Nonlinear analysis of 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 $\pm$ 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: Resultado dos modelos aditivos generalizados utilizados para ajustar os dados de Salvo et al. (2017).

Distribuição	RMSE	% var. explicada	Variáveis mais importantes
Normal	19.82	70.50	Temperatura, vento, umidade, radiação e tendência
Gama	20.07	69.50	Temperatura, vento, umidade, radiação e tendência
Normal inversa	29.28	45.30	Temperatura, radiação, umidade, vento e tendência

Os resultados dos modelos Normal e Gama ficaram muito próximos do modelo de regressão linear ajustado pelos autores. Já o modelo Normal Inversa se mostrou inferior, mostrando que essa distribuição (com função de ligação $1/\mu^2$) não é adequada aos dados. Observamos que, para esses modelos, a tendência entrou como uma das cinco variáveis mais importantes para explicar a variabilidade da concentração de ozônio, o que não aconteceu para o modelo de regressão linear. Isso provavelmente se deve pela maior flexibilidade que os modelos aditivos possuem para representar a não-linearidade desse componente temporal. No modelo com distribuição normal, a proporção estimada de carros rodando a gasolina foi considerada a décima terceira mais importante, no modelo Gama foi a nona mais importante e no modelo Normal inversa foi a nona mais importante.

A maneira usual de interpretar os modelos aditivos generalizados é construir gráficos de cada preditor pela sua função não-linear estimada[1]. Na Figura 1, apresentamos esse gráfico para o preditor referente à proporção

[1] Os preditores que entraram no modelo de forma linear podem ser interpretados de maneira análoga a um modelo de regressão linear.

estimada de carros à gasolina utilizando o modelo com distribuição Normal. O gráfico indica que conforme a proporção de carros à gasolina aumenta, a concentração de ozônio tende a diminuir, mesma conclusão encontrada no modelo de regressão linear.

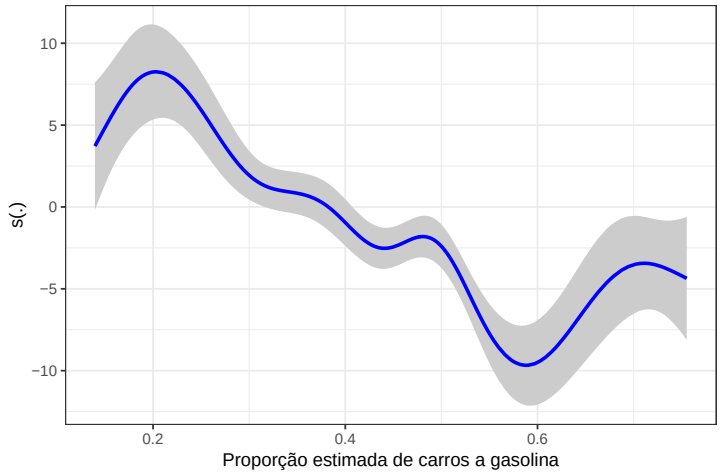

Figure 1: Função não-linear estimada pelo modelo aditivo generalizado com distribuição Normal para a proporção estimada de carros rodando a gasolina. A área cinza em volta da curva representa o intervalo de confiança com 2 erros-padrão para cima e para baixo.

Para avaliar a variabilidade das estimativas, criamos 200 amostras de *bootstraping* e ajustamos o modelo aditivo generalizado com distribuição Normal (que apresentou o melhor desempenho) para cada uma delas. Na Figura 2, apresentamos o resultado para a função estimada da variável referente à proporção estimada de carros à gasolina. As curvas cinzas são as 200 funções estimadas, uma para cada amostra de *bootstrapping*, e representam a variabilidade da função apresentada na Figura 1. A curva azul é a curva suavizada por *splines* cúbicos. Podemos notar que a tendência de diminuição da concentração de ozônio conforme a proporção de carros a gasolina aumenta é consistente para o modelo aditivo generalizado. Podemos observar também que há uma maior variabilidade nos extremos desse preditor. Isso acontece porque temos poucos dias nos quais a proporção de carros a gasolina foi estimada muito baixa ou muito alta (Figura **??**).

Em busca de resultados mais precisos, ajustamos também uma *random forest* aos dados. Os resultados estão resumidos na Tabela 3.

Observamos que a *random forest* apresentou um menor erro de teste (RMSE = 14.11) do que o modelo de regressão linear, além de explicar uma maior porcentagem da variação da concentração de ozônio. Os cinco

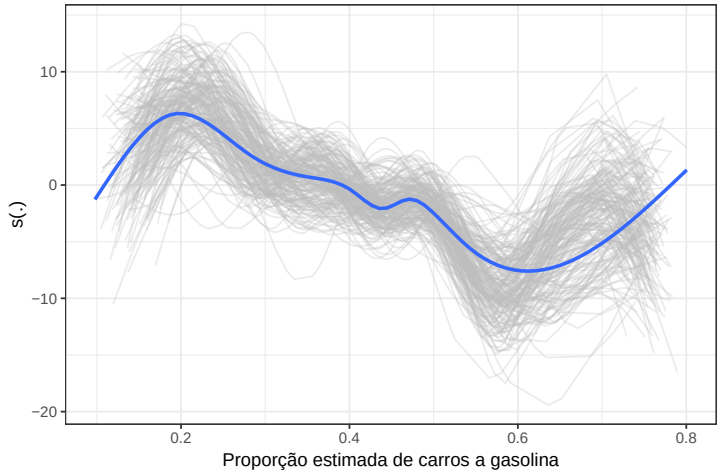

Figure 2: Em cinza, as funções estimadas da variável referente à proporção estimada de carros à gasolina para cada uma das 200 amostras de *bootstrapping*. Em azul, a curva suavizada por *splines* cúbicos.

Table 3: Resultado do modelo *random forest* aplicado aos dados de Salvo et al. (2017). Os hiperparâmetros referentes ao tamanho mínimo de cada nó e o número de preditores sorteados em cada amostra foram definidos por validação cruzada.

Tamanho mínimo dos nós	Número de preditores em cada amostra	RMSE	% var. explicada	Variáveis mais importantes
1	48	14.11	85.72	Temperatura, umidade, radiação, vento e tendência

preditores mais importantes foram temperatura, umidade, radiação, vento e tendência, iguais aos encontrados nos modelos aditivos generalizados. A proporção estimada de carros a gasolina foi o sexto preditor mais importante.

Apesar de termos um ajuste mais preciso, a partir deste modelo não conseguimos interpretar a diretamente a relação entre a proporção estimada de carros a gasolina e a concentração de ozônio. Não sabemos se esse preditor é estatisticamente significativo e, em caso positivo, em qual direção ela está associada à resposta. Sem essa interpretação, não conseguimos responder a pergunta de interesse do estudo.

Uma maneira alternativa de interpretar os dados é utilizar o LIME, discutido na Seção **??**. Como exemplo, vamos avaliar as predições para os 100 dias com maiores níveis de ozônio e para os 100 dias com menores dias de

ozônio. A Figura 3 sugere que, para os dias com maiores médias de ozônio, o efeito protetor da proporção estimada de carros à gasolina acontece com maior frequência quando temos cerca de 50% da frota ou mais rodando a gasolina. O mesmo acontece para os dias com menores médias. Esse comportamento reforça a relação não-linear encontrado evidenciada pelo modelo aditivo generalizado ajustado na seção anterior.

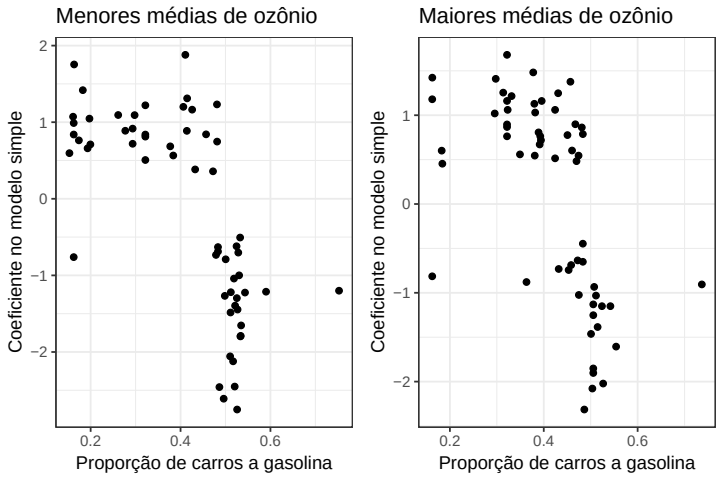

Figure 3: Proporção estimada de carros gasolinas contra o coeficiente estimado para esse preditor no modelo simples (regressão *ridge*). À esquerda, utilizamos os 100 dias com as menores médias de ozônio e, à direita, os 100 dias com as maiores médias.

Como a suposição de que a proporção de carros a gasolina não está associada diretamente a nenhum outro preditor, podemos predizer o ozônio substituindo o verdadeiro valor da proporção por um valor hipotético, que simule um cenário com poucos carros a gasolina rodando na cidade e outro com muito carros a gasolina. Na Figura 4, apresentamos as curvas suavizadas do ozônio predito para esses dois cenários, além do cenário efetivamente observado. Os valores fixados para a proporção de carro a gasolina foram: 20% para representar o cenário com poucos carros a gasolina e 70% para o cenário com muitos carros. Esses valores foram escolhidos com base na distribuição da variável original. Observamos pelo gráfico que, os cenários observado e com baixa proporção são bem parecidos, enquanto o cenário com alta proporção apresenta menores concentrações em algumas estações.

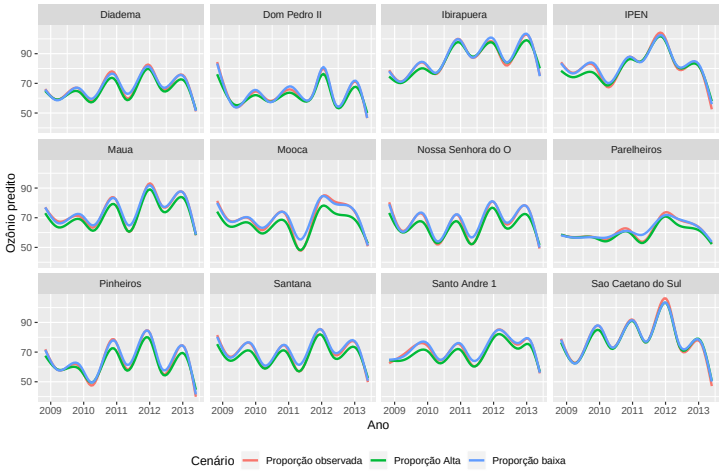

Figure 4: Curvas suavizadas do ozônio predito para os cenários observado, com alta proporção de carros a gasolina (70% durante todo o período) e baixa proporção de carros a gasolina (20% durante todo o período).

4 Discussion

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.

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.