Estratégias para análise de dados de poluição do ar

William Nilson de Amorim

Projeto de pesquisa

Programa: Doutorado em Estatística
Orientador: Prof. Dr. Antonio Carlos Pedroso de Lima
Coorientador: Prof. Dr. Julio da Motta Singer

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES e do CNPq.

São Paulo, 2 junho de 2017

Introdução

A poluição do ar, considerada pela Organização Mundial da Saúde (OMS) o maior risco ambiental à saúde humana, é responsável por aproximadamente 7 milhões de mortes por ano, um oitavo do total global (Jasarevic *et al.*, 2014). Poluentes como óxidos de carbono, nitrogênio e enxofre, ozônio e material particulado trazem diversos prejuízos à nossa qualidade de vida e ao equilíbrio do planeta. Eles são agentes sistemáticos no desenvolvimento de irritação dos olhos, obstrução nasal, tosse, asma e redução da função pulmonar. À exposição contínua, estão associadas diversas doenças respiratórias e cardiovasculares, problemas digestivos e do sistema nervoso, câncer e aumento da mortalidade infantil (European Commission, 1999). Além disso, vários deles estão diretamente ligados ao aquecimento global e ao efeito estufa.

Nas últimas décadas, diversos estudos têm alertado sobre os riscos da poluição atmosférica. Seus principais objetivos compreendem a descrição dos níveis de poluição em determinada região, o acompanhamento das concentrações dos poluentes ao longo do tempo, a busca de relação entre a mortalidade (ou morbidade) e a concentração de poluentes e o desenvolvimento de soluções mais limpas (ou menos poluentes) economicamente viáveis.

Carslaw *et al.* (2007), por exemplo, modelaram concentrações diárias de óxidos e dióxidos de nitrogênio, monóxido de carbono, benzeno e 1,3-butadieno para avaliar a tendência das concentrações desses poluentes durante o período de 1998 a 2005 no movimentado centro de Londres. Já Beer *et al.* (2011) avaliaram os impactos à saúde ao se utilizar etanol como aditivo na gasolina em regiões urbanas da Austrália. Os autores examinaram emissões de escapamento e evaporativas de veículos leves em câmaras de poluição, medindo níveis de ozônio, dióxido de nitrogênio e material particulado. Kloog *et al.* (2012) utilizaram medidas de profundidade óptica de aerossóis feitas por satélites para prever concentrações diárias de material particulado na costa leste dos Estados Unidos. Belusic *et al.* (2015) estudaram a relação das concentrações horárias de monóxido de carbono, dióxido de enxofre, dióxido de nitrogênio e material particulado com as condições climáticas da cidade de Zagrebe, na Croácia.

Os estudos citados — e muitos outros, como Chang *et al.* (2017), Conceição *et al.* (2001a,b), Lin *et al.* (1999), Schwartz *et al.* (1996), Katsouyanni *et al.* (1996), Schwartz (1994, 1996), Saldiva *et al.* (1994, 1995), Schwartz e Dockery (1992), Schwartz e Marcus (1990) e Shumway e Stoffer (1982) —, embora abordem temas diferentes, concordam sobre a importância da diminuição da emissão de poluentes. Devido à forte dependência de combustíveis fósseis, o setor de transportes é considerado pela União Europeia o mais resiliente aos esforços para a redução de emissões (European Commission, 2011). Como soluções que visam controlar o tamanho da frota de veículos ou restringir o seu uso são limitadas por fatores políticos e econômicos, os estudos nessa área buscam encontrar combustíveis menos poluentes, alternativas ao diesel e à gasolina.

Em um experimento controlado na cidade de Fairbanks, no Alasca, Mulawa *et al.* (1997) coletaram amostras de material particulado de carros a gasolina e as compararam com dados de emissões de carros abastecidos com E10 (gasolina com 10% de álcool). Os autores constataram que os carros com E10 emitiam menos material particulado e que os níveis desse poluente aumentavam conforme a temperatura dos dois combustíveis diminuía. Yoon *et al.* (2009) conduziram uma investigação similar e concluíram que a combustão de etanol e da mistura E85 (85% etanol e 15% gasolina) emitiam concentrações inferiores de hidrocarbonetos, monóxido de carbono e óxidos de nitrogênio quando comparados com a gasolina sem aditivos sob diversas condições experimentais. Já Pereira

et al. (2004) expuseram ao sol câmaras contendo etanol puro e gasool (mistura de 22-24% etanol em gasolina) para estudar a formação do ozônio e concluíram que as concentrações máximas do poluente eram, em média, 28% maiores para o álcool do que para o gasool.

As grandes cidades funcionam como um laboratório natural para os estudos de poluição do ar. Com a disponibilidade de dados meteorológicos e de tráfego, é possível avaliar grande parte dos fatores que influenciam a formação dos poluentes. Muitos trabalhos vêm trocando os ambientes controlados por dados de poluição urbana. Nesse contexto, Jacobson (2007) relacionou a substituição da gasolina por etanol (E85) com dados epidemiológicos em Los Angeles, em particular, e nos Estados Unidos como um todo. Os resultados do estudo mostraram que a utilização de E85 aumentou as taxas de mortalidade, hospitalização e asma devido a maiores concentrações de ozônio. Salvo e Geiger (2014) utilizaram uma mudança real na preferência por gasolina ocasionada em flutuações de larga escala no preço do etanol para analisar a associação entre as proporções de carros a gasolina rodando na cidade de São Paulo[1] com os níveis de ozônio medidos no começo da tarde durante os anos de 2008 a 2011. Os autores concluíram que o uso do etanol está associado a maiores concentrações do poluente.

Uma das grandes dificuldades associadas ao estudo de dados de poluição atmosférica é o grande número de efeitos confundidores. Em áreas urbanas, pode existir uma complexa mistura de fatores que contribuem de maneiras diferentes para a formação e dispersão dos poluentes. Diversas variáveis podem ser consideradas, como aquelas relacionadas ao clima, ao tráfego, à química atmosférica local, às mudanças climáticas sazonais, a eventos esporádicos que podem alterar o trânsito em algumas áreas da cidade, ao tamanho e idade da frota de veículos, às emissões evaporativas, entre outras. Além disso, a relação entre essas variáveis pode não ser muito simples, o que exige modelos menos restritivos, e não é rara a presença de dados omissos ou grande períodos sem observação, complicando ainda mais a análise. Embora diferentes técnicas estatísticas venham sendo empregadas na modelagem desses dados, nenhuma delas é robusta o suficiente para atacar sozinha todos esses problemas. Na maioria dos casos, a abordagem mais adequada dependeria de uma formulação de estratégias que envolvessem a combinação de duas ou mais técnicas, além de ferramentas estatísticas ainda não utilizadas nesse contexto.

O objetivo desta tese é comparar estratégias para analisar séries temporais de poluentes atmosféricos, avaliando o desempenho dos diferentes modelos na presença de variáveis com erro de medida, períodos sem informação, diferentes especificações para a distribuição dos dados e observações correlacionadas.

Técnicas estatísticas empregadas na análise e suas deficiências

As principais metodologias para abordar dados de poluição e epidemiologia ambiental envolvem modelos lineares (Saldiva *et al.*, 1995; Salvo e Geiger, 2014), modelos lineares generalizados (Conceição *et al.*, 2001b; Lin *et al.*, 1999; Saldiva *et al.*, 1994; Schwartz e Dockery, 1992), modelos aditivos generalizados (Carslaw *et al.*, 2007; Conceição *et al.*, 2001a,b; Schwartz *et al.*, 1996; Schwartz, 1994, 1996) e séries temporais (Katsouyanni *et al.*, 1996; Schwartz e Marcus, 1990; Shumway e Stoffer, 1982). A escolha de uma estratégia de análise adequada é muito importante, pois dados de poluição

[1] Proporção de carros a gasolina entre os carros bicombustíveis.

do ar[2] usualmente violam as suposições associadas a esses modelos. Salvo e Geiger (2014), por exemplo, utilizam um modelo linear, que supõe independência entre as observações, para ajustar uma série horária de concentração de ozônio, possivelmente autocorrelacionada. No trabalho, a variável explicativa de maior interesse, a proporção de carros a gasolina na cidade de São Paulo, é estimada, isto é, existe um erro associado a essas observações, tornando um modelo com erros nas variáveis mais adequado para essa análise. Além disso, a escolha dos autores por variáveis indicadoras para a hora, o dia da semana e a semana do ano pode não ser a mais adequada para controlar a sazonalidade da série. De uma forma geral, autocorrelação, heteroscedasticidade, superdispersão, tendência, sazonalidade, componentes espaciais, variáveis com erro de medida e grandes períodos sem observação, acarretando um grande número de observações omissas, são características comuns em dados de poluição do ar e precisam ser contempladas pelo modelo escolhido.

Modelos lineares são uma boa alternativa devido à facilidade de implementação e interpretação de seus coeficientes. Além disso, grandes intervalos sem observações não geram maiores problemas no processo de estimação. Por outro lado, a concentração de poluentes é uma medida positiva e, em muitos casos, assimétrica e autocorrelacionada, tornando as suposições de normalidade e de independência muito restritivas. Os modelos lineares generalizados flexibilizam a suposição de normalidade e de homoscedasticidade, permitindo modelar também a dispersão dos dados, mas ainda estão restritos a observações independentes. Em geral, os efeitos temporais são representados por variáveis indicadoras para a hora do dia, dia da semana, semana do ano etc. Nessa abordagem, muitas vezes é difícil especificar quais termos devem ser incluídos no modelo e determinar se eles realmente controlam esses componentes.

Os modelos aditivos generalizados são uma boa alternativa nesse contexto. Eles permitem incluir um componente não-paramétrico para ajustar uma curva suavizada da resposta em função do tempo, controlando os efeitos sazonais e de tendência, e um componente paramétrico para associar as variáveis explicativas à variável resposta. No entanto, a especificação do modelo pode ser complicada, pois a escolha dos parâmetros que controlam a suavização nem sempre é trivial. Grandes períodos sem observação também atrapalham o ajuste, pois a curva suavizada ao longo do tempo é considerada contínua durante todo o intervalo.

Os modelos de séries temporais também sofrem com dados omissos, mas são mais adequados para controlar a tendência e a sazonalidade, além de permitirem especificar o grau de autocorrelação entre as observações. Contudo, as técnicas usuais regridem a variável resposta em função apenas de suas observações defasadas, não sendo possível associar a ela fatores confundidores ou de controle. Dessa maneira, os resíduos dos modelos de séries temporais podem ser ajustados contra as variáveis explicativas a partir de um modelo linear sem a necessidade de controlar o efeito do tempo.

Modelos lineares

Seja Y_t a variável resposta sob estudo, $\mathbf{X_t} = (X_{1t}, ..., X_{pt})$ um vetor de variáveis explicativas cuja associação com Y_t estamos interessados em avaliar, $\mathbf{Z_t} = (Z_{1t}, ..., Z_{st})$ um vetor de variáveis de confundimento cuja associação com Y_t é, a priori, conhecida e t o instante no qual essas variáveis foram medidas, com $t = 1, ..., n$. Aqui não são feitas suposições sobre a natureza dos preditores $\mathbf{X_t}$

[2]Consideraremos tanto séries de poluentes quanto dados epidemiológicos, como número de mortes ou casos de doenças associadas à poluição do ar.

e $\mathbf{Z_t}$, isto é, essas variáveis podem ser fixas ou aleatórias, qualitativas ou quantitativas. Dado os vetores de parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ e $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_s)$, a formulação mais simples para um modelo linear é dada por

$$Y_t = \alpha + \beta_1 X_{1t} + ... + \beta_p X_{pt} + \gamma_1 Z_{1t} + ... + \gamma_s Z_{st} + \epsilon_t. \tag{1}$$

Em geral, supõe-se que os erros $(\epsilon_1, ..., \epsilon_n)$ sejam independentes, homoscedásticos e, em certos casos, normalmente distribuídos. Além disso, a especificação (1) impõe que a relação entre a resposta Y_t e os preditores $\mathbf{X_t}$ e $\mathbf{Z_t}$ seja linear e aditiva.

Os coeficientes $(\beta_1, ..., \beta_p)$ e $(\gamma_1, ..., \gamma_s)$ podem ser estimados pelo método de mínimos quadrados (Hastie *et al.*, 2008) ou de máxima verossimilhança (Sprott, 2000). Sob a suposição de normalidade, as duas abordagens são equivalentes. O método de mínimos quadrados não exige a distribuição Normal no processo de estimação, mas ela pode ser necessária para se fazer inferências sobre os parâmetros (intervalos de confiança e testes de hipóteses) em amostras pequenas[3].

Apesar de suas restrições, alguns autores optam por modelos lineares para analisar dados de poluição atmosférica. Saldiva *et al.* (1995) utilizaram um modelo linear para modelar a associação entre poluição do ar e mortalidade diária, justificando que a escolha do modelo gaussiano (em detrimento de um modelo Poisson) é razoável devido ao elevado número médio de mortes diárias (maior que 60). Salvo e Geiger (2014), como citado anteriormente, ajustaram um modelo linear para estudar a associação entre a proporção de carros a gasolina e a concentração de ozônio em São Paulo.

Como as estimativas do modelo são calculadas com base na suposição de independência, variáveis $Y_1, ..., Y_n$ muito correlacionadas podem levar a estimativas enviesadas e conclusões equivocadas. Mais especificamente, se as observações apresentarem correlação, os erros-padrão estimados tenderão a subestimar a verdadeira variabilidade, o que comprometeria toda a inferência feita a partir deste modelo, já que os valores p associados seriam menores do que deveriam ser. Variáveis indicadoras para a unidade de tempo em que as medidas foram realizadas (hora, dia da semana, semana do mês etc) podem ajudar a reduzir o efeito da correlação.

Em modelos lineares, uma forma de tratar observações heteroscedásticas é ajustar certas transformações da variável resposta Y, como $\log Y$ e $\sqrt{Y}$, que são boas tentativas para estabilizar a variância das observações. Observe que, nesse caso, a relação entre Y e os preditores deixa de ser linear, e a modelagem é toda feita com base na variável resposta, inclusive a interpretação dos parâmetros. Essa estratégia também pode ser utilizada quando a violação da suposição de normalidade é um problema (Hastie *et al.*, 2008). Uma outra alternativa consiste em ponderar as observações com pesos proporcionais ao inverso de sua variância, mas se limita aos casos em que a variabilidade é conhecida ou pode ser estimada com precisão. Fazendo o gráfico de resíduos para um dos modelos lineares ajustados por Salvo e Geiger (2014), observa-se pela Figura 1 que a variância não parece ser constante em todo o período observado.

A suposição de linearidade estabelece que a variação esperada em Y_t causada pelo acréscimo de uma unidade em X_{it} é constante e igual a β_i, independentemente do valor de X_{it}. O mesmo vale para as variáveis $\mathbf{Z_t}$ e os coeficientes $(\gamma_1, ..., \gamma_s)$, mas, em geral, não há interesse em interpretá-los. Para contornar problemas com a suposição de linearidade no modelo linear, costuma-se aplicar transformações nos preditores, tais como $\log(X)$, $\sqrt{X}$ e X^2. Quando há interesse inferencial, é

[3]Em geral, o tamanho amostral não é um problema em estudos de poluição do ar.

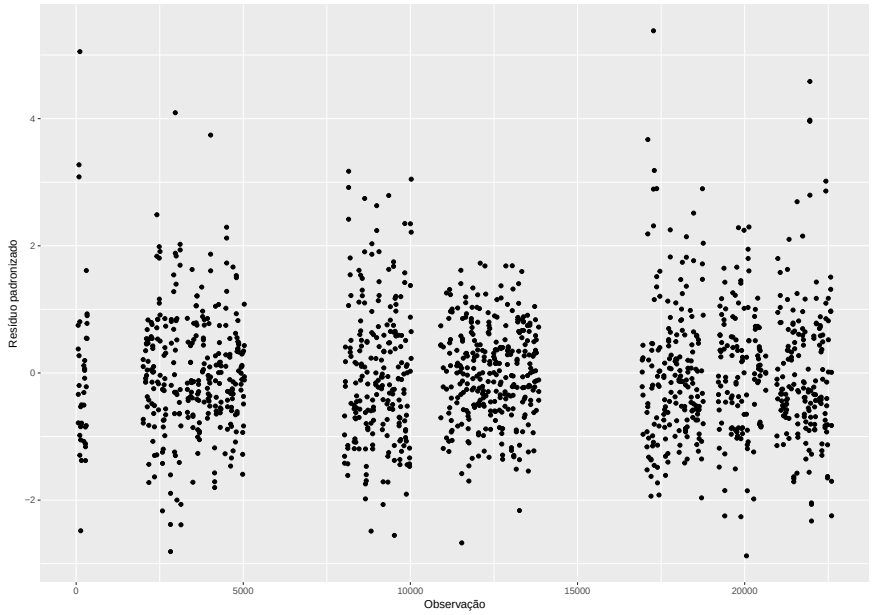

Figura 1: *Gráfico dos resíduos padronizados contra o índice das observações para o modelo VI ajustado por Salvo e Geiger (2014) para a estação Dom Pedro II. As faixas sem pontos se referem a períodos sem observação.*

preciso ter cuidado com a nova interpretação dos parâmetros após aplicar essas transformações. Já suposição de aditividade implica que a variação esperada na resposta Y_t causada por uma mudança no preditor X_{it} (ou Z_{jt}) independe do valor dos outros preditores. Essa suposição pode ser relaxada com a introdução de termos de interação.

É natural pressupor que séries de variáveis ambientais ou epidemiológicas não sejam estacionárias, isto é, que as suas médias não são constantes ao longo do tempo. Para acrescentar um termo de tendência linear ao modelo (1), podemos especificar $Z_{1t} = t$, $t = 1, ..., n$. Assim, um coeficiente γ_1 positivo indica que a resposta cresce linearmente com o tempo, enquanto um coeficiente negativo indica que a resposta decresce linearmente. Podemos definir outras formas para a tendência, como quadrática, $Z_{1t} = t^2$, ou logarítmica, $Z_{1t} = \log(t)$.

Uma outra forma de controle da tendência de uma série é transformar os dados originais para gerar uma série estacionária. A transformação mais comum é dada pela diferenciação

$$\Delta Y_t = Y_t - Y_{t-1}.$$

Caso o cálculo dessa primeira diferença não tenha sido suficiente para alcançar a estacionariedade, diferenças de maior grau costumam ser aplicadas. Para $j < n$, define-se

$$\Delta^j Y_t = \Delta[\Delta^{j-1} Y_t].$$

Usualmente, uma ou duas diferenças tornam a série estacionária (Morettin e Toloi, 2004).

A média da variável resposta também pode variar segundo fatores que se alteram periodicamente dentro de um intervalo de tempo. Esses padrões, conhecidos como sazonalidade, se repetem ao longo das semanas, meses e anos, e também precisam ser controlados. O termo sazonal pode ser incluído no modelo linear (1) a partir de variáveis indicadoras. Se, por exemplo, acredita-se haver um efeito sazonal de mês, pode-se então adicionar ao modelo 11 variáveis indicadoras Z_{it}, $i = 1, ..., 11$, tais

que

$$Z_{it} = \begin{cases} 1, & \text{se a observação } t \text{ pertence ao } i\text{-ésimo mês do ano; e} \\ 0, & \text{caso contrário.} \end{cases}$$

Com essa formulação, o mês de dezembro será tomado como referência.

Modelos lineares generalizados

O modelo linear definido em (1) tem três importantes restrições:

i. para se fazer inferência, principalmente em amostras pequenas, a variável Y precisa ter distribuição Normal;

ii. a variância de Y é considerada constante durante todo o período de observação; e

iii. as observações precisam ser independentes.

Formalmente, o modelo admite que $\epsilon_t \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$, $t = 1, ..., n$, isto é, os erros aleatórios são normalmente distribuídos, independentes e homoscedásticos[4].

Como discutido anteriormente, essas suposições podem não ser compatíveis com dados de poluição atmosférica, justificando a necessidade de modelos mais flexíveis, que contemplem a distribuição e a estrutura de variância e covariância das observações e continue interpretável.

Os modelos lineares generalizados foram introduzidos por Nelder e Wedderburn (1972) e englobam uma série de modelos de regressão. Eles são uma das principais alternativas para o ajuste de dados não-normais, permitindo a utilização de distribuições como a Gama e a Log-normal para dados assimétricos, a Poisson e a Binomial negativa para dados de contagem e Binomial para dados categorizados.

O modelo linear generalizado pode ser definido como

$$Y_t | \mathbf{X_t}, \mathbf{Z_t} \overset{ind}{\sim} \mathcal{D}(\mu_t, \phi)$$

$$g(\mu_t) = \alpha + \beta_1 X_{1t} + ... + \beta_p X_{pt} + \gamma_1 Z_{1t} + ... + \gamma_s Z_{st}, \tag{2}$$

em que $\mathcal{D}$ é uma distribuição pertencente à família exponencial, μ_t é um parâmetro de posição e ϕ é um parâmetro de precisão.

Os parâmetros deste modelo podem ser estimados por máxima verossimilhança, e os cálculos envolvem o uso de procedimentos iterativos, como o Newton-Raphson e escore de Fisher (Dobson, 1990).

Os modelos Gama e Log-normal são usualmente utilizados para ajustar dados positivos assimétricos, sendo, em geral, mais adequados para concentrações de poluentes do que a distribuição Normal. Como essas distribuições não assumem valores negativos, concentrações próximas de zero podem gerar problemas de convergência no algoritmo de estimação. Para ajustar um modelo Gama para os dados analisados por Salvo e Geiger (2014), por exemplo, é preciso fazer algumas transformação na série de ozônio para lidar com as concentrações muito baixas.

[4] A notação $X_i \overset{ind}{\sim} G$, $i = 1, ..., n$, indica que as variáveis $X_1, \ldots, X_n$ são independentes e todas possuem a mesma distribuição G.

Dados de contagem, como o número de casos de uma doença ou mortalidade, são usualmente ajustados pelo modelo Poisson. Conceição *et al.* (2001b), por exemplo, utilizaram esse modelo para avaliar a associação entre poluição atmosférica e marcadores de mortalidade em idosos na cidade de São Paulo. No entanto, a distribuição Poisson impõe que a média e a variância das observações são iguais e pode não se ajustar bem quando os dados apresentam sobredispersão (variância maior que a média). O modelo com resposta binomial negativa é uma alternativa nesses casos, já que permite a modelagem conjunta dos parâmetros de posição e dispersão.

A primeira expressão em (2) impõe que as variáveis sejam homoscedásticos e independentes. A suposição de homoscedasticidade pode ser relaxada modelando-se a dispersão dos dados por meio dos modelos lineares generalizados duplos (Smyth, 1989). Já as equações de estimação generalizadas (Liang e Zeger, 1986) — utilizadas por Schwartz e Dockery (1992) para modelar dados de mortalidade nos Estados Unidos a partir de um modelo de Poisson com efeito aleatório de ano — e os modelos lineares generalizados mistos (Breslow e Clayton, 1993) são uma extensão para dados não-gaussianos correlacionados.

Aqui, a especificação dos termos de tendência e sazonalidade pode ser feita da mesma forma que no modelo linear.

Modelos aditivos generalizados

Modelos aditivos generalizados são uma extensão do modelo linear generalizado que permite associar cada um dos preditores à variável resposta a partir de funções não-lineares, mantendo a suposição de aditividade. Eles são muito utilizados no estudo de poluição do ar, em que a relação entre as variáveis muitas vezes não é linear. Conceição *et al.* (2001b), por exemplo, ajustaram um modelo aditivo Poisson para a série de morbi-mortalidade e compararam os resultados com os do modelo linear Poisson. Os autores mostraram que a primeira classe de modelos apresentou mais poder para detectar os efeitos significantes, embora as conclusões de ambos terem sido coerentes. Carslaw *et al.* (2007) ajustaram modelos aditivos com distribuição Normal para estudar a série de diversos poluentes no centro de Londres, controlando por variáveis meteorológicas. A partir desses modelos, os autores analisaram a tendência das concentrações dos poluentes no período de 1998 a 2005.

Formalmente, o modelo aditivo generalizado pode ser escrito como

$$Y_t|\mathbf{X_t}, \mathbf{Z_t} \overset{ind}{\sim} \mathcal{D}(\mu_t, \phi)$$

$$g(\mu_t) = \beta_0 + f_1(X_{1t}) + ... + f_p(X_{pt}) + f_{p+1}(Z_{1t}) + ... + f_{p+s}(Z_{st}), \tag{3}$$

em que $\mathcal{D}$ é uma distribuição pertencente à família exponencial. No caso mais simples, assim como nos modelos lineares generalizados, supõe-se que as variáveis Y_t são homoscedásticas, independentes e normalmente distribuídas.

Existem diversas propostas a respeito de como f deve ser representada, incluindo o uso de splines naturais, splines penalizados, splines suavizados e loess (Hastie e Tibshirani, 1990). Outra questão importante recai no grau de suavização de f, pois funções muito suaves geram maior viés e funções muito flexíveis geram maior variância. Em geral, utiliza-se validação cruzada para determinar a suavidade ótima de f (Hastie *et al.*, 2008).

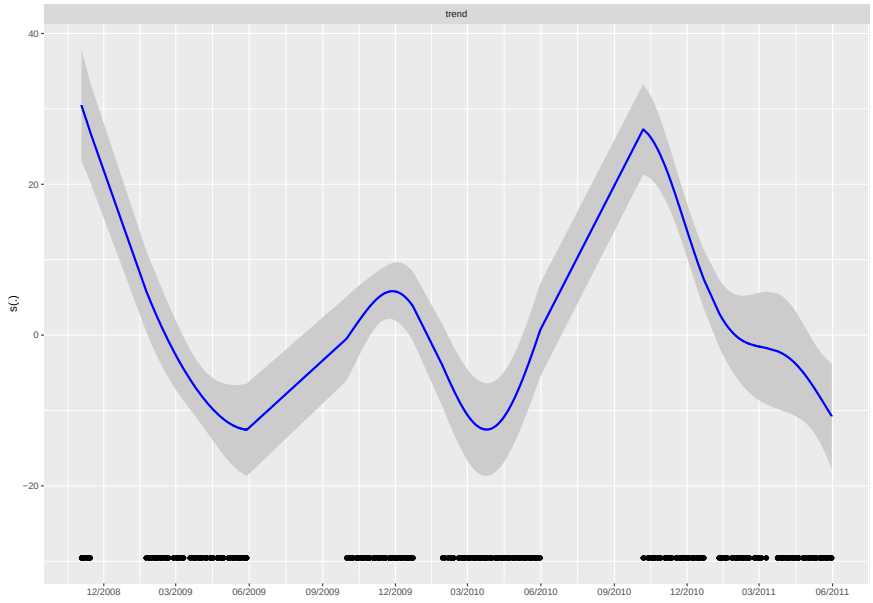

Figura 2: *Função suavizada do tempo para os dados de Salvo e Geiger (2014) (estação Dom Pedro II).*

O ajuste dos modelos aditivos generalizados consiste em incorporar um desses métodos de alisamento ao procedimento de estimação usual de um modelo linear generalizado. Para mais informações, consulte Hastie e Tibshirani (1990) ou Hastie *et al.* (2008).

A expressão (3) considera que todos os preditores serão estimados a partir de funções não-lineares. No entanto, em muitos casos, pode ser interessante considerar o alisamento de apenas algumas variáveis explicativas, enquanto outras são tratadas com estrutura paramétrica. Portanto, um ponto importante no ajuste destes modelos é a escolha de quais variáveis serão alisadas. Em estudos de poluição atmosférica, a utilização de modelos aditivos tem como objetivo o alisamento do tempo, sendo que, em alguns casos, outras variáveis confundidoras cujo efeito é supostamente não-linear, como temperatura, umidade, congestionamento etc, também são alisadas. A escolha dos parâmetros de suavização é outro ponto crucial. Em geral, os algorítimos de estimação permitem que os próprios dados estimem uma suavização ótima, que visa balancear a relação entre viés e variância das estimativas. No entanto, há casos em que a estrutura dos dados (sazonal, por exemplo) demanda que esses valores sejam pré-fixados.

Na Figura 2, pode-se observar a curva suavizada do tempo, representada pela variável *trend*, dada pelo modelo aditivo generalizado ajustado aos dados de Salvo e Geiger (2014). Os pontos paralelos ao eixo x representam a faixa com observações disponíveis. Note que o intervalo de confiança representado pela sombra em volta da curva tende a aumentar na presença de dados faltantes.

Séries temporais

Os modelos utilizados para descrever séries temporais são processos estocásticos, isto é, famílias de variáveis aleatórias definidas em um mesmo espaço de probabilidades e indexadas por um conjunto de índices inteiros ou reais. De uma forma geral, há dois enfoques na análise de séries temporais. O primeiro se encontra no domínio temporal, com base em modelos paramétricos. O segundo, situado no domínio de frequências, se baseia em modelos não-paramétricos. Ambas as abordagens têm como objetivo investigar o mecanismo gerador da série temporal, fazer previsões

de valores futuros e descrever tendências, ciclos, variações sazonais e periodicidades relevantes.

Os modelos paramétricos mais utilizados para a análise de séries temporais são os modelos auto-regressivos integrados e de médias móveis (ARIMA) e os modelos sazonais (SARIMA). Dentre os modelos não-paramétricos, destaca-se a análise espectral.

O modelo ARIMA é uma generalização do modelo auto-regressivo e de média móvel (ARMA). Mais especificamente, é um caso particular de um modelo de filtro linear, que supõe que as séries de tempo são geradas por um sistema linear cuja entrada é um ruído branco. Formalmente,

$$Y_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots, \tag{4}$$

em que $\{\psi_j, j \geq 1\}$ é uma sequência de pesos correspondente a uma função de transferência e μ é um parâmetro que determina o nível da série. Dessa forma, o modelo ARIMA pode ser escrito como

$$Y_t - \alpha_1 X_{t-1} - \ldots - \alpha_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}, \tag{5}$$

sendo $(\alpha_1, \ldots, \alpha_p)$ os parâmetros do modelo auto-regressivo e $(\theta_1, \ldots, \theta_q)$ os parâmetros do modelo de médias móveis.

Essa classe de modelos é frequentemente empregada na análise de séries não-estacionárias, para as quais uma ou mais diferenciações precisam ser aplicadas inicialmente (correspondente ao "integrado" no nome do modelo). A parte auto-regressiva implica que a variável será regredida ou explicada pelos seus próprios valores defasados, como indicado por (4). A parte de médias móveis resulta que os erros aleatórios são uma combinação de termos de erro que ocorreram em vários tempos no passado.

Os modelos ARIMA são geralmente denotados por ARIMA(p, d, q), em que os parâmetros p, d e q são inteiros não-negativos e correspondem, respectivamente, à ordem da defasagem do modelo auto-regressivo, do número de graus de diferenciação da série e da ordem do modelo de médias móveis. A estimação pode ser feita segundo a abordagem de Box e Jenkins (1970).

Quando há evidências de um efeito sazonal na série, o modelo ARIMA pode ser generalizado para o modelo auto-regressivo integrado e de médias móveis sazonal (SARIMA). Esses modelos são usualmente denotados por ARIMA$(p, d, q)(P, D, Q)_m$, em que m se refere ao número de períodos em cada ciclo e os parâmetros P, D e Q referem-se, respectivamente, aos termos auto-regressivo, de diferenciação e de médias móveis da parte sazonal do modelo.

Como os modelos ARIMA e SARIMA não envolvem variáveis explicativas, os seus resíduos podem ser utilizados como variável resposta em um modelo linear para avaliar a relação com variáveis de interesse e controlar os confundidores. Se os modelos de séries temporais estiverem bem ajustados, espera-se que os resíduos resultantes não apresentem mais tendência, volatilidade ou componentes sazonais.

A análise espectral caracteriza a frequência de séries temporais estacionárias, sendo essencial em estudos cujo objetivo é encontrar periodicidade nos dados. A técnica consiste em utilizar funções periódicas, como o seno, para decompor a série em partes mais simples, com diferentes frequências. O periodograma e a transformada de Fourier discreta podem ser utilizados para a estimação espectral.

O periodograma e a função de autocorrelação parcial (Morettin e Toloi, 2004) também são técnicas muito utilizadas na análise de séries temporais. A primeira visa investigar a existência

de padrões cíclicos em diferentes tamanhos de períodos da série. A segunda descreve a correlação serial das observações em defasagens de ordem $1, 2, 3, \ldots$, com os valores corrigidos pelas defasagens anteriores. Katsouyanni *et al.* (1996) utilizaram essas técnicas para analisar séries de poluentes, mortalidade e admissões hospitalares emergenciais em quinze cidades europeias de dez diferentes países. Os autores também utilizaram gráficos das observações contra o tempo, da série temporal predita e da série residual para encontrar estruturas temporais, como tendência e sazonalidade, o efeito de variáveis climáticas, como temperaturas altas e baixas, e a ocorrência de epidemias.

A metodologia usual para análise de séries temporais supõe que as observações são medidas em intervalos igualmente espaçados de tempo, sendo que dados omissos geralmente se tornam um grande problema. Para contornar esse problema, Shumway e Stoffer (1982) propuseram um procedimento recursivo para estimação dos parâmetros via máxima verossimilhança que combina os estimadores suavizados de Kalman (Kalman, 1960) com o algoritmo EM descrito em Dempster *et al.* (1977).

Proposta da tese

De uma maneira geral, este trabalho pretende abordar os seguintes tópicos:

1. revisão bibliográfica para a identificação das diferentes técnicas estatísticas utilizadas na análise de dados de poluição do ar;
2. avaliação das suposições inerentes às técnicas estatísticas utilizadas e quais as consequências de sua violação nos trabalhos estudados;
3. proposta de soluções para contornar essas violações e outras limitações; e
4. aplicação dos resultados ao problema de avaliação do efeito de biocombustíveis na concentração atmosférica de diferentes poluentes.

Para os modelos lineares e os modelos lineares generalizados, avaliar-se-á como contemplar os efeitos de sazonalidade e tendência e em quais situações as suposições desses modelos são razoáveis.

Para os modelos aditivos generalizados, propõe-se discutir a escolha dos parâmetros de suavização, quais variáveis devem ser suavizadas, alternativas para lidar com grandes períodos sem observação e técnicas para avaliar a qualidade do ajuste.

A abordagem baseada em séries temporais será estudada como uma maneira de decompor as séries em seus componentes, como autocorrelação, periodicidade, sazonalidade e tendência. Serão considerados modelos para retirar esses efeitos da série, de tal forma que seus resíduos possam ser ajustados por modelos de regressão sem a necessidade de controlar as variáveis temporais. Também pretende-se estudar formas de contemplar conjuntos de dados com grandes intervalos de observações omissas.

No contexto geral de regressão, serão introduzidos termos que modelem a correlação entre as observações, comparando os resultados desse expediente com aqueles gerados sob a suposição de independência. Também serão discutidas técnicas de regularização, como regressão *ridge* e LASSO (James *et al.*, 2013), que reduzem a variância das estimativas, gerando modelos com menor erro preditivo. Esses procedimentos se baseiam em penalizar a soma de erros quadráticos ou a verossimilhança de tal forma que os coeficientes tendem a ser encolhidos na direção do zero. No caso

do LASSO, coeficientes associados a variáveis pouco significativas podem ser estimados exatamente como zero, permitindo a utilização desse método para seleção de variáveis.

Por fim, esta tese também é motivada pela potencial aplicação de seus resultados na reanálise dos dados de Salvo e Geiger (2014) e de um projeto de pesquisa financiado pela FAPESP, sob responsabilidade do professor José Goldemberg, que visa investigar a real associação entre a concentração de poluentes atmosféricos e a proporção de carros a gasolina na região metropolitana de São Paulo. Além disso, esses resultados também poderão ser utilizados no desenvolvimento de um projeto em colaboração com a CETESB cujo objetivo é estudar a série histórica da concentração de diversos poluentes em São Paulo e relacioná-la ao uso de biocombustíveis.

Referências

Beer *et al.***(2011)** T. Beer, J. Carras, D. Worth, N. Coplin, P. K. Campbell, B. Jalaludin, D. Angove, M. Azzi, S. Brown, I. Campbell, M. Cope, O. Farrell, I. Galbally, S. Haiser, B. Halliburton, R. Hynes, D. Jacyna, M. Keywood, S. Lavrencic, S. Lawson, S. Lee, I. Liepa, J. McGregor, P. Nancarrow, M. Patterson, J. Powell, A. Tibbett, J. Ward, S. White, D. Williams e R. Wood. The health impacts of ethanol blend petrol. *Energies*, (4): 352–367. Citado na pág. 2

Belusic *et al.***(2015)** A. Belusic, I. Herceg-Bulic e Z. B. Klaic. Using a generalized additive model to quantify the influence of local meteorology on air quality in zagreb. *Geofizika*, 32: 48–78. Citado na pág. 2

Box e Jenkins(1970) G. E. P. Box e G. M. Jenkins. *TIme series analysis: forecasting and control*. Holden-Day, San Francisco. Citado na pág. 10

Breslow e Clayton(1993) N. E. Breslow e D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88: 9–25. Citado na pág. 8

Carslaw *et al.***(2007)** D. C. Carslaw, S. D. Beevers e J. E. Tate. Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach. *Atmospheric Environment*, 41: 5289–5299. Citado na pág. 2, 3, 8

Chang *et al.***(2017)** S. Y. Chang, W. Vizuete, M. Serre, L. P. Vennam, M. Omary, V. Isakov, M. Breen e S. Arunachalam. Finely resolved on-road PM2.5 and estimated premature mortality in central north carolina. *Risk Analysis*. doi: 10.1111/risa.12775. URL http://dx.doi.org/10.1111/risa.12775. Citado na pág. 2

Conceição *et al.***(2001a)** G. M. Conceição, S. G. Miraglia, H. S. Kishi, P. H. N. Saldiva e J. da Motta Singer. Air pollution and child mortality: a time-series study in São Paulo, Brazil. *Environmental Health Perspectives*, 109(3): 347–350. Citado na pág. 2, 3

Conceição *et al.***(2001b)** G. M. Conceição, P. H. N. Saldiva e J. da Motta Singer. Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo. *Revista Brasileira de Epidemiologia*, 4(3): 206–219. Citado na pág. 2, 3, 8

Dempster *et al.***(1977)** A. P. Dempster, N. M. Laird e D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.*, B 39: 1–38. Citado na pág. 11

Dobson(1990) A. J. Dobson. *An introduction to generalized linear models*. Chapman and Hall, New York. Citado na pág. 7

European Commission(1999) European Commission. EU focus on clean air. *Office for Official Publications of the European Communities*. Citado na pág. 2

European Commission(2011) European Commission. Climate action. https://ec.europa.eu/clima/policies/strategies/2050_en, 2011. [Online; acessado 15-03-2017]. Citado na pág. 2

Hastie e Tibshirani(1990) T. Hastie e R. Tibshirani. *Generalized additive models*. London:Chapman & Hall. Citado na pág. 8, 9

Hastie *et al.***(2008)** T. Hastie, R. Tibshirani e J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer. Citado na pág. 5, 8, 9

Jacobson(2007) M. Z. Jacobson. Effects of ethanol (E85) versus gasoline vehicles on cancer and mortality in the united states. *Environmental Science & Technology*, 41(11): 4150–4157. Citado na pág. 3

James *et al.***(2013)** G. James, D. Witten, T. Hastie e R. Tibshirani. *An Introduction to Statistical Learning*. Springer Series in Statistics. Springer, New York. Citado na pág. 11

Jasarevic *et al.***(2014)** T. Jasarevic, G. Thomas e N. Osseiran. 7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/, 2014. [Online; acessado 13-03-2017]. Citado na pág. 2

Kalman(1960) R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME J. of Basic Eng.*, 8: 35–45. Citado na pág. 11

Katsouyanni *et al.***(1996)** K. Katsouyanni, J. Schwartz, C. Spix, G. Touloumi, D. Zmirou, A. Zanobetti, B. Wojtyniak, J. M. Vonk, A. Tobias, A. Pönkä, S. Medina, L. Bachárová e H. R. Anderson. Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of Epidemiology & Community Health*, 50(Suppl 1): S12–S18. ISSN 0143-005X. doi: 10.1136/jech.50.Suppl_1.S12. URL http://jech.bmj.com/content/50/Suppl_1/S12. Citado na pág. 2, 3, 11

Kloog *et al.***(2012)** I. Kloog, F. Nordio, B. A. Coull e J. Schwartz. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the mid-atlantic states. *American Chemical Society*, 46: 11913–11921. Citado na pág. 2

Liang e Zeger(1986) K. Y. Liang e S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73: 13–22. Citado na pág. 8

Lin *et al.***(1999)** C. A. Lin, M. A. Martins, S. C. Farhat, C. A. Pope, G. M. Conceição, V. M. Anastácio, M. Hatanaka, W. C. Andrade, W. R. Hamaue, G. M. Bohm e P. H. Saldiva. Air pollution and respiratory illness of children in São Paulo, Brazil. *Paediatric and Perinatal Epidemiology*, 13(4): 475–488. ISSN 1365-3016. doi: 10.1046/j.1365-3016.1999.00210.x. URL http://dx.doi.org/10.1046/j.1365-3016.1999.00210.x. Citado na pág. 2, 3

Morettin e Toloi(2004) P. A. Morettin e C. M. Toloi. *Análise de Series Temporais*. ABE - Projeto Fisher e Editora Edgard Blucher, São Paulo. Citado na pág. 6, 10

Mulawa *et al.***(1997)** P. A. Mulawa, S. H. Cadle, K. Knapp, R. Zweidinger, R. Snow, R. Lucas e J. Goldbach. Effect of ambient temperature and E10 fuel on primary exhaust particulate matter emissions from light-duty vehicles. *American Chemical Society: Environ. Sci. Technol.*, 31 (5): 1302–1307. Citado na pág. 2

Nelder e Wedderburn(1972) J. A. Nelder e R. W. M. Wedderburn. Generalized linear models. *Stat Soc A*, 135: 370–384. Citado na pág. 7

Pereira *et al.***(2004)** P. A. Pereira, L. M. B. Santos, E. T. Sousa e J. B. de Andrade. Alcohol- and gasohol-fuels: a comparative chamber study of photochemical ozone formation. *Journal of the Brazilian Chemical Society*, 15(5): 646–651. Citado na pág. 2

Saldiva *et al.***(1994)** P. H. N. Saldiva, A. J. F. C. Lichtenfels, P. S. O. Paiva, I. A. Barone, M. A. Martins, E. Massad, J. C. R. Pereira, V. P. Xavier, J. M. Singer e G. M. Bohm. Association between air pollution and mortality due to respiratory diseases in children in São Paulo, Brazil: a preliminary report. *Environmental Research*, 65 (2): 218 – 225. ISSN 0013-9351. doi: http://dx.doi.org/10.1006/enrs.1994.1033. URL http://www.sciencedirect.com/science/article/pii/S0013935184710334. Citado na pág. 2, 3

Saldiva *et al.***(1995)** P. H. N. Saldiva, C. A. Pope, J. Schwartz, D. W. Dockery, A. J. Lichtenfels, J. M. Salge, I. Barone e G. M. Bohm. Air pollution and mortality in elderly people: a time-series study in São Paulo, Brazil. *Archives of Environmental Health: An International Journal*, 50: 159–163. Citado na pág. 2, 3, 5

Salvo e Geiger(2014) A. Salvo e F. M. Geiger. Reduction in local ozone levels in urban São Paulo due to a shift from ethanol to gasoline use. *Nature Geoscience*, 7: 450–458. Citado na pág. 3, 4, 5, 6, 7, 9, 12

Schwartz e Dockery(1992) J. Schwartz e D. W. Dockery. Particulate air pollution and daily mortality in Steubenville, Ohio. *Am J Epidemiol.*, 1(135): 12–19. Citado na pág. 2, 3, 8

Schwartz *et al.*(1996) J. Schwartz, D. W. Dockery e L. M. Neas. Is daily mortality associated specifically with fine particles? *J Air Waste Manag Assoc*, 10(46): 927–939. Citado na pág. 2, 3

Schwartz(1994) J. Schwartz. Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Statistical Society of Canada*, 22(4): 471–487. Citado na pág. 2, 3

Schwartz(1996) J. Schwartz. Air pollution and hospital admissions for respiratory disease. *Epidemiology*, 1(7): 20–28. Citado na pág. 2, 3

Schwartz e Marcus(1990) J. Schwartz e A. Marcus. Mortality and air pollution in London: a time series analysis. *American Journal of Epidemiology*, 131(1): 185. doi: 10.1093/oxfordjournals.aje.a115473. URL +http://dx.doi.org/10.1093/oxfordjournals.aje.a115473. Citado na pág. 2, 3

Shumway e Stoffer(1982) R. H. Shumway e D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4): 253–264. ISSN 1467-9892. doi: 10.1111/j.1467-9892.1982.tb00349.x. URL http://dx.doi.org/10.1111/j.1467-9892.1982.tb00349.x. Citado na pág. 2, 3, 11

Smyth(1989) G. K. Smyth. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society*, (51): 47–60. Citado na pág. 8

Sprott(2000) D. A. Sprott. *Statistical Inference in Science*. Springer Series in Statistics. Springer. Citado na pág. 5

Yoon *et al.*(2009) S. H. Yoon, S. Y. Ha, H. G. Roh e C. S. Lee. Effect of bioethanol as an alternative fuel on the emissions reduction characteristics and combustion stability in a spark ignition engine. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 223: 941–951. Citado na pág. 2