5-9]. These models and methods have developed from using a single air pollutant as the risk factor in early work, to attempts at using multiple ones simultaneously in later works. In these multiple-pollutant health risk assessment studies, the adverse effects of combinations of different kinds of air pollutants are discussed [10-15]. Although the multiple-pollutant risk estimation models are closer to the actual environment, as people are generally exposed to different air pollutants simultaneously, challenges still exist in their modelling and analysing. On the one hand, different air pollutants are physically and chemically mixed from their emission sources to the human interaction during their spreading. These mixtures are not easy to capture and explain in the multi-pollutant statistical models. On the other hand, correlations between different pollutants make the selection of primary risk factors for some particular heath issue difficult, especially with some highly correlated pollutants like NO, NO_2 and NO_x or PM_{10} and $PM_{2.5}$. Moreover, collinearity brought by correlated air pollutants reduces their interpretation and statistical significance in regression analysis.

With the aforementioned issues, correlation analysis of different air pollutants is usually done to support the risk assessment modelling process, especially the variable selection component. As the associations between different air pollutants are usually nonlinear for different datasets and problems, the Pearson correlation coefficient (PCC) used in our previous work [14, 15] for screening cannot always capture the pollutants' relationship properly. In order to facilitate the modelling process and the analysis of the adverse effects, mutual information based correlation analysis is applied to the ambient air pollutants data obtained from the National Air Pollution Surveillance database managed by Environment and Climate Change Canada [16]. In Section 2, the air pollutant dataset used for analysis is introduced and pre-processed. In Section 3, mutual information-based maximal information coefficient is used to capture and analyse the relationship of

different air pollutants and comparisons with Pearson correlation coefficients are presented. A conclusion is made in Section 4.

2. Data Preparation

The ambient air pollutants data used in the following analysis are from the Canada-Wide Air Quality Database of the National Air Pollution Surveillance program. This program aims to evaluate air quality related health issues and Canadian environmental sustainability. The air pollutants are monitored continuously and include NO, NO_2, NO_X, CO, SO_2, O_3, $PM_{2.5}$ and PM_{10}. Based on their original hourly concentration, 24-hour mean concentrations are usually used as the human exposure factor for investigating their health risk impacts. In this study, daily means of the air pollutants are first calculated and then imputation [17] is used to deal with the missing days caused by sensor malfunction. As a majority of PM_{10} data are not available and the data for 2022 have not been updated for the impact of the COVID-19 pandemic, the other seven air pollutants are used in this work with the measurement timescale set to 2000-2021. In addition, historical daily mean temperature from the Canadian Centre for Climate Services [18] is also used as a health risk factor as in most models, and missing days are also imputed.

The datasets of temperature (T) and seven air pollutants from 2000-2021 for Toronto, ON, Canada are shown in Fig. 1. It can be seen that temperature and ozone have obvious periodicities on a yearly basis. People are prone to be exposed to higher O_3 concentrations in high temperature months as the O_3 concentration is closely related to solar irradiance due to the chemistry of the formation. While the daily concentrations of CO, NO, NO_2, NO_X and SO_2 all have downward trends for the years to be analysed, due to increasing air pollution standards and decreases in polluting sources. It can also be seen that the concentrations of $PM_{2.5}$ are more stable over this time span.

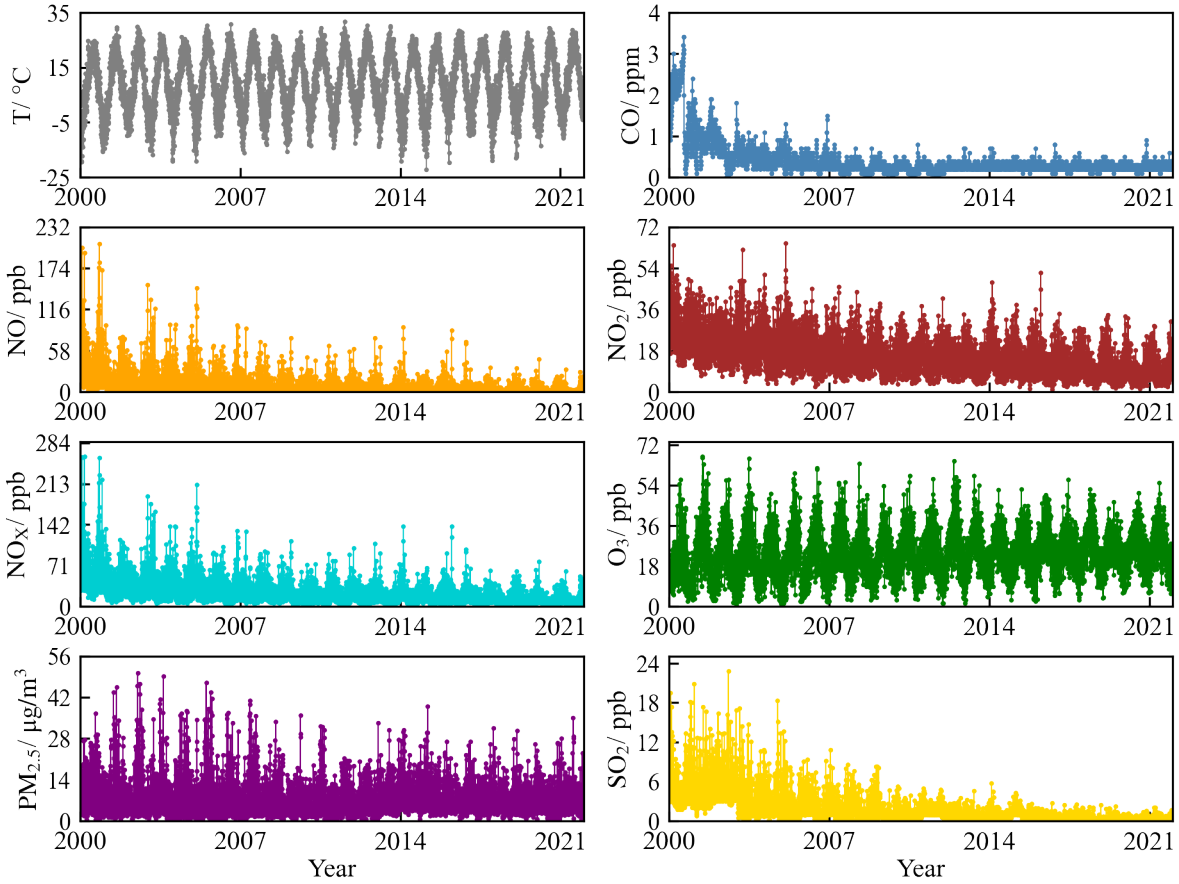

Fig. 1: Daily mean of temperature and air pollutants for Toronto, ON, Canada from 2000 to 2021

It is known from the physical and chemical formation process that the aforementioned air pollutants have complex associations and cannot be seen as independent factors in the health risk assessment models. As the air pollutant time series are usually pre-processed to remove their time dependency before being applied to health risk assessment, the relations of several time-residual air pollutants are visualized in Fig. 2 using the 22 years of data. In the first subfigure of Fig. 2, the relation between time-residual NO and time-residual NO_X is approximately linear, which tracks given NO as a component within NO_X. In the other three subfigures, the long timescale-independent relations of O_3 and temperature, O_3 and $PM_{2.5}$ as well as NO_2 and CO are not obviously linear. Using the Pearson correlation to evaluate the degree of their relevance may not be able to capture their true relations, especially when choosing air pollutants that contribute to the health risk in statistical assessment models like GAMs.

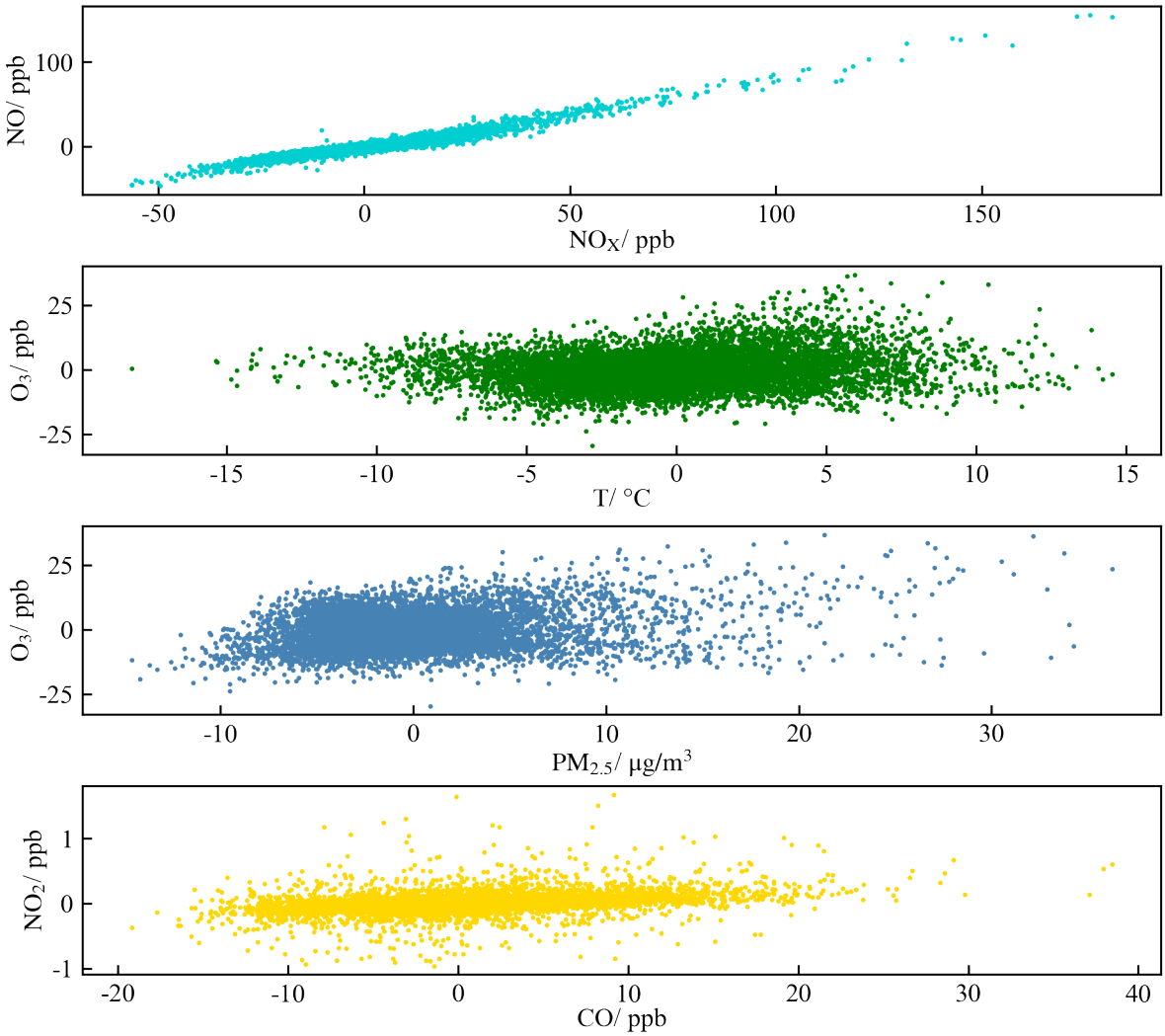

Fig. 2: Relations of different air pollutants (short timescale only residuals) for Toronto, ON, Canada.

3. Mutual Information Based Correlation Analysis of Air Pollutants

In order to effectively measure the correlations of these air pollutants in a general form and facilitate the analysis of their contributions in health risk assessment models, mutual information (MI) is used to capture both the linear and nonlinear relations. MI is a quantity that can evaluate the mutual dependence of two random variables by using entropy to measure the uncertainty [19]. A higher absolute value of MI indicates a stronger relation (positively or negatively correlated) between two variables, and a lower one indicates that they are weakly associated. Zero value indicates that there is no statistical correlation. For two different air pollutants X and Y, their MI can be calculated using discrete samples as follows:

$$MI(X,Y) = \sum_{x}\sum_{y} p_{XY}(x,y) log_2 \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \quad (1)$$

where $p_{XY}(x, y)$ is their joint probability distribution, and $p_X(x)$ and $p_Y(y)$ are their marginal distributions.

Although mutual information in Eqn. (1) can capture various relations between different air pollutants, there still exist some limitations for our datasets. The main problem is that the monitored air pollutants are measured with different units on different scales, and mutual information between different variables cannot be adjusted to a common scale, the result of which is that comparison of mutual information for different air-pollutant pairs is not available. Therefore, mutual information based maximal information coefficient (MIC) [20] is used in our analysis. MIC is a quantity developed from MI to measure the association between two random variables in a normalized form. Its main idea is to use a binning method to partition the variables into different bins and choose the gridding with maximum normalized MI. MIC used in the following correlation analysis is written as:

$$MIC(X,Y) = \max_{n_X \times n_Y < N} \frac{MI(X,Y)}{log_2 \min(n_X, n_Y)} \quad (2)$$

where n_X and n_Y are the number of grids for air pollutants X and Y, and N is a value related to the number of samples. It is set to $S^{0.6}$ in the following analysis (S is the sample number). Note also that the $MI(X, Y)$ in Eqn. (2) is the mutual information calculated based on partition of two air-pollutant samples to $n_X \times n_Y$, where their joint and marginal probability distributions are calculated. The scale of MIC is from 0 to 1 and this normalized form can be used to compare the associations of different air-pollutant pairs with various measurements.

With Eqns. (1) and (2), the relations of different air pollutants (including temperature) for Toronto, ON, Canada are evaluated using pre-processed data and presented in the heat map of Fig. 3, where PCC stands for Pearson correlation coefficient (absolute value) and MIC is the maximal information coefficient. It can be seen from the figure that the correlations measured using MICs are relatively milder than using PCCs in general. Compared to PCCs, the influence of T on concentrations of air pollutants is smaller evaluated using MICs except for NO and NO_X. Both PCCs and MICs show that CO, NO, NO_2, NO_X and SO_2 are weakly associated with T, while MICs show a relatively stronger relationship between NO or NO_X and T. This implies that PCCs are not able to capture all the associations and cannot fully reflect their relevance for screening and variable selection. CO and $PM_{2.5}$ are correlated with the other six air pollutants and temperature, evaluated using both PCCs and MICs, with MICs showing weaker relations. SO_2 and O_3 are also shown to have some relevance with other five pollutants using both PCCs and MICs. However, MIC shows a mild relationship between SO_2 and O_3, where their PCC shows they are almost independent. This is another case that PCC does not fully capture the relations between different pollutants. Although the MIC values are relatively smaller, both PCCs and MICs show strong correlations of NO, NO_2, and NO_X, as these pollutants usually come from human activity and have the same sources like industrial emission, fossil fuel electric utilities and motor vehicles.

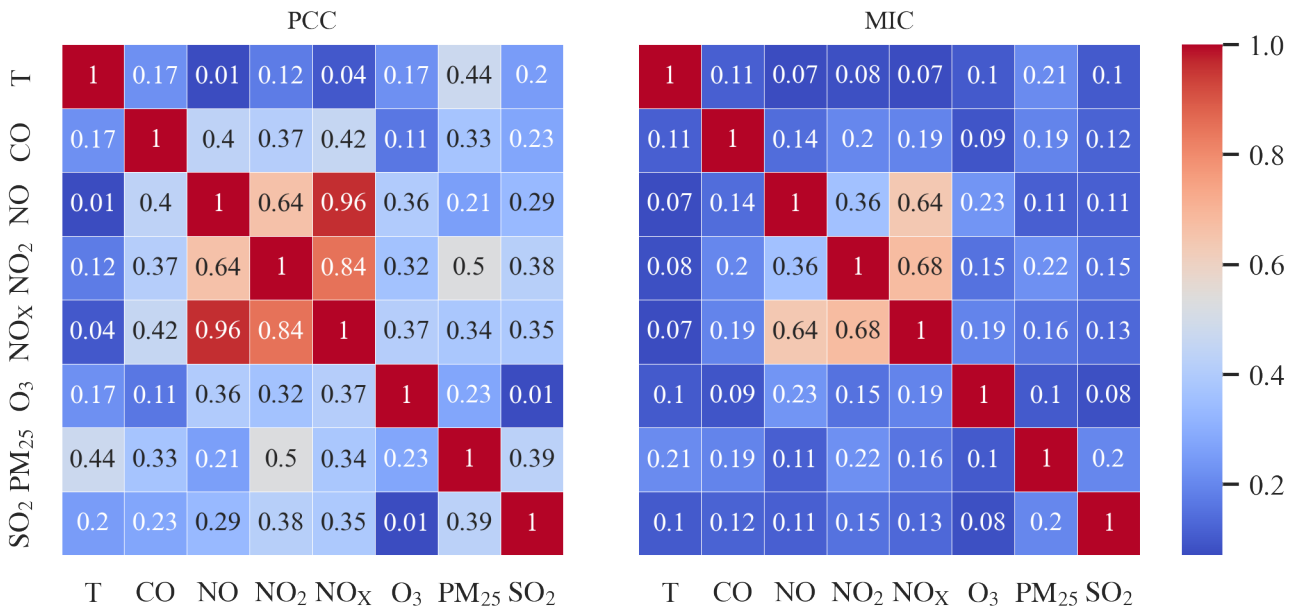

Fig. 3: Comparison of PCCs and MICs with data of all seasons for Toronto, ON, Canada.

Furthermore, as the health risk in air pollutant-related risk assessment models like GAMs is usually evaluated for different seasons, relations of different air pollutants using data of both warm (from April to September) and cold (from October to March) seasons are evaluated and presented in Fig. 4 and Fig. 5, respectively. It can be seen from the figures that in both cold and warm seasons associations evaluated using MICs are milder than using PCCs in general, following the patterns in Fig. 3. However, MIC values for T and NO, T and NO_2, and T and NO_X in cold seasons, as well as MIC values for T and NO_X, CO and O_3, NO and $PM_{2.5}$, NO_2 and O_3 in warm seasons are all larger than their PCC values. This means that in these different season scenarios MIC still shows its capability to capture more associations between different air pollutants compared to PCC, and that PCC may lead to incorrect association conclusions in nonlinear situations.

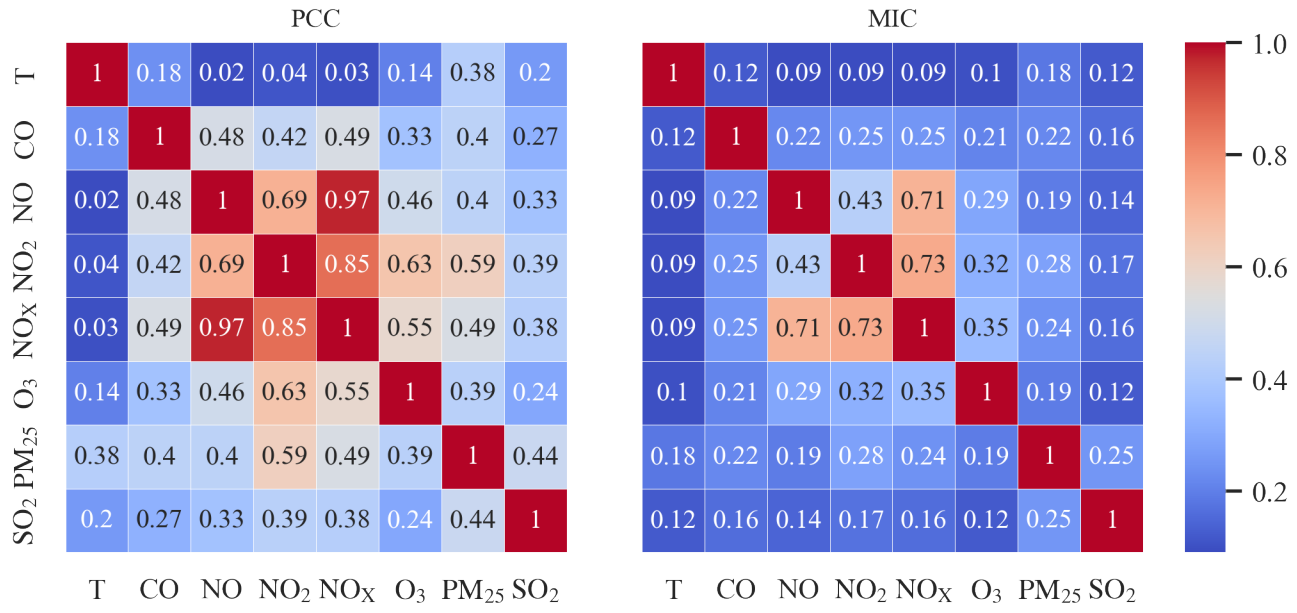

Fig. 4: Comparison of PCCs and MICs with data of cold seasons (from October to March) for Toronto, ON, Canada.

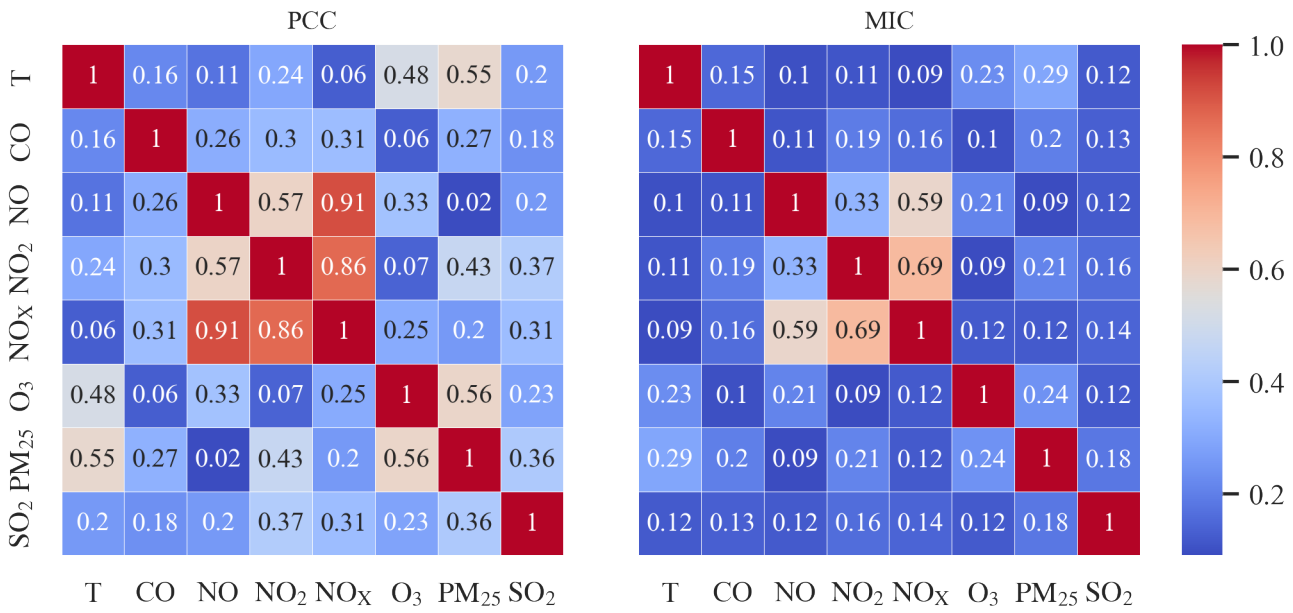

Fig. 5: Comparison of PCCs and MICs with data of warm seasons (from April to September) for Toronto, ON, Canada.

Note that in the above analysis, the values of PCC are mainly used for comparison and they may be not effective in situations where the relations between different air pollutants are not linear. Compared to PCC, as MIC can capture multiple types of associations instead of only linearity captured by PCC, this provides a more effective approach for dependency analysis and has the potential to facilitate risk factors selection for air pollutant related health models like GAMs as well as analysis of their risk contributions therein.

4. Conclusion

Research of ambient air pollutants on human health has been of interest for years in environmental epidemiology studies. As different air pollutants are usually correlated with complex physical and chemical processes, analysing their associations is helpful for selection of risk factors and variable inputs to the models, and hence analysis of their contributions in health risk models. In this work, mutual information-based maximal information coefficient is used to evaluate the relations of different ambient air pollutants. Using air pollutant data from 2000 to 2021 for Toronto, ON, Canada, comparisons with Pearson correlation are made and results show that Pearson correlation is less effective in general relevance evaluation and even provides wrong measurements under nonlinear dependencies, while maximal information coefficient can capture different kinds of associations and provides a more effective approach for the objective of correlation analysis. In conclusion, maximal information coefficient is an effective association metric and has the potential to facilitate risk factors selection and risk contribution analysis for statistical health risk models in the future work.

References

[1] Y. Hong, J. Lee, H. Kim, and H. Kwon. “Air pollution: a new risk factor in ischemic stroke mortality,” *Stroke*, vol. 33, no. 9, pp. 2165-2169, 2002.

[2] V. N. Likhvar, M. Pascal, K. Markakis, A. Colette, D. Hauglustaine, M. Valari, Z. Klimont, S. Medina and P. Kinney, "A multi-scale health impact assessment of air pollution over the 21st century." *Science of the Total Environment* 514 (2015): 439-449.

[3] C. A. Pope Iii, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston, "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution." *Jama* 287, no. 9 (2002): 1132-1141.

[4] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, “The contribution of outdoor air pollution sources to premature mortality on a global scale,” *Nature*, vol. 52.5, no. 7569, p. 367–371, 2015.
[5] F. Dominici, A. McDermott, S. L. Zeger, and J. M. Samet, “On the use of generalized additive models in time-series studies of air pollution and health,” *American Journal of Epidemiology*, vol. 156, no.3, pp.193–203, 2002.
[6] F. Dominici, A. McDermott, T. J. Hastie, “Improved semiparametric time series models of air pollution and mortality,” *Journal of the American Statistical Association*, vol. 99, no.468, pp. 938–948, 2004.
[7] J. B. Souza, A. R. Valdério, C. F. Glaura, I. Márton, B. Pascal, and J. M. Santos, “Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data,” *Journal of the Royal Statistical Society Series C*: *Applied Statistics*, vol. 67, no. 2, pp. 453-480, 2018.
[8] K. K. Mokoena, C. J. Ethan, Y. Yu, K. Shale, and F. Liu, “Ambient air pollution and respiratory mortality in Xi’an, China: a time-series analysis,” *Respiratory research*, vol. 20, no. 1, pp. 1-9, 2019.
9] C. Liu, R. Chen, F. Sera, A. M. Vicedo-Cabrera, Y. Guo, S. Tong, M. S. Coelho, P. H. Saldiva, E. Lavigne, P. Matus and N. Valdes Ortega,"Ambient particulate air pollution and daily mortality in 652 cities." *New England Journal of Medicine* 381, no. 8 (2019): 705-715.
[10] S. Roberts, “A new model for investigating the mortality effects of multiple air pollutants in air pollution mortality time-series studies,” *Journal of Toxicology and Environmental Health*, *Part A*, vol. 69, no. 6, pp. 417–435, 2006.
[11] S. Roberts and M. A. Martin, “Investigating the mixture of air pollutants associated with adverse health outcomes,” *Atmospheric Environment*, vol. 40, no. 5, pp. 984-991, 2006.
[12] F. Dominici, R. D. Peng, C. D. Barr, and M. L. Bell, “Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach,” *Epidemiology*, vol. 21, no. 2, pp. 187-194, 2010.
[13] W. S. Burr, “Air pollution and health: Time series tools and analysis,” Ph.D. dissertation, Dept. of Mathematics & Statistics, Queen’s Univ., Kingston, ON, Canada, 2012.
[14] S. Jarvis and W. S. Burr, “Development of a multi pollutant model to assess air pollution association with human health effects,” in *Proceedings of the* 4th *International Conference on Statistics*: *Theory and Applications* (*ICSTA*'22), Prague, Czech Republic, 2022, DOI: 10.11159/icsta22.151.
[15] S. Jarvis, “Particulate matter component analyses in relation to public health in Canada,” M.S. thesis, Dept. of Mathematics, Trent Univ., Peterborough, ON, Canada, 2022.
[16] Environment and Climate Change Canada. (2023). National Air Pollution Surveillance Program [Online]. Available: https://open.canada.ca/data/en/dataset/1b36a356-defd-4813-acea-47bc3abd859b.
[17] S. Castel and W. S. Bur, “Assessing Statistical Performance of Time Series Interpolators,” *Engineering Proceedings*, vol. 5, no. 1, pp. 1-11, 2021, DOI: doi.org/10.3390/engproc2021005057.
[18] Government of Canada. (2023) Historical Climate Data. Open Government License, https://climate.weather.gc.ca/.
[19] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, pp. 066138-1-16, 2004.
[20] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S.R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.