Nonlinear models and machine learning techniques for accessing the effect of the gasoline/ethanol share on ozone concentration

William Nilson Amorim, Antonio Carlos Pedroso de Lima,
Julio M. Singer, Carmen D.S. André,
Departamento de Estatística, Universidade de São Paulo, Brazil
Maria de Fátima Andrade,
Departamento de Ciências Atmosféricas, Universidade de São Paulo, Brazil
Paulo H.N. Saldiva
Instituto de Estudos Avançados, Universidade de São Paulo, Brazil

Bio-ethanol, a quasi-renewable fuel widely used in some countries as a cleaner option than gasoline, has been in the focus of the energy agenda since the European Commission stated that by 2050 the European Union should cut 20% of the greenhouse gas emissions relatively to 1990 levels and increase to 20% the amount of renewable energy used.

The ethanol/gasoline shift as vehicles fuel is directly associated with the atmospheric balance of nitrate oxides (NOx) and organic volatile composts (VOCs), since gasoline burning generates more NOx and ethanol evaporation and partial burning generates more VOCs. This balance is an important component to describe the tropospheric ozone formation along the day Madronich (2014).

In the past years, some have discussed the actual impact of shifting gasoline to ethanol as primary fuel in bi-fuel light-duty vehicles. Salvo and Geiger (2014) showed that ozone concentration decreased as the share of bi-fuel vehicles burning gasoline rose from 14 to 76% in São Paulo, Brazil, analyzing data from 2008 to 2011, and shed a valid concern about whether ethanol is a safer substitute for gasoline with respect to its relation with ozone concentration.

Besides ozone, particulate matter has also been subject of many air pollution studies, mostly for the lack of regulation, lack of a safe threshold, and its effects on public health. Salvo et al. (2017) extended this work analyzing data from 2008 to 2013 and other pollutants, like fine particles. The authors reached the same conclusion for ozone concentration, but they observed that ambient number concentrations of 7?100nm diameter particles (ultra-fine particles) rise along with the use of gasoline.

This work has as primary goals (1) analyze if the association between ozone and the estimated proportion of vehicles burning gasoline is in fact linear and (2) investigate the association between mortality and the estimated proportion of vehicles burning gasoline. To reach this goals, we used more sophisticated non-linear models, such as generalized addictive models and random forests, as well the LIME method to interpret the results of the latter.

1 Pollution, weather and traffic data

The data used in this study was collected, organized and provide by Salvo et al. (2017). It consist of air pollution, meteorologic and traffic data, as well the *shareE25*, the estimated proportion of bi-fuel vehicles burning gasoline with 25% ethanol (E25) over pure ethanol (E100).

The *shareE25* variable was estimated using information on the price of ethanol at the pump and the motorist-level revealed-choice survey data (Salvo and Huse, 2013). Such values were estimated weekly for the entire city, implying that the proportion of bi-fuel vehicles running on E25 was the same for all the monitoring stations where the pollutants were recorded hourly.

The air pollution and meteorologic data was measured hourly in several monitoring stations along the city maintained by the environmental authority of the state of São Paulo (CETESB). The traffic data, also measured hourly, was provided by the city traffic authority (CET).

More information about the data can be found in supplementary information of Salvo and Geiger (2014) and Salvo et al. (2017).

2 Statistical analysis

To analyze the effect of the *shareE25* variation on the ozone levels, Salvo et al. (2017) used a linear regression model, which assumes that the relation between this to variables is the same for all the values of *shareE25*.

To investigate the linearity assumption, we applied two non-linear models on the exactly same variables considered by the authors. The first is the well established *generalized additive model* (Hastie et al., 2008), which allows one to associate each one of the predictors to the response variable using a non-linear function. The second one is a *bagging* tree model known as *random* *forest*. This model consider a high number of regression trees and average its results to make predictions. Due to the lack of interpretation, this type of predictive model is very little used in studies where the main goal is to make associations between variables. To work around this problem, we used a interpretation technique called LIME (Ribeiro et al., 2016).

More details about this models can be found in James et al. (2013).

Following Salvo et al. (2017), we used as response the daily 12pm to 7pm ozone concentration average, excluding the cold months from June to September. We have ozone data available from 12 monitoring stations in the city. We also considered the same predictors as the authors presented in the Table 1.

Table 1: Predictors considered for the ozone models

Type	Variables	Number of parameters
Ethanol	Estimated proportion of cars running by gasoline (E25)	1
Station	Monitoring station dummies.	11
Calendar	Dummies for day of week, week of year, vacation periods and public holidays.	44
Trend	General and station specific trend term.	12
Climate	Temperature, solar radiation, humidity, wind speed and dummies for precipitation and thermal inversion..	9
Traffic	Dummies for station specific and city vehicle traffic intensity, as well inauguration of important beltway.	18
Total	**16 predictors + intercept term**	**96 parameters***

*95 parameters from predictors + 1 parameters from intercept term.

The model considered by Salvo et al. (2017) estimated a coefficient of -16.66 ± 10.01 to the *shareE25* variable, suggesting that the increase of the proportion of cars burnning E25 in the city is associated with the decrease of ozone levels. To compare the models, we used the root mean square error estimated by cross validation. We also compute the proportion of variance explained for each model. For the model used by Salvo et al. (2017), these two quantities were, respectively, 19.74 and 70.65%.

3 Results

We fitted three generalized additive models, the first using the Normal distribution, the second using the Gamma distribution and the last using the Inverse Normal distribution. The non-linear functions were assign to all numeric predictors: *shareE25*, temperature, radiation, humidity, wind speed and trend. The other predictors were linear associated with the response. Smoothing splines were used to estimate the functions and the smoothing degree was chosen by cross validation. In the three models, the *shareE25* variable was considered statistically significant to explain the ozone concentration. The performance of each model is described in the Table 2.

Table 2: Distribution used, root mean square error (RMSE), proportion of the variance explained and five more important variables for the generalized additive models.

Distribution	RMSE	% var. explained	More important variables
Normal	19.82	70.50	Temperature, wind speed, humidity, radiation e trend
Gamma	20.07	69.50	Temperature, wind speed, humidity, radiation e trend
Inverse Normal	29.28	45.30	Temperature, radiation, humidity, wind speed e trend

The results from the Normal and Gama models were close to those for the linear regression fitted by Salvo et al. (2017), while the Inverse Normal model got bigger RMSE and lower R^2, indicating that this distribution is not appropriated to fit the data. Contrary to the linear regression model, the trend term was held as one of the five most important variables to explain the variability of the ozone concentration in these three models. It presumably happened due to the flexibility of the GAM to represent the non-linearity of this temporal component. The *shareE25* variable was the 13th more important[1] in the Normal model, and the 9th in the Gama model and Inverse Normal model.

The usual method to interpret the generalized additive models is to plot each predictor by its estimated non-linear function. Figure 1 shows this plot for the *shareE25* variable, suggesting a non-linear relation between this variable and the concentration of ozone: positive in the extremes (10% - 20% and 60% - 80%) and negative in the center (20% - 60%).

[1] The variable importance measure for these models is the reduction in the generalized

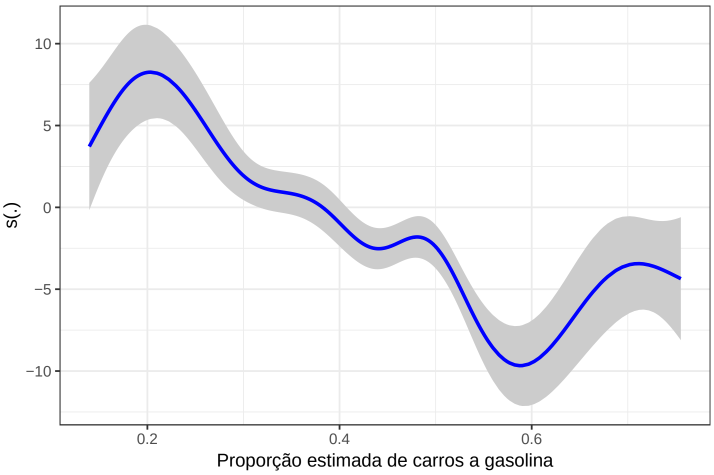

Figure 1: Non-linear function estimated by the Normal generalized additive model to the *shareE25* variable. The gray area represents a 95% confidence interval.

To evaluate the variability of the estimates, we also followed the strategy in Salvo et al. (2017) and fitted the Normal GAM (which had the best performance) in 200 bootstrap samples. Figure 2 shows the 200 estimated curves (in gray) for the *shareE25* variable in the bootstrap models and the cubic splines smoothed curve (in blue). The non-linear pattern found is the same as described earlier. The high variability in the extremes is result of few data with low or high estimated proportion of cars burning E25.

The table 3 shows the results of the random forest e XGBoost models. As expected we obtained lower test errors (14.11 and 12.24) and higher R^2 (85.72 and 88.56) than the linear regression and the additive models. The five most important predictors for both models were: temperature, humidity, radiation, wind speed e trend, just like in the GAM. The *shareE25* was the 6th most important variable.

Although we have more precise fits, we can not directly access the relation between the *shareE25* and the ozone concentration, i.e., we don't know if this predictor is statistically significant to explain the variability of the ozone levels and the direction of this supposed association. To solve this problem, we used a graphic technique called *accumulated local effects* (ALE) plot (Apley, 2016).

The ALE plot for the random forest in the left panel of the Figure 3 shows a *shareE25* effect very similar of the one suggest by the additive model.

cross-validation estimate of error when each predictor's feature is added to the model.

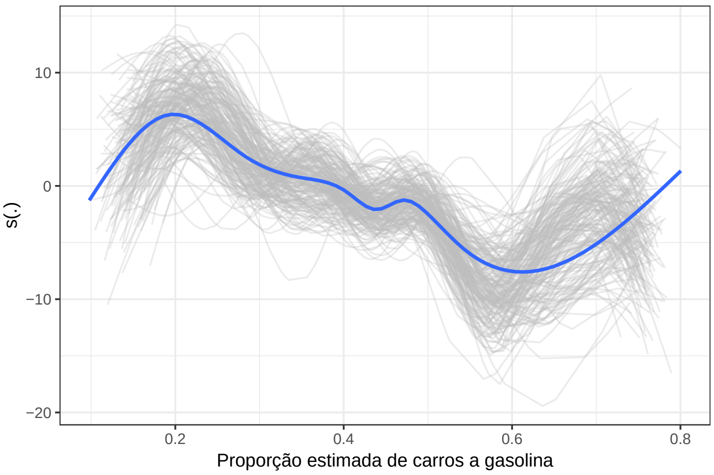

Figure 2: In gray, the 200 estimated curves for the *shareE25* variable in the bootstrap models. In blue, the cubic splines smoothed curve.

Table 3: Root mean square error (RMSE), proportion of the variance explained and five most important variables for the random forest and the XGBoost models. The hyperparameters of this models were set using 5-fold cross validation.

Model	RMSE	% var. explained	Most important variables
Random forest	14.11	85.72	Temperature, humidity, radiation, wind speed and trend
XGBoost	12.24	88.56	

However, the ALE plot for the XGBoost (in the right panel) does not show a clear relation between the variables. Moreover, it suggests now that the variables could be positively associated.

To evaluate the interpretations suggested by the ALE plots, the Figure 4 shows these plots for the climates variables: temperature, humidity, radiation and wind speed. The interpretation found for each predictor is plausible with the available knowledge about the tropospheric ozone formation: temperature and radiation are positively associated while wind speed and humidity are negatively associated.

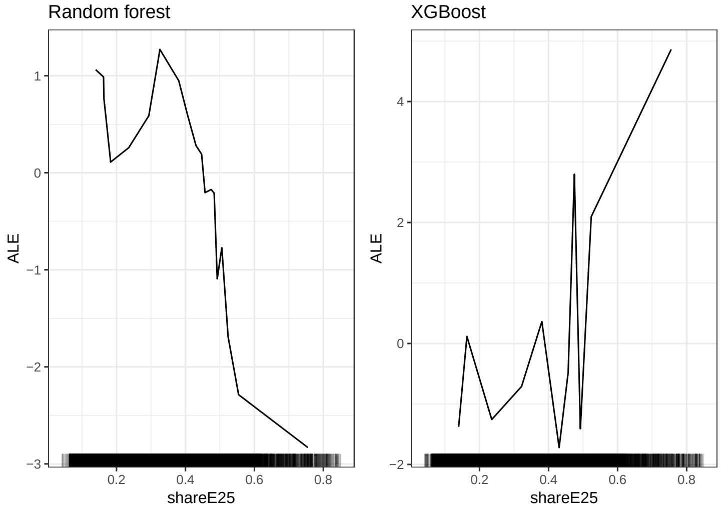

Figure 3: ALE plots of the *shareE25* predictor for the random forest and XGBoost. The y-axis gives us the main effect of the *shareE25* at a certain value compared to the average prediction of the data.

4 Discussion

The analysis conducted in this work suggests that the relation between the ozone concentration in the city of Sao Paulo and the *shareE25* is highly non-linear, and then it is not appropriate to use a linear model on the attempt to explain it.

All the non-linear models apply to the data indicated a non-linear association, with abrupt increases and decreases of the effect of the *shareE25* in the ozone concentration. Also, different models point in different directions: while the additive models and the random forest suggested a general negative association, the XGBoost suggested a positive association.

The findings here advocate for two hypotheses: (1) the relation between the use of ethanol/gasoline and ozone formation is too complex and cannot be fitted by the models considered and/or it does not have a practical explanation; or (2) the *shareE25* variable does not quite capture the real proportion of cars burning gasoline in the city and the associations found by the models are spurious.

Supporting the first hypothesis is known complex process behind the tropospheric ozone formation and the non-linear effect suggest by the models. The second hypothesis is supported by the different results found throughout the analysis and the possible measurement error in the *shareE25* variable and the strong supposition of considering only one measure for the whole city.

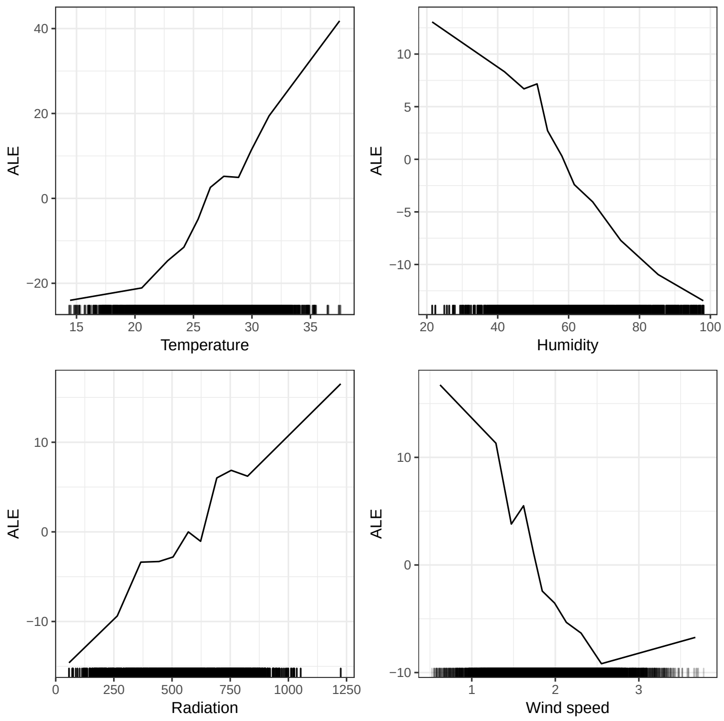

Figure 4: ALE plots of the climate predictors for the XGBoost. The y-axis gives us the main effect of these variables at a certain value compared to the average prediction of the data.

Considering also the practical relevance of the study, a decrease of 8.3 $\mu g/m^3$ in the daily mean ozone concentration pointed by Salvo et al. (2017) when *shareE25* raises from 30% to 80% would decrease the mean ozone concentration found in the sample only from 72.2 to 63.9 $\mu g/m^3$. Even if the linear model is in fact expressing the true relation between these variables, while a very important finding, it could not be sufficient to motivate public policies to reduce ethanol emissions. Furthermore, the impact of a average 8.3 $\mu g/m^3$ decrease in ozone levels in morbidity and mortality must be evaluated to better access this issue.

References

S. Madronich, *Ethanol and ozone*, Nature Geoscience: news & views **7** (2014), 395–397.

A. Salvo and F. M. Geiger, *Reduction in local ozone levels in urban São Paulo* *due to a shift from ethanol to gasoline use*, Nature Geoscience **7** (2014), 450–458.

A. Salvo, J. Brito, P. Artaxo, and F. M. Geiger, *Reduced ultrafine particle levels in São Paulo's atmosphere during shifts from gasoline to ethanol use*, Nature Communications **8** (2017), 1–14.

A. Salvo and C. Huse, *Build it, but will they come? Evidence from consumer choice between gasoline and sugarcane ethanol*, Journal of Environmental Economics and Management (2013), 251–279.

T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, 2008.

M. T. Ribeiro, S. Singh, and C. Guestrin, *"Why should I trust you?" Explaining the predictions of any classifier*, arXiv:1602.04938v3 [cs.LG] (2016).

G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer Series in Statistics, Springer, New York, 2013.

D. Apley, *Visualizing the effects of predictor variables in black box supervised learning models*, arXiv (2016).

Acknowledgements

This research received financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant 3304126/2015-2) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant 2013/21728-2), Brazil.

Author contributions

W.N.A., A.C.P.L., J.M.S. and C.D.S.A. analyzed the data; M.F.A. contributed with climate and pollutant chemistry knowledge; P.H.N.S. suggested the theme; all authors wrote the paper.

Additional information

Correspondence and request for materials should be addressed to W.N.A. All the figures and the R codes used in the statistical analysis may be obtained respectively at http://bit.do/amorim_et_al_figures and http://bit.do/amorim_et_al_codes.

Competing financial interests

The authors declare no competing financial interests.