2022 19th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). Mexico City, Mexico. November 9-11, 2022

Exploring nonlinear effects of air pollution on hospital admissions by disease using gradient boosting machines

Carlos Minutti-Martinez
UPIITA
National Polytechnic Institute
Mexico City, Mexico
carlos.minutti@iimas.unam.mx

Antonio Galindo
UPIITA
National Polytechnic Institute
Mexico City, Mexico
mcantoniog@gmail.com

Luis F. Valdez-Garduño
UPIITA
National Polytechnic Institute
Mexico City, Mexico
luisfvaldezg@gmail.com

Miguel F. Mata-Rivera
UPIITA
National Polytechnic Institute
Mexico City, Mexico
migfel@gmail.com

***Abstract*—Air pollution has been linked to premature mortality and reduced life expectancy, with acute and chronic effects on human health. These effects can be difficult to measure because of possible interactions and nonlinear relationships with other variables such as age, weight, sex, and socioeconomic status.**

Multi-dimensional relationships are difficult to model using conventional statistical methods. However, modern machine learning techniques have been quite successful in this domain.

In this study, gradient boosting regression trees are used to predict the severity/mortality of the leading causes of hospitalization in Mexico City for 91,964 patients during the years 2015-2020 to measure the impact due to different air pollutants. The results show multiple nonlinear relationships and a significant effect of air pollutants on some of the most prevalent diseases.

***Index Terms*—artificial intelligence, air pollution, fine particulate matter, nonlinear effects, hospitalizations, health**

I. Introduction

Predictive analytics play an important role in clinical research. Conventionally, predictive analytics is performed using parametric modeling which comes with a number of assumptions. For example, generalized linear regression models require linearity and additivity to hold for the underlying data. However, these assumptions may not hold in practice [22].

Conventional modeling methods have trouble capturing high-dimensional relationships. However, some sophisticated machine learning techniques (ML) have been invented to handle this situation. Gradient Boosting Machine (GBM) is one of these techniques which is able to recursively fit a weak learner to the residual so as to improve model performance with a gradually increasing number of iterations. It can automatically discover complex data structures, including nonlinearity and high-order interactions [22].

The GBM model has already been successfully used in predictive analytics in medicine against other ML and statistical models, e.g., Kong *et al.* [12] used 86 predictor variables to predict in-hospital mortality of patients with sepsis using different modes: least absolute shrinkage and selection operator (LASSO), random forest (RF), GBM and logistic regression (LR), with the GBM model showing the best performance.

Air pollutants (AP), such as carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_x), volatile organic compounds ($VOCs$), ozone (O_3), heavy metals, and respirable particulate matter ($PM_{2.5}$ and PM_{10}), differ in their chemical composition, reaction properties, emission, time of disintegration, and ability to diffuse over long or short distances. Air pollution has both acute and chronic effects on human health, affecting a number of different systems and organs. Short and long-term exposures have also been linked with premature mortality and reduced life expectancy [11]. In both short-term and long-term studies, air pollution has an effect on cardiac deaths and hospital admissions in addition to respiratory effects [4].

The GBD 2019 Risk Factors Collaborators [8] studied the trends in exposure to leading risk factors on human health from 1990 to 2019 in 204 countries finding that the largest increases in risk exposure were for ambient particulate matter pollution. They use meta-regression for risk functions that might not be log-linear to allow for monotonically increasing or decreasing but potentially nonlinear functions.

Liu *et al.* [14] evaluated the associations of PM_{10}, $PM_{2.5}$ with daily all-cause, cardiovascular, and respiratory mortality across multiple countries using overdispersed generalized additive models to control for potentially nonlinear confounding effects of weather conditions, founding a consistent increase in daily mortality with increasing PM concentration, with steeper slopes at lower PM concentrations. Brown *et al.* [3] mention how various toxic exposures have been associated with the incidence of diabetes or with health outcomes associated with diabetes, such as cardiovascular disease, kidney disease, and hypertension.

In addition, some studies (see [7], [9]) showed a relationship between socioeconomic status (SES) and AP exposure, but also between SES and AP sensitivity, making evident the need to include AP and SES as confounding effects in health studies to avoid incorrectly estimating the true effect of AP and SES.

In this regard, a meta-analysis (see [10]) presents multiple

Funding by the Mexican National Council of Science and Technology (CONACYT) through the project "Data Science and Health".

978-1-6654-5508-4/22/$31.00 ©2022 IEEE

examples of studies in which, after confounding by SES, the estimated effect of the AP changed. Despite a vast air pollution epidemiology literature to date and the recognition that lower SES populations are often disproportionately exposed to pollution, there is little research identifying optimal means of adjusting for confounding by SES in air pollution epidemiology, nor is there a strong understanding of biases that may result from improper adjustment [10].

In this study, we analyzed data from 91,964 patients hospitalized in Mexico City from 2015-2020 to model the severity of the leading causes of hospitalizations by using GBM to automatically consider nonlinear confounding effects as well as interactions to estimate the effect of each variable.

The paper is structured as follows: Section II presents the source data, the methodology for constructing the SES and AP factors, and the model. Section III presents the results, including the performance and importance of the explanatory variables, as well as graphs of their nonlinear effects. Finally, Section IV includes the conclusion and future work.

II. Materials and Methods

To determine the relevance of the AP in the leading causes of hospitalization, we used as predictor variables for each patient those most closely related to the severity of hospitalization and included SES and AP indicators associated with the patient's locality of residence.

Severity is determined by the number of days of hospitalization and whether or not death occurred for each patient. The contribution of each variable is estimated by means of a score using the relative feature importance given by the GBM model and its ROC-AUC value, contrasting the score obtained for the different AP factors in each of the causes of hospitalization.

The following sections discuss the related details of each of the components mentioned here for the modeling, analysis, and interpretation process. For all variables, a 0-1 scale was applied, so it is possible to compare the coefficients and weights.

A. Data

There are three main data sources used in the study:

a) Hospitalizations: Anonymized data from public hospitals were provided by the Ministry of Health of Mexico City (SEDESA) in accordance with CONACYT project 7051, where the data set includes information on each patient such as age, weight, gender, origin, an indicator of first or subsequent hospitalization, entitlement to health services, date of admission and discharge, days of hospitalization, conditions, locality of residence, and International Classification of Diseases (ICD) codes for the initial diagnosis of hospitalization and the main condition, additionally, in the case of death, ICD code for the base cause of death.

b) Air pollutant concentrations: AP measures were obtained from Mexico City's Automatic Air Quality Monitoring Network [18]. A fifteen-year average (2005-2020) for each monitoring station was calculated for the concentrations of PM_{10}, $PM_{2.5}$, CO, NO_X, NO_2, SO_2, NO, and O_3. The mean concentrations corresponding to the locality of residence of the patients were obtained by kriging and QGIS, using the centroid of each locality.

c) The Census of Population and Housing: This data set is the official 2020 Census of Mexico which contains housing and population variables at the locality level that were used to construct official SES indicators.

B. SES and AP factors

Census data are widely used for constructing neighborhood-level composite SES indicators by using Principal Component Analysis (PCA) or Factor Analysis (FA) to weigh each variable's contribution [16], [21].

The SES indicators for this study were derived using FA to have economic and social factors as separate effects, resulting in factors F_ECONOM and F_SOCIAL, where high values of the factors represent less favorable circumstances. To validate these factors, they were used as predictors in a regression analysis to estimate the Social Gap Index (SGI) from the National Council for the Evaluation of Social Development Policy [5], Social Development Index (SDI) from the Mexico City Social Development Evaluation Council [6] and the Human Development Index (HDI) from United Nations Development Programme [19], resulting in coefficients of determination $R^2 > 0.9$, for all of them.

Regarding the AP, the number and distribution of mountains make Mexico City and its metropolitan area a highly complex terrain, influencing meteorology and how pollutants behave in the atmosphere. Local winds influence air quality and the distribution of pollutants [17]. This result in spatially correlated AP, so an analysis that does not consider this component could lead to incorrect estimates of the effects of each pollutant, therefore factors are constructed by grouping highly correlated pollutants by location.

PCA was used to determine which pollutants to group (Fig. 1). To determine the weights of each pollutant within its group, additional PCAs were performed for each group, resulting in the following factors and weights:

$$\begin{aligned} PM\_CO &= 0.35 \cdot PM_{10} + 0.39 \cdot PM_{2.5} + 0.26 \cdot CO \\ NO2\_NOx &= 0.54 \cdot NO_X + 0.46 \cdot NO_2 \\ SO2\_NO\_O3 &= 0.35 \cdot SO_2 + 0.33 \cdot NO + 0.32 \cdot O_3 \end{aligned}$$

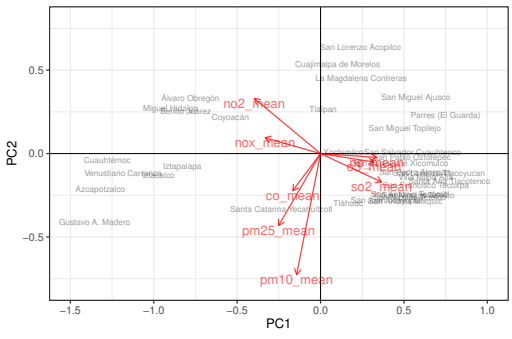

Fig. 1. Principal component analysis for air pollutants at different localities of Mexico City.

C. Severity model

For the analyses, the main causes of hospitalization that were selected are as follows:

- Renal insufficiency (RI).
- Diabetes mellitus (DM).
- Liver diseases (LD).
- Cerebrovascular diseases (CVD).
- Heart diseases (HD).
- Influenza and Pneumonia (I&P).
- COVID-19, virus identified (COVID-IV).

which are among the causes of hospitalization with the highest number of deaths during the period studied.

The severity of hospitalization for each patient is measured as a composite indicator expressing whether death occurred and the number of days of hospitalization. Thus, the severity Y of patient i is represented as follows:

$$Y_i = \frac{1}{1+X_i};\ X_i = \frac{\text{h\_days}_\text{i}}{\max_\text{d}}, \text{if death occurred.},$$
$$X_i = \frac{\max_\text{d}}{\text{h\_days}_\text{i}},\ \text{otherwise}$$

where $\max_\text{d}$ is the maximum number of hospitalization days in the data set and h_days_i number of days of hospitalization for patient i.

Therefore, if $Y_i > 0.5$ implies that death occurred and the faster the death occurred, Y_i is closer to 1 indicating greater severity. $Y_i < 0.5$ indicates that death did not occur and the lower the number of hospitalization days the discharge occurred, the Y_i is closer to 0, indicating less severity. This also can be considered as a classification problem, with two classes, high severity (death) and low severity (not death), but with higher weights for extreme cases.

At the patient level, the variables considered relevant for modeling severity are age, weight, gender, and origin (external, emergency room, referred, other, unspecified). At the level of the locality of residence, the patient's municipality of residence and date of admission, expressed as months 1-72 are considered.

The municipality of residence is relevant since each municipality may be correlated with higher levels of AP or SES, have a different hospital infrastructure, health policy, or factors not considered that could present a spurious correlation with other variables.

The GMB implementation used for the analysis was the *GradientBoostingRegressor* in the Python package *scikit-learn*, using a random sample of 85% of the records as the training data set, and 15% for the validation data set.

D. Scoring of relevant variables

To select how relevant each variable is to predict the severity of a cause of hospitalization, the following score was used:

$$\text{Score}_i = \frac{\text{var\_imp}_i}{\sum_{j=1}^{n} \text{var\_imp}_j} \cdot \text{AUC-val} \qquad (1)$$

where var_imp_i the Gini importance of the variable i, estimated by the GBM algorithm, and AUC-val is the AUC score of the model for the validation data set. If the variable i has a $\text{Score}_i = 1$, it can be used to perfectly categorize if a patient will have a low or high severity hospitalization.

III. Results

For each cause of hospitalization, Table I presents the score of Eq. (1) for the variables age (AGE), weight (WEIGHT), economic factor (F_ECO), social factor (F_SOC), NO_2 and NO_x pollutants factor (NO2_NOx), PM and CO pollutants factor (PM_CO), SO_2, NO_2 and O_3 pollutants factor (SO2_NO_O3), the number of hospitalizations (N_HOSP), the validation AUC of the model (AUC), and the absolute effect of the AP (AP_ABS), which results from adding the score of each AP factor and multiplying it by the annual number of hospitalizations. Results are ordered from highest to lowest AP_ABS.

When comparing results by cause of hospitalization, Cerebrovascular diseases and Liver diseases are the causes that showed greater sensitivity to exposures of (PM, CO), as for (NO_2, NO_x) it was Renal insufficiency the most sensitive to exposures, and Liver diseases for exposures of (SO_2, NO O_3).

Within each cause of hospitalization, interactions and nonlinear effects are explored using partial dependence plots. The following results are observed:

A. Renal insufficiency

Fig. 2 presents the most important variables for renal insufficiency severity. The most important variable is when the patient comes from the emergency department (PROCED_2), followed by age and weight. The factor (NO_2, NO_x) are the most relevant AP followed by (SO_2, NO O_3), however, in the validation set it is observed that it is (NO_2, NO_x) that maintains its importance. It is also observed that despite the inclusion of SES and AP factors, some municipalities of residence (E_MUN prefix) have greater importance, indicating other severity effects associated with the place of residence.

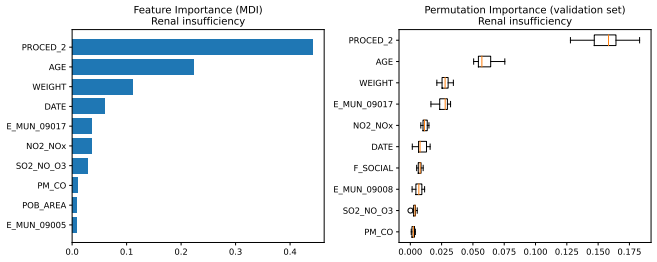

Fig. 2. Renal insufficiency - Feature importance.

Fig. 3 shows the joint effect of selected variable pairs on severity. Both age and weight have nonlinear effects, with approximately 50 years of age and older showing a more rapid increase in severity, and both high and low weight values also have a greater severity effect.

The main AP factor, (NO_2, NO_x), also shows a nonlinear effect where severity increases from exposures above the area

TABLE I
CAUSE OF HOSPITALIZATION AND SCORE OF RELEVANT VARIABLES FOR SEVERITY

	AGE	WEIGHT	F_ECO	F_SOC	NO2_NOx	PM_CO	SO2_NO_O3	N_HOSP	AUC	AP_ABS
Renal insufficiency	0.182	0.092	0.005	0.004	0.029	0.008	0.023	24847	0.820	249.4
COVID-19, virus identified	0.320	0.086	0.047	0.020	0.009	0.018	0.011	2632	0.714	134.7[a]
Diabetes mellitus	0.234	0.084	0.013	0.016	0.015	0.009	0.003	26656	0.763	123.1
Heart diseases	0.200	0.090	0.006	0.010	0.015	0.017	0.018	12254	0.736	103.5
Liver diseases	0.151	0.111	0.006	0.016	0.013	0.023	0.035	5839	0.615	69.7
Influenza and Pneumonia	0.736	0.021	0.002	0.004	0.002	0.014	0.004	14837	0.864	47.9
Cerebrovascular diseases	0.106	0.139	0.013	0.017	0.013	0.024	0.015	4899	0.609	42.8

N_HOSP = Number of hospitalizations, AUC = AUC of the validation data set, AP_ABS = (NO2_NOx + PM_CO + SO2_NO_O3) · N_HOSP / N_YEARS.
[a] N_YEARS=0.75 for COVID-19, and N_YEARS=6 for the other causes of hospitalization.

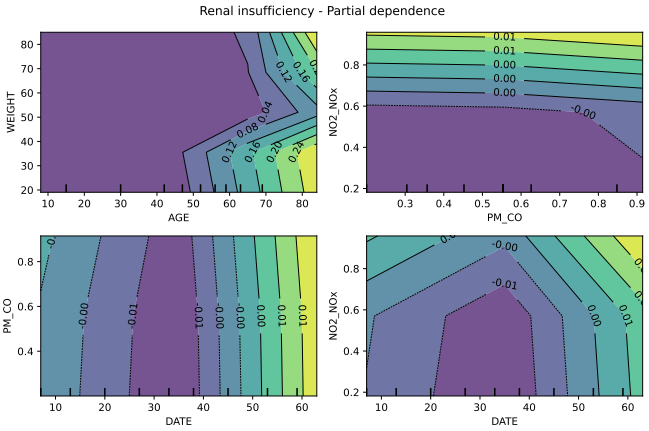

Fig. 3. Renal insufficiency - Nonlinear effects on severity for pairs of variables

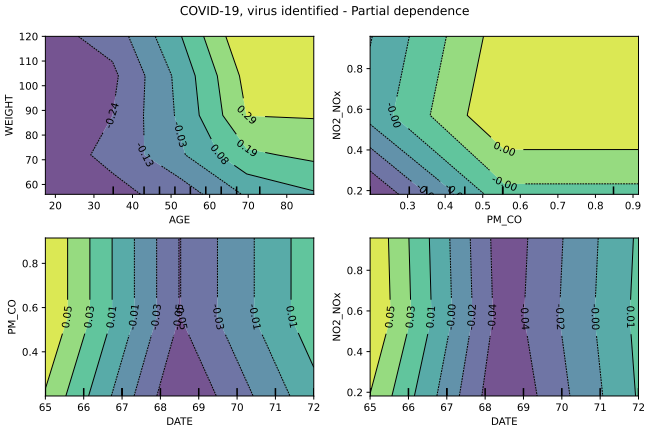

Fig. 5. COVID-19, virus identified - Nonlinear effects on severity for pairs of variables

average value. The analysis of the variable DATE shows an increase in severity in recent years and a slight correlation of that increased effect with exposure to (NO_2, NO_x). The main AP are consistent with the literature (see [1]).

B. COVID-19, virus identified

For COVID-19 the greatest effect on severity came from age and weight (Fig. 4). As for AP factors, (PM, CO) were the most important.

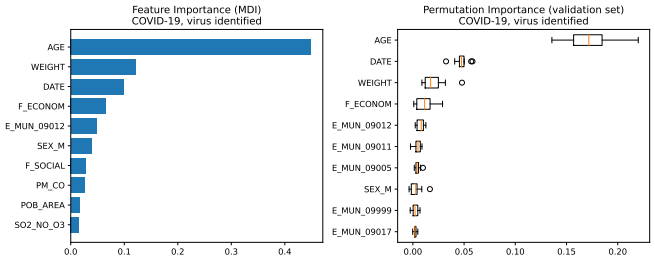

Fig. 4. COVID-19, virus identified - Feature importance.

As for the interaction between variables (Fig. 5), the higher the age and weight, the faster the severity increases, an interaction between contaminants is also observed, being higher the severity at higher exposures in conjunction of (PM, CO) and (NO_2, NO_x).

The highest severity is shown in the first months of hospitalizations. In the last months of 2020, an increase in the effect of (PM, CO) is observed in localities with lower values of this factor, indicating a possible higher sensitivity.

C. Diabetes mellitus

Fig. 6 shows that age, weight, and date are among the most important variables for diabetes mellitus. The most relevant AP factor is (NO_2, NO_x), which is consistent with other studies (see [1], [2], [13], [15], [20]).

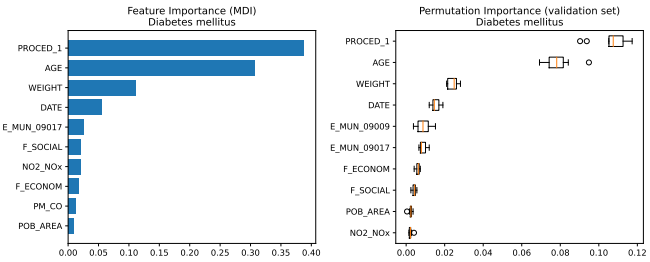

Fig. 6. Diabetes mellitus - Feature importance.

As for the nonlinear effects (Fig. 7), although low weight increases severity, the largest effect is found for high weight. For the joint effect of (NO_2, NO_x) and (PM, CO), (NO_2,

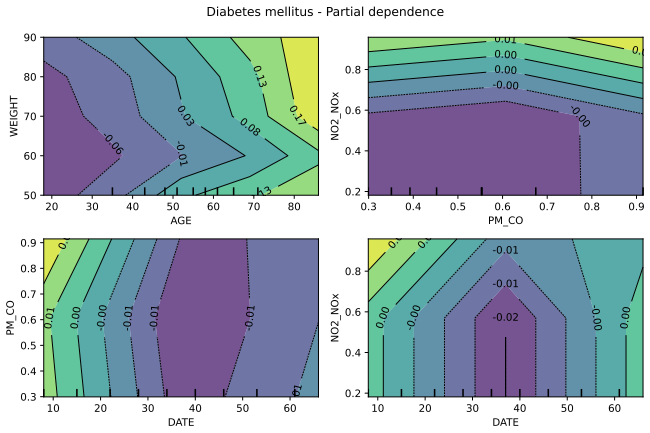

Fig. 7. Diabetes mellitus - Nonlinear effects on severity for pairs of variables

NO_x) dominates, and only for high (PM, CO) exposures does severity increase.

D. Heart diseases

In Fig. 8, for age and weight, we can differentiate two weight ranges that have lower severity, presumably corresponding to the male and female gender, where values below and above these weight values increase the severity of hospitalization.

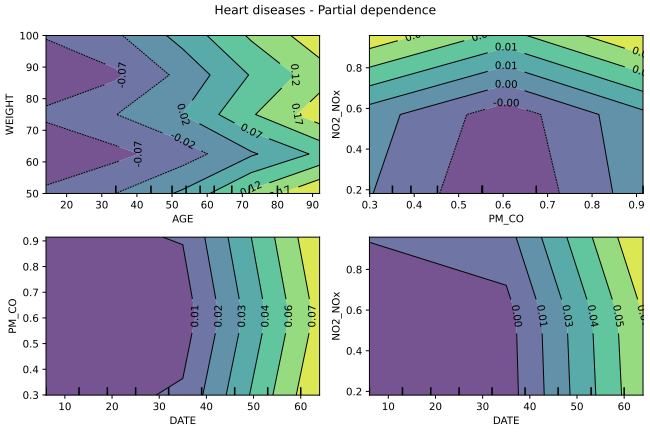

Fig. 8. Heart diseases - Nonlinear effects on severity for pairs of variables

From Table I, it can be observed a contribution to the severity of all the AP, and in the nonlinear effect, it is (NO_2, NO_x) that is shown to be the most relevant. In addition, an increase in severity is observed in recent years.

E. Liver diseases

Among the causes of hospitalization studied, Liver disease is the one that shows more nonlinear effects (Fig. 9), having an increase in sensitivity of (PM, CO) with respect to time (DATE) being a higher risk over the years for high concentrations and having greater resistance to low concentrations. A change in sensitivity is also observed when (PM, CO) and (NO_2, NO_x) have high exposures and a higher risk in the older and underweight population.

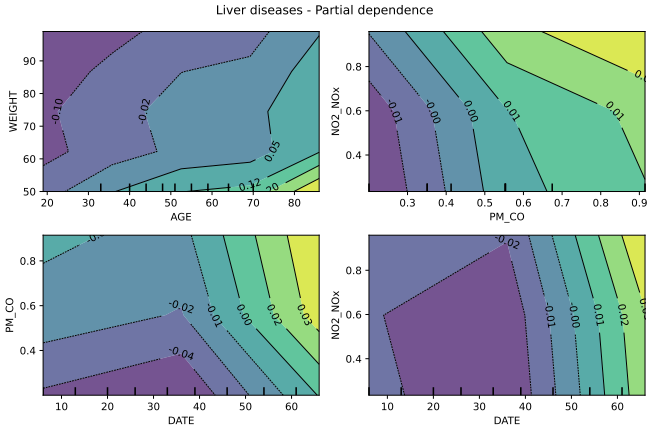

Fig. 9. Liver diseases - Nonlinear effects on severity for pairs of variables

F. Influenza and Pneumonia

It is observed from the results (Fig. 10) that Influenza and Pneumonia are the cause of hospitalization where age is more relevant in terms of severity, with a significantly higher value than the rest.

For the AP, it is clearly observed that exposure to (PM, CO) is the only one that shows relevant effects. Additionally, among the causes of hospitalization studied, this is the only one that has shown a decrease in severity with respect to time.

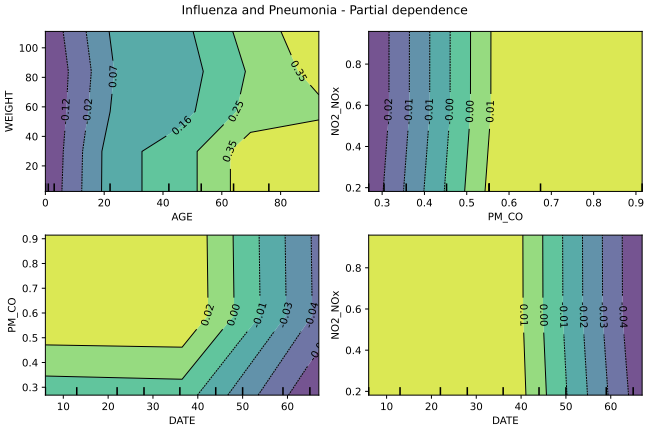

Fig. 10. Influenza and Pneumonia - Nonlinear effects on severity for pairs of variables

G. Cerebrovascular diseases

As for Cerebrovascular diseases, severity is affected by a low weight and a nonlinear effect of the interaction of (PM, CO) and (NO_2, NO_x) is observed (Fig. 11), where the increased exposure of (PM, CO) at low concentrations is more relevant, but at high exposures shows a greater effect of (NO_2, NO_x). An increase in severity over time is also

observed, but with a greater sensitivity to (NO_2, NO_x) than that observed for (PM, CO).

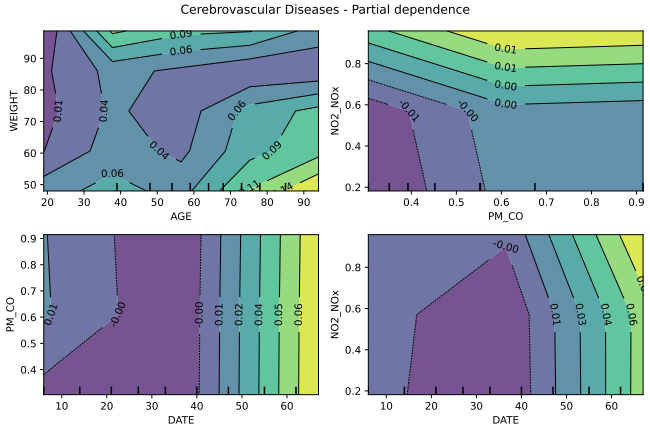

Fig. 11. Cerebrovascular diseases - Nonlinear effects on severity for pairs of variables

IV. Conclusion

Although the effects of the different variables influencing a disease are typically modeled with statistical methods, other models such as GBM allow for the exploration of highly nonlinear relationships and interactions between variables previously suspected or even unknown.

If a researcher is interested in modeling the relevant effects, the knowledge acquired through this type of model can be incorporated into the statistical models to validate the results. Instead, if the interest is prediction, models such as GBM can be more accurate by incorporating different effects simultaneously.

For future work, we plan to include these nonlinear effects in the statistical models and contrast the results when only linear effects are used.

Acknowledgment

The authors thank the Mexican National Council of Science and Technology (CONACYT) which made this research possible by providing funding through the project 7051 “Data observatory for discoveries of social-spatial-temporal patterns in health, mobility and air quality” and to the Ministry of Health of Mexico City (SEDESA) for providing their data and knowledge.

References

[1] Baris Afsar, Rengin Elsurer Afsar, Asiye Kanbay, Adrian Covic, Alberto Ortiz, and Mehmet Kanbay. Air pollution and kidney disease: review of current evidence. *Clinical Kidney Journal*, 12(1):19–32, 11 2018.

[2] Zorana J. Andersen, Ole Raaschou-Nielsen, Matthias Ketzel, Steen S. Jensen, Martin Hvidberg, Steffen Loft, Anne Tjønneland, Kim Overvad, and Mette Sørensen. Diabetes Incidence and Long-Term Exposure to Air Pollution: A cohort study. *Diabetes Care*, 35(1):92–98, 12 2011.

[3] Arleen F. Brown, Susan L. Ettner, John Piette, Morris Weinberger, Edward Gregg, Martin F. Shapiro, Andrew J. Karter, Monika Safford, Beth Waitzfelder, Patricia A. Prata, and Gloria L. Beckles. Socioeconomic Position and Health among Persons with Diabetes Mellitus: A Conceptual Framework and Review of the Literature. *Epidemiologic Reviews*, 26(1):63–77, 07 2004.

[4] Bert Brunekreef and Stephen T Holgate. Air pollution and health. *The Lancet*, 360(9341):1233–1242, 2002.

[5] CONEVAL. Índice de Rezago Social (IRS), 2020. https://www.coneval.org.mx/Medicion/IRS/Paginas/Indice_Rezago_Social_2020.aspx, 2021. [Online; accessed 17-July-2022].

[6] EvaluaCDMX. Índice de Desarrollo Social de la Ciudad de México, 2020. https://evalua.cdmx.gob.mx/principales-atribuciones/medicion-del-indice-de-desarrollo-social-de-las-unidades-territoriales/medicion-del-indice-de-desarrollo-social-de-las-unidades-territoriales, 2021. [Online; accessed 17-July-2022].

[7] Francesco Forastiere, Massimo Stafoggia, Carola Tasco, Sally Picciotto, Nerina Agabiti, Giulia Cesaroni, and Carlo A. Perucci. Socioeconomic status, particulate air pollution, and daily mortality: Differential exposure or differential susceptibility. *American Journal of Industrial Medicine*, 50(3):208–216, 2007.

[8] GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. *Lancet*, 396(10258):1223–1249, October 2020.

[9] Anjum Hajat, Charlene Hsia, and Marie S. O'Neill. Socioeconomic disparities and air pollution exposure: a global review. *Current Environmental Health Reports*, 2(4):440–450, Dec 2015.

[10] Anjum Hajat, Richard F. MacLehose, Anna Rosofsky, Katherine D. Walker, and Jane E. Clougherty. Confounding by socioeconomic status in epidemiological studies of air pollution and health: Challenges and opportunities. *Environmental Health Perspectives*, 129(6):065001, 2021.

[11] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367, 2008. Proceedings of the 4th International Workshop on Biomonitoring of Atmospheric Pollution (With Emphasis on Trace Elements).

[12] Guilan Kong, Ke Lin, and Yonghua Hu. Using machine learning methods to predict in-hospital mortality of sepsis patients in the icu. *BMC Medical Informatics and Decision Making*, 20(1):251, Oct 2020.

[13] Yongze Li, Lu Xu, Zhongyan Shan, Weiping Teng, and Cheng Han. Association between air pollution and type 2 diabetes: an updated review of the literature. *Ther. Adv. Endocrinol. Metab.*, 10:2042018819897046, December 2019.

[14] Cong Liu, Renjie Chen, Francesco Sera, Ana M. Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline S.Z.S. Coelho, Paulo H.N. Saldiva, Eric Lavigne, Patricia Matus, Nicolas Valdes Ortega, Samuel Osorio Garcia, et al. Ambient particulate air pollution and daily mortality in 652 cities. *New England Journal of Medicine*, 381(8):705–715, 2019. PMID: 31433918.

[15] S A Meo, A N Memon, S A Sheikh, F A Rouq, A Mahmood Usmani, A Hassan, and S A Arian. Effect of environmental air pollution on type 2 diabetes mellitus. *Eur. Rev. Med. Pharmacol. Sci.*, 19(1):123–128, January 2015.

[16] Lynne C. Messer, Barbara A. Laraia, Jay S. Kaufman, Janet Eyster, Claudia Holzman, Jennifer Culhane, Irma Elo, Jessica G. Burke, and Patricia O'Campo. The development of a standardized neighborhood deprivation index. *Journal of Urban Health*, 83(6):1041–1062, Nov 2006.

[17] Carlos Minutti-Martinez, Magali Arellano-Vázquez, and Marlene Zamora-Machado. A hybrid model for the prediction of air pollutants concentration, based on statistical and machine learning techniques. *Lecture Notes in Computer Science*, 13068:252–264, 2021.

[18] RAMA. Automatic Air Quality Monitoring Network. http://www.aire.cdmx.gob.mx/default.php?opc=\%27aKBh\%27, 2022. [Online; accessed 17-July-2022].

[19] UNDP. Informe de Desarrollo Humano Municipal 2010-2015. https://www.undp.org/es/mexico/publications/idh-municipal-2010-2015, 2019. [Online; accessed 17-July-2022].

[20] Bo-Yi Yang, Shujun Fan, Elisabeth Thiering, Jochen Seissler, Dennis Nowak, Guang-Hui Dong, and Joachim Heinrich. Ambient air pollution and diabetes: A systematic review and meta-analysis. *Environmental Research*, 180:108817, 2020.

[21] Mandi Yu, Zaria Tatalovich, James T. Gibson, and Kathleen A. Cronin. Using a composite index of socioeconomic status to investigate health disparities while protecting the confidentiality of cancer registry data. *Cancer Causes & Control*, 25(1):81–92, Jan 2014.

[22] Zhongheng Zhang, Yiming Zhao, Aran Canes, Dan Steinberg, and Olga Lyashevska. Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine*, 7(7):152–152, April 2019.