Estratégias para análise de dados de poluição do ar

William Nilson de Amorim

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

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, fevereiro de 2019

Estratégias para análise de dados de poluição do ar

Esta é a versão original da tese elaborada pelo candidato William Nilson de Amorim, tal como submetida à Comissão Julgadora.

Agradecimentos

Texto texto. Texto opcional.

Ao Starbucks por todas as vezes que eu trabalhei lá sem consumir nada.

Resumo

AMORIM, W. N. **Estratégias para análise de dados de poluição do ar**. 2019. ?? f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 201?.

Elemento obrigatório, constituído de uma sequência de frases concisas e objetivas, em forma de texto. Deve apresentar os objetivos, métodos empregados, resultados e conclusões. O resumo deve ser redigido em parágrafo único, conter no máximo 500 palavras e ser seguido dos termos representativos do conteúdo do trabalho (palavras-chave). Texto texto. Texto texto.

Palavras-chave:

Abstract

Amorim, W. N. **Strategies for air pollution data modelling**. 201?. ?? f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

Elemento obrigatório, elaborado com as mesmas características do resumo em língua portuguesa. De acordo com o Regimento da Pós- Graduação da USP (Artigo 99), deve ser redigido em inglês para fins de divulgação. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

Keywords:

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A poluição do ar, considerada pela Organização Mundial da Saúde (OMS) como 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 em afecções como 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 no sistema nervoso, câncer e aumento da mortalidade infantil (European Commission, 1999). Além disso, vários poluentes estão diretamente ligados ao aquecimento global e ao efeito estufa.

As taxas elevadas de poluição do ar geralmente são produto de políticas não sustentáveis em setores como transporte, energia, saneamento e indústria. A escolha de estratégias favoráveis à saúde pública e ao meio ambiente costuma esbarrar em fatores econômicos, mesmo quando a redução a longo prazo nos gastos com tratamentos de saúde poderia gerar números positivos nesse balanço. Nas últimas décadas, diversos estudos vêm sendo realizados para nos alertar sobre os riscos da poluição atmosférica. Seus principais objetivos compreendem a descrição dos níveis locais de poluição, o acompanhamento das concentrações dos poluentes ao longo do tempo, a busca por associações entre a concentração de poluentes e mortalidade ou morbidade e o desenvolvimento de soluções mais limpas (ou menos poluentes) que ainda sejam economicamente viáveis.

Nesse contexto, como discutido em Zannetti (1990), a análise de dados de poluição do ar é de extrema importância, pois, a partir dela, podemos:

- levantar evidências para a criação de leis de controle de emissão;
- avaliar os impactos de novas legislações;
- determinar e responsabilizar fontes atuais de poluição;
- selecionar regiões para futuras fontes de poluição, minimizando o impacto ambiental;
- controlar episódios severos de poluição a partir de estratégias de intervenção;
- investigar o efeito da concentração de poluentes na saúde pública, principalmente em grupos de risco como crianças, gestantes e idosos. .

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. Beer *et al.* (2011) analisaram dados de morbidade e mortalidade para estudar os impactos à saúde ao se utilizar etanol como aditivo na gasolina em regiões urbanas da Austrália, medindo níveis de ozônio, dióxido de nitrogênio e material particulado em câmaras de poluição. 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), Saldiva *et al.* (1994, 1995), Schwartz (1994, 1996), Schwartz e Dockery (1992) e Schwartz e Marcus (1990) —, embora abordem diferentes temas, concluem sobre a importância da diminuição da emissão de poluentes.

Embora a formação de poluentes envolva reações químicas complicadas, cuja análise demanda ambientes controlados, as grandes cidades podem ser pensadas como laboratórios naturais 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 na formação dos poluentes. Nesse sentido, podemos citar Jacobson (2007), Salvo e Geiger (2014) e Salvo *et al.* (2017).

Uma das maiores dificuldades associadas ao estudo de dados de poluição atmosférica está no grande número de efeitos confundidores. Em áreas urbanas, pode existir uma complexa mistura de fatores que contribuem para a formação e dispersão dos poluentes. Diversas variáveis podem ser consideradas: clima, tráfego, química atmosférica local, mudanças climáticas sazonais, feriados ou eventos esporádicos na cidade (que podem alterar o fluxo do trânsito), tamanho e idade da frota de veículos, emissões evaporativas, entre outras. Além do grande número de variáveis, a relação entre elas pode não ser muito simples, exigindo o uso de modelos mais flexíveis, que geralmente deixam a modelagem mais desafiadora. E por fim, também não é rara a presença de dados omissos ou grandes períodos sem observação, dificultando ainda mais a análise.

Embora diferentes técnicas estatísticas venham sendo empregadas na modelagem de dados de poluição atmosférica, nenhuma delas é robusta o suficiente para atacar sozinha todos esses problemas. Dependendo do problema, precisamos formular estratégias que envolvam a combinação de duas ou mais técnicas, e, às vezes, até metodologias que ainda não são usadas no contexto de poluição do ar.

Na literatura, os principais estudos de poluição do ar envolvem a utilização de modelos lineares, como em Salvo e Geiger (2014) e Salvo *et al.* (2017), modelos generalizados (Conceição *et al.*, 2001b; Lin *et al.*, 1999; Saldiva *et al.*, 1994, 1995; Schwartz e Dockery, 1992), modelos aditivos generalizados (Carslaw *et al.*, 2007; Conceição *et al.*, 2001a,b; Schwartz *et al.*, 1996; Schwartz, 1994, 1996) e modelos para 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 do ar[1] usualmente violam as suposições associadas a esses modelos. Salvo e Geiger (2014), por exemplo, utilizam um modelo de regressão linear, cujo ajuste assume suposições nem sempre razoá-

[1] Consideraremos tanto séries de poluentes quanto dados epidemiológicos (número de mortes ou casos de doenças associadas à poluição do ar).

veis em problemas reais. De uma forma geral, autocorrelação, heteroscedasticidade, superdispersão, tendência, sazonalidade, componentes espaciais, variáveis com erro de medida e grandes períodos sem observação (ou muitas observações omissas) são características comuns em dados de poluição do ar e precisam ser identificadas e contempladas pelo modelo escolhido.

Modelos lineares (Hastie *et al.*, 2008) 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, já que as observações não são interpretadas como uma série. Por outro lado, a concentração de poluentes costuma ser autocorrelacionada, tornando a suposição de observações independentes muito restritiva. Os modelos lineares generalizados (Nelder e Wedderburn, 1972) flexibilizam a suposição 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 (Hastie e Tibshirani, 1990) são uma boa alternativa nesse contexto. Eles permitem incluir termos não-paramétricos 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 das funções de suavização e hiperparâmetros não é trivial. Grandes períodos sem observação também atrapalham o ajuste, pois a curva suavizada ao longo do tempo é suposta contínua durante todo o intervalo.

Modelos para séries temporais acomodam bem os componente temporais, como sazonalidade e autocorrelação, mas diversos deles assumem que a série é estacionária, sendo preciso realizar alguma transformação nos dados antes do ajuste.

O objetivo desta tese é criar e discutir estratégias robustas para a análise de dados de poluição do ar que considerem as seguintes situações:

1. grande número de covariáveis e relações não-lineares entre as covariáveis e a resposta;
2. presença de dados omissos;
3. e presença de variáveis não-observáveis ou com erro de medida.

Os resultados deste trabalho visam auxiliar pesquisadores de qualquer área, mesmo aqueles com pouca experiência em modelagem estatística, na análise de dados oriundos de estudos de poluição do ar. O desenvolvimento adequado desses trabalhos é essencial para a criação de políticas públicas que promovam melhorias na saúde pública das grandes cidades e na nossa relação com o meio ambiente.

Com essa finalidade, utilizaremos conjuntos de dados reais para apontar as vantagens e desvantagens de cada metodologia, gerando estratégias de análise que contemplem as principais dificuldades encontradas na prática. Também aplicaremos técnicas de aprendizado estatístico pouco ou ainda não utilizadas no estudo de séries de poluição do ar, como validação cruzada e regularização.

No Capítulo 2, discutiremos a análise exploratória de dados de poluição, com o objetivo principal de diminuir a complexidade do problema para facilitar a busca de informações relevantes sobre o fenômeno estudado (Wickham e Grolemund, 2017). No Capítulo 3, apresentaremos resumidamente

os principais modelos utilizados no ajuste de séries de poluição, assim como outras técnicas interessantes para a análise desses dados. Nos Capítulos 4, 5 e 6 desenvolveremos estratégias para a análise de dados de poluição do ar envolvendo a descrição do poluente ao longo do tempo, o uso de combustíveis e a associação com dados de saúde pública. No Capítulo 7, apresentaremos formas de se extrair da internet dados públicos de poluição do ar. Por fim, no Capítulo 8, concluiremos a tese discutindo os principais resultados.

O texto a seguir busca um equilíbrio entre formalismo matemático, interpretação e aplicabilidade. O propósito dessa tentativa é produzir um trabalho acessível a pesquisadores de todas as áreas, tendo em vista os objetivos propostos. Apesar disso, um certo grau de conhecimento estatístico será exigido em muitos pontos. Hastie e Tibshirani (1990); James *et al.* (2013) são ótimos livros para consulta. A parte computacional deste trabalho foi realizada integralmente no programa estatístico R (R Core Team, 2016), sendo Wickham e Grolemund (2017) uma excelente referência.

Capítulo 2

Análise exploratória

> The greatest value of a picture
> is when it forces us to notice
> what we never expected to see
> — John Tukey

A análise exploratória é a primeira tentativa de se extrair informação dos dados. Seu objetivo é gerar conhecimento acerca do fenômeno sob estudo para guiar as próximas etapas da análise. Existem diversas maneiras de conduzir uma análise exploratória, e a estratégia aplicada a cada problema depende do tipo de variável com que estamos trabalhando.

Como estudos de poluição do ar geralmente envolvem *séries temporais*, apresentaremos diversas técnicas para explorar variáveis dessa natureza. Uma visão mais geral sobre a análise exploratória de dados pode ser encontrada em Wickham e Grolemund (2017).

Séries temporais compreendem variáveis observadas repetidas vezes ao longo de grande períodos de tempo. As metodologias usuais para a análise de séries temporais supõem que as observações são realizadas em intervalos equidistantes, principalmente pela facilidade computacional que essa propriedade proporciona. Por pragmatismo, dado que esse é o cenário mais comum na prática, o enfoque deste trabalho será na análise de séries com essa característica. Para a análise de séries com observações não igualmente espaçadas, recomendamos a leitura de Eckner (2018).

O efeito do tempo nas observações é a grande peculiaridade das séries temporais, gerando características como *tendência*, *sazonalidade* e *autocorrelação*, que influenciam diretamente a escolha do melhor modelo para os dados. A identificação dessas características é fundamental para a análise, o que torna a análise exploratória uma etapa de extrema importância no estudo de séries temporais. Discutiremos esse tópico na Seção 2.2.

Neste texto, representaremos séries temporais pela notação $\{Y_t, t \geq 0\}$ ou, de forma simplificada, Y_t. O índice inteiro não-negativo t representa a ordem em que as observações foram realizadas. As variáveis serão denotadas por letras maiúsculas do nosso alfabeto e os índices por letras minúsculas: Y_t, X_s, Z_r etc.

Sob o contexto de estudos de poluição do ar, apresentaremos a seguir as principais técnicas para análise exploratória de séries temporais. Utilizaremos como exemplo as séries horárias de concentração de ozônio (O_3), óxido de nitrogênio (NO), dióxido de nitrogênio (NO_2) e temperatura, todas medidas na cidade de São Paulo de 2008 a 2013, disponibilizadas por Salvo e Geiger (2014)

e Salvo *et al.* (2017) nos respectivos endereços: http://bit.do/salvo_geiger_data e https://goo.gl/9tNzvj.

2.1 Gráficos

> The simple graph has brought more information to the data analyst's mind than any other device. – John Tukey

Nós construímos gráficos para elucidar informações sobre as variáveis que estão "escondidas"na base de dados. Para cumprir esse objetivo, um gráfico precisa ser facilmente compreendido, dado que gráficos muito verbosos podem ser mal interpretados e gerar mais confusão do que esclarecimento.

Embora o conceito de gráfico estatístico seja amplamente conhecido, não há um consenso sobre o que realmente é um gráfico e, por consequência, quais as melhores práticas para construí-lo. Wilkinson (2005) atacou esse problema definindo um gráfico estatístico como o mapeamento de variáveis em atributos estéticos de formas geométricas. Essa definição, conhecida como "a gramática dos gráficos", é interessante pois contempla os principais modelos gráficos já conhecidos e abre caminho para a criação de *frameworks* bem estruturados de construção de gráficos.

Wickham (2010), por exemplo, utilizou a as ideias propostas por Wilkinson (2005) e definiu uma "gramática dos gráficos por camadas"[1], acrescentando que cada elemento de um gráfico representa uma camada e que o gráfico em si é a sobreposição de todas as suas camadas. O resultado dessa definição foi a origem ao pacote de R `ggplot2`, sendo uma das melhores ferramentas atuais para criação de gráficos estáticos.

A visualização mais comum para séries temporais é o *gráfico da série*. Com base na definição criada por Leland, as variáveis mapeadas serão o par (t, Y_t), as formas geométricas são retas e o atributo estético é a posição dessas retas em um eixo coordenado (com t, o tempo, no eixo x e Y_t no eixo y). A seguir, apresentamos alguns exemplos de como construir e interpretar esses gráficos.

2.1.1 O gráfico da série

O gráfico da série é uma visualização da variável Y_t contra o tempo. A partir dele, podemos observar a existência de diversos comportamentos, como tendência, sazonalidade e heteroscedasticidade[2], sendo a principal técnica de visualização de séries temporais.

Apesar de ser uma ferramenta de fácil construção e interpretação, quando o volume de dados é muito grande, a simples construção do gráfico da série pode não trazer toda a informação disponível nos dados. Uma boa estratégia nesse cenário é tentar diminuir a complexidade do problema, trabalhando inicialmente com casos particulares e, em seguida, buscar os padrões encontrados nos casos mais gerais.

Como exemplo de como explorar os dados utilizando o gráfico da série, vamos analisar a concentração horária de ozônio medida na Grande São Paulo, no período de 2008 a 2013, disponibilizada por Salvo *et al.* (2017).

A base de dados contém medições de ozônio de 12 estações de monitoramento espalhadas pela cidade. A princípio, vamos analisar o gráfico de apenas uma delas, por exemplo, a estação Dom Pedro II (Figura 2.1). Podemos observar alguns períodos sem observação e, com a ajuda da série suavizada (por *splines* cúbicos, ver Seção 3.3.2), uma sazonalidade anual, com picos no início de cada ano.

[1] *A layered grammar of graphics.*
[2] Variância não-constante ao longo do tempo.

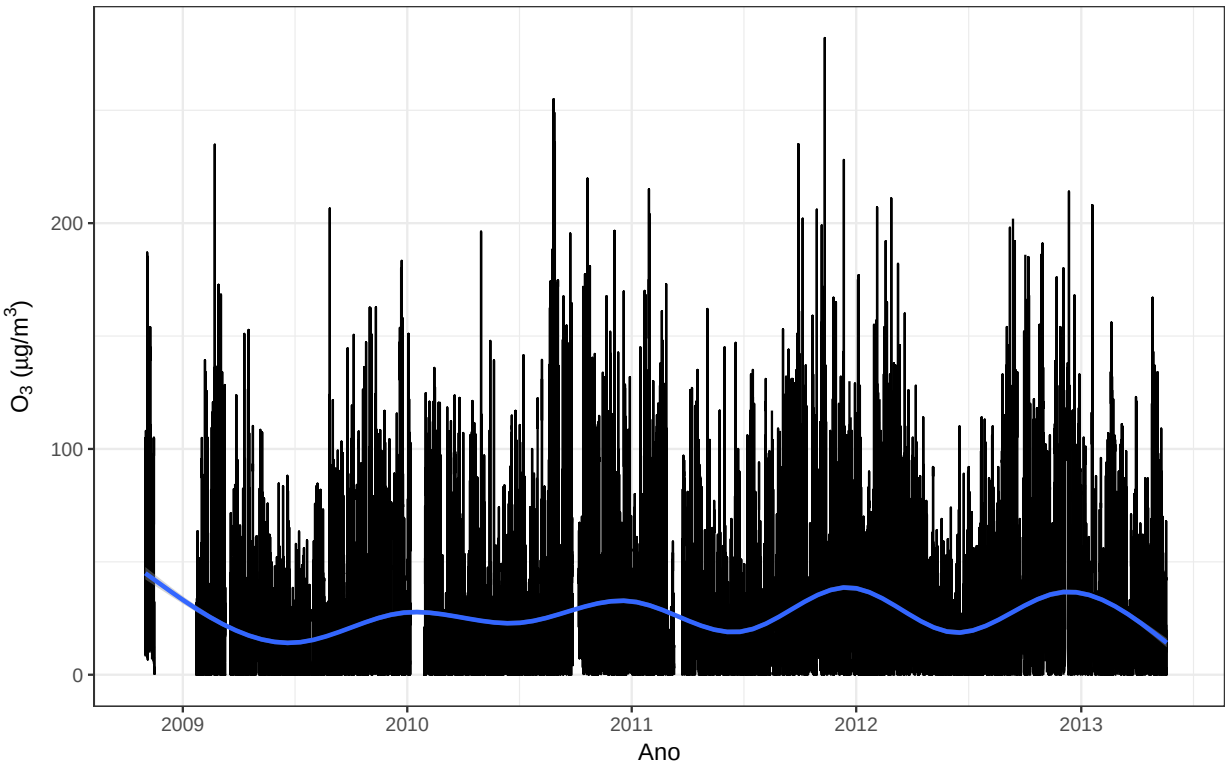

Figura 2.1: *Série da concentração de ozônio para a estação Dom Pedro II, na cidade de São Paulo, no período de 2008 a 2013. Em azul, a série suavizada usando splines cúbicos.*

Como a série é horária, o grande volume de observações pode ocultar alguns padrões. Para avaliar o comportamento da concentração de ozônio ao longo do dia, vamos analisar a concentração média horária dentro do período analisado (Figura 2.2). Observamos que o pico de ozônio, em geral, acontece no começo da tarde, entre o meio-dia e 16 horas.

Podemos então considerar a média diária dentro desse período para avaliar apenas o horário em que a concentração de ozônio geralmente está alta. Observe pela Figura 2.3 que fica mais fácil observar o padrão sazonal. O padrão parece não ser o mesmo em 2009, mas essa diferença provavelmente se deve à falta de informação no período. Como indicado na Figura 2.4, esse padrão se repete para todas as 12 estações.

Note que conduzir a análise exploratória na direção de casos particulares facilita a obtenção de informações importantes sobre o fenômeno. No exemplo, essa particularização poderia ainda ser feita em várias direções, como avaliar as diferenças entre os dias da semana ou as estações do ano.

Muitas vezes, também temos interesse em estudar a relação entre duas séries. Os gráficos dessas séries, avaliadas em um mesmo período, podem então ser construídos na mesma figura como uma tentativa de encontrar padrões no comportamento conjunto das duas curvas. Na Figura 2.5, construímos gráficos das séries horárias de ozônio e de óxido de nitrogênio (NO), ambos medidos na estação Dom Pedro II, em São Paulo, no período de 2008 a 2011. Podemos observar que períodos de menor concentração de ozônio parecem estar associados a períodos de maior concentração de NO.

Os gráficos da série geram bastante intuição sobre o comportamento do fenômeno sob estudo, mas seria interessante dispormos de medidas mais objetivas. Nas próximas seções, discutiremos os conceitos de estacionariedade e autocorrelação e como identificar essas características. Além disso,

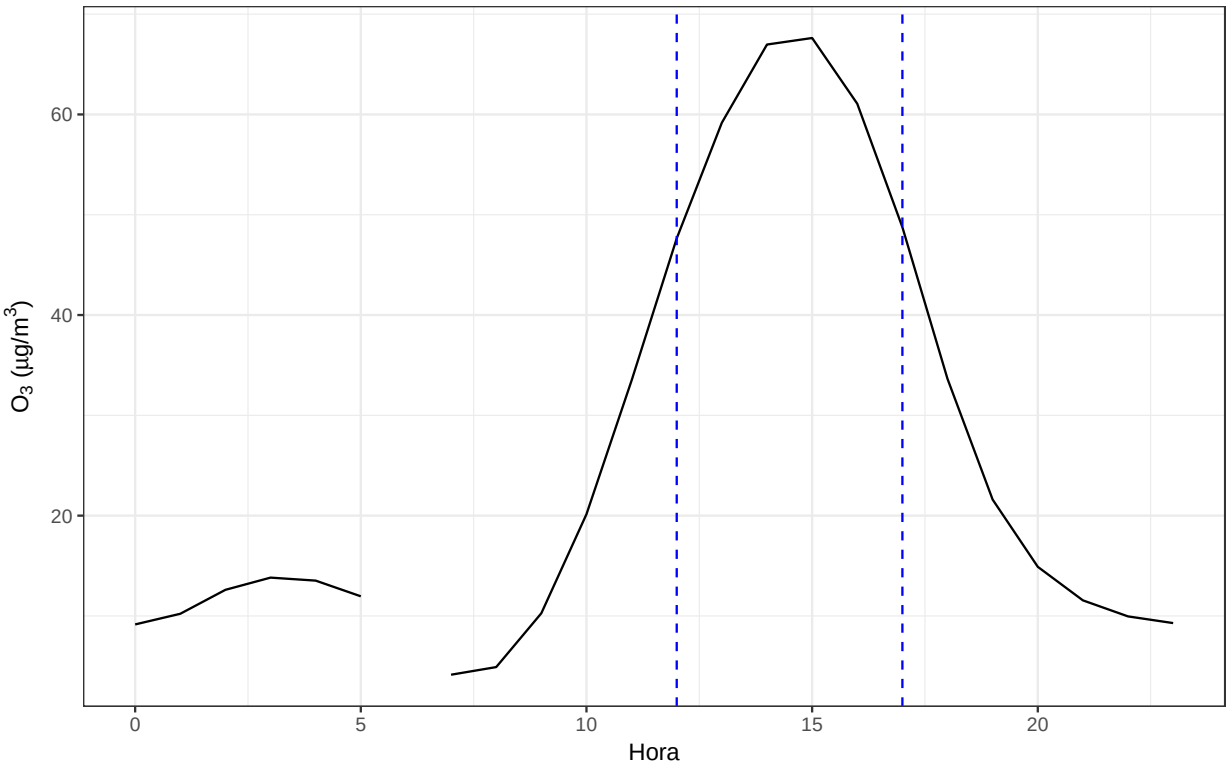

Figura 2.2: *Série da concentração média de ozônio ao longo do dia para a estação Dom Pedro II, na cidade de São Paulo, no período de 2008 a 2013. Não existe informação para as 6 da manhã pois é o horário em que o equipamento sofre manutenção.*

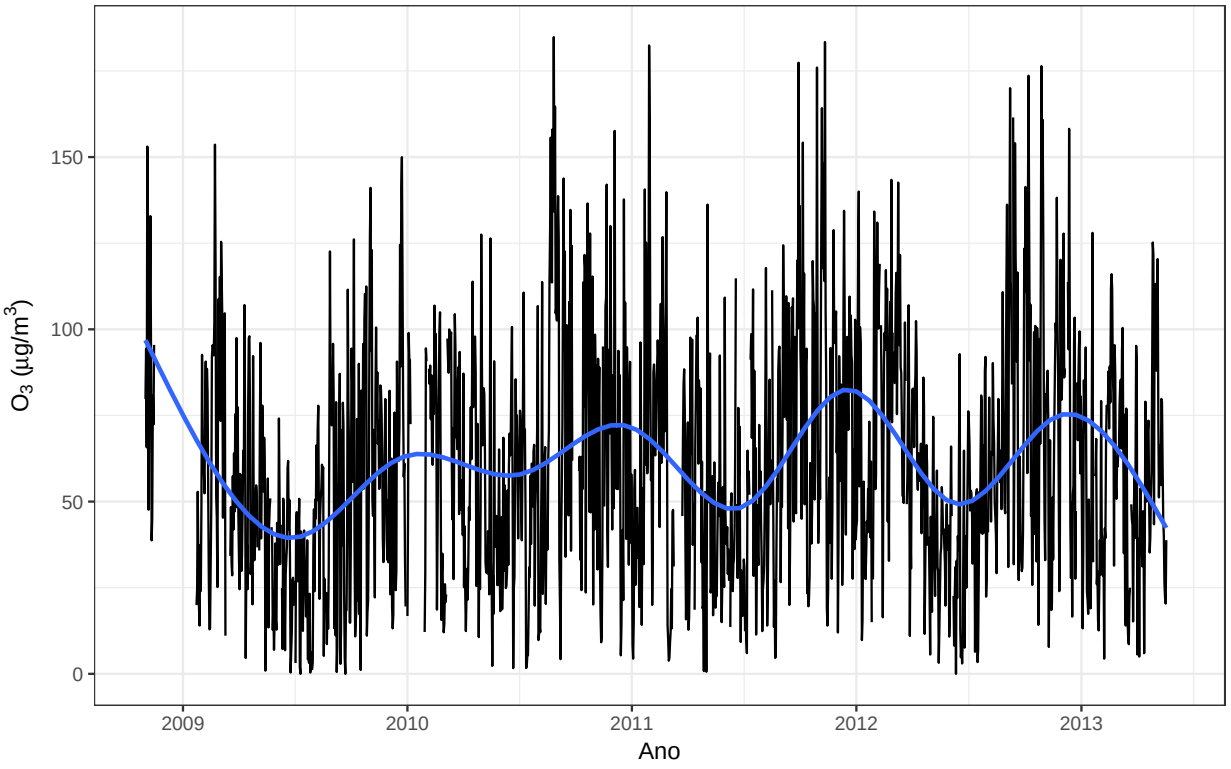

Figura 2.3: *Série diária da concentração média de ozônio medido no começo da tarde para a estação Dom Pedro II, na cidade de São Paulo, no período de 2008 a 2013. Em azul, a série suavizada usando splines cúbicos.*

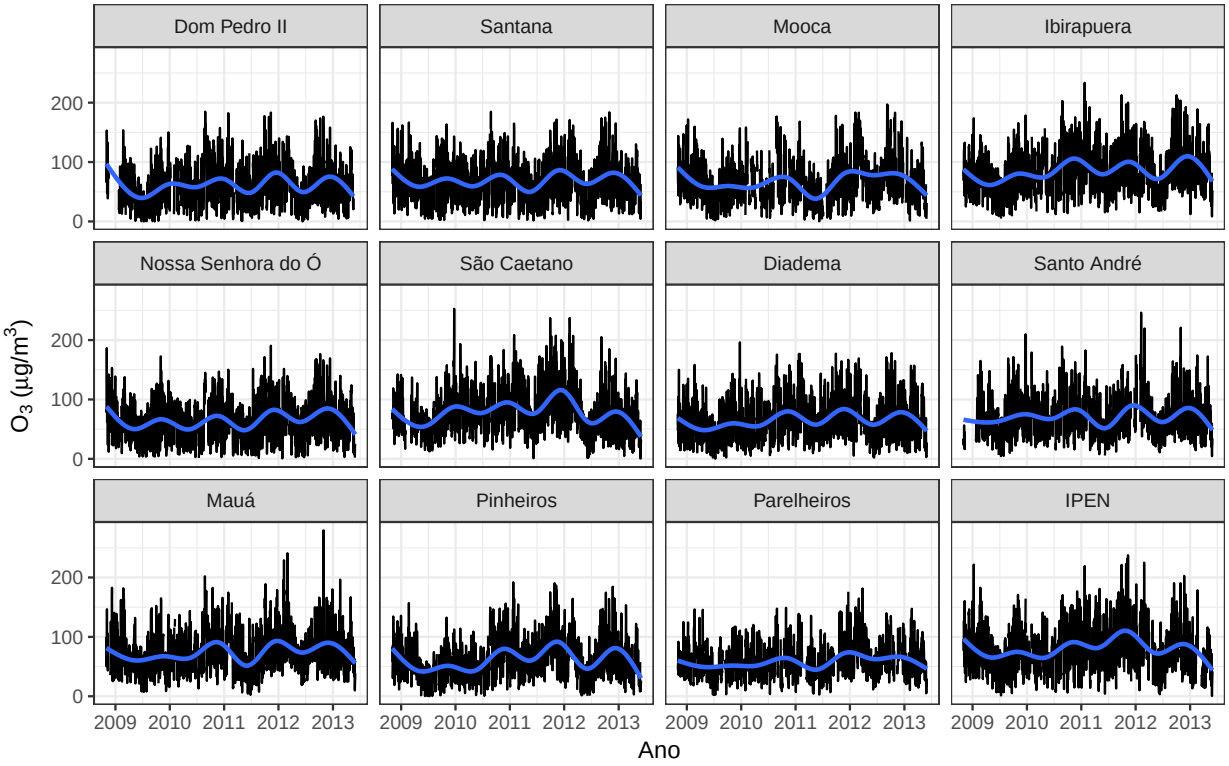

Figura 2.4: *Série diária da concentração média de ozônio medido no começo da tarde para todas as estações, na cidade de São Paulo, no período de 2008 a 2013. Em azul, as séries suavizadas usando splines cúbicos.*

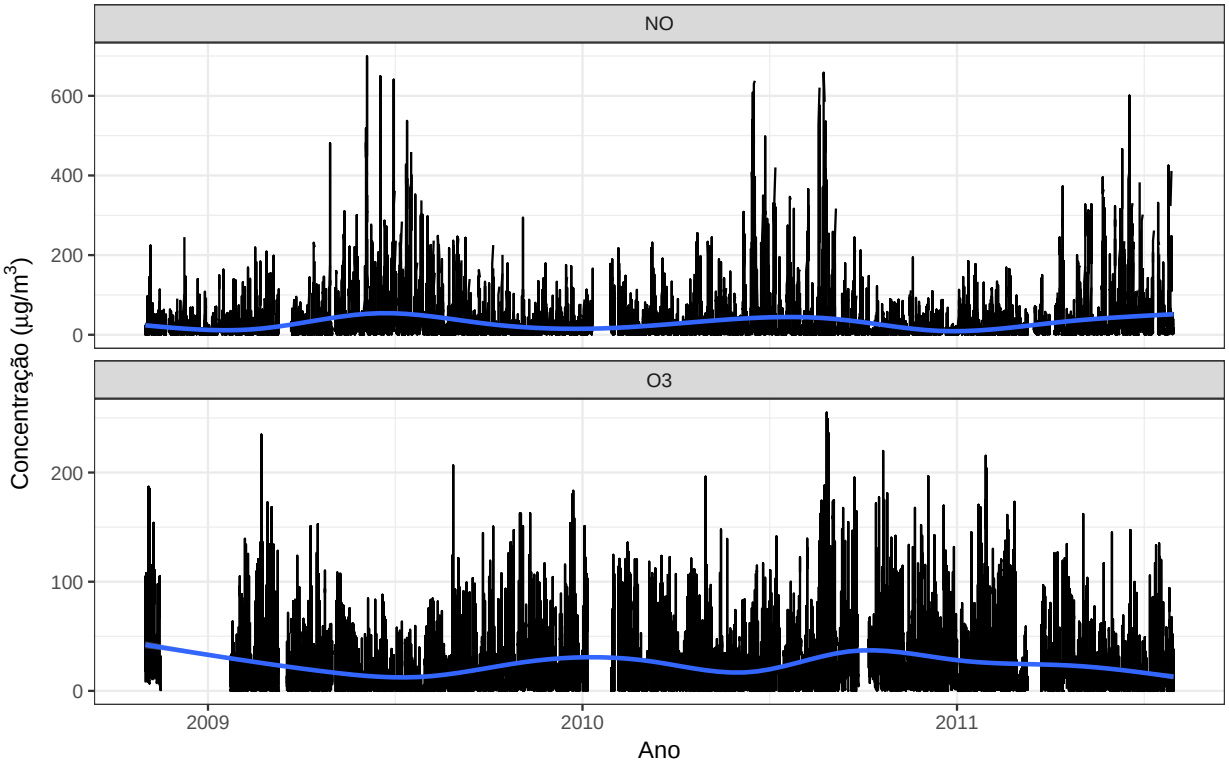

Figura 2.5: *Séries horárias de ozônio e de óxido de nitrogênio (NO), ambos medidos na estação Dom Pedro II, em São Paulo, no período de 2008 a 2011. Em azul, as séries suavizadas usando splines cúbicos.*

apresentaremos estratégias para conduzir a análise na presença de tendência e sazonalidade.

2.1.2 Gráficos de dispersão

Gráficos de dispersão são amplamente utilizados na Estatística. Sua principal função é estudar a associação entre duas variáveis, sendo possível levantar indícios sobre a forma, intensidade e direção dessa associação caso ela exista. Construímos esses gráficos posicionando pontos em um eixo cartesiano, sendo a variável resposta mapeada no eixo y e a variável explicativa no eixo x. Podemos também adicionar curvas suavizadas para facilitar a identificação da associação.

Na Figura 2.6, apresentamos o gráfico de dispersão da concentração de ozônio contra a concentração de óxido de nitrogênio, ambas medidas das 12 às 16 horas, de 2008 a 2011. Observamos que a concentração de ozônio decresce exponencialmente conforme a concentração de NO aumenta. É conhecido que o ozônio ao longo da tarde reage com o NO, portanto espera-se que dias de alta concentração de ozônio tenham baixa concentração de NO e vice-versa.

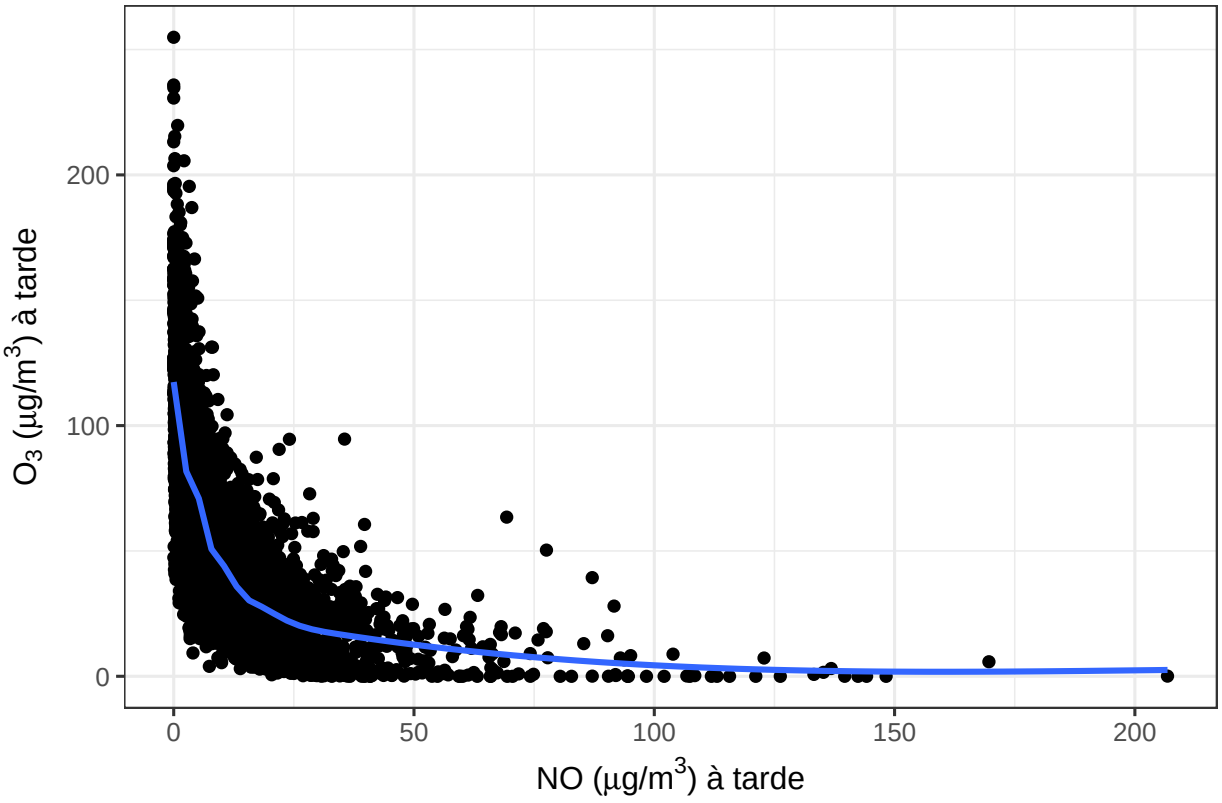

Figura 2.6: *Gráfico de dispersão da concentração de ozônio contra a concentração de óxido de nitrogênio medidas das 12 às 16 horas na estação de monitoramento Dom Pedro II, em São Paulo, de 2008 a 2011.*

Apresentamos agora, na Figura 2.7, o gráfico de dispersão da concentração de ozônio contra a concentração de dióxido de nitrogênio, ambas também medidas das 12 às 16 horas, de 2008 a 2011. Observe que não há indícios de associação entre as duas variáveis. No entanto, sabe-se que a fotólise do NO_2 pela manhã faz parte do processo gerador do ozônio ao longo da tarde. Na Figura 2.8, apresentamos o gráfico de dispersão da concentração de ozônio, medida à tarde, contra a concentração e dióxido de nitrogênio, agora medida pela manhã, das 7 às 11 horas. Observe que, neste caso, encontramos indícios de uma relação positiva entre as duas variáveis.

Uma limitação dos gráficos de dispersão é não levar em conta o efeito de outras variáveis. Muitas vezes a associação dentre duas variáveis pode ser induzida ou mascarada pela ação de uma terceira. Portanto, é importante termos em mente que a interpretação desses gráficos levanta apenas indícios

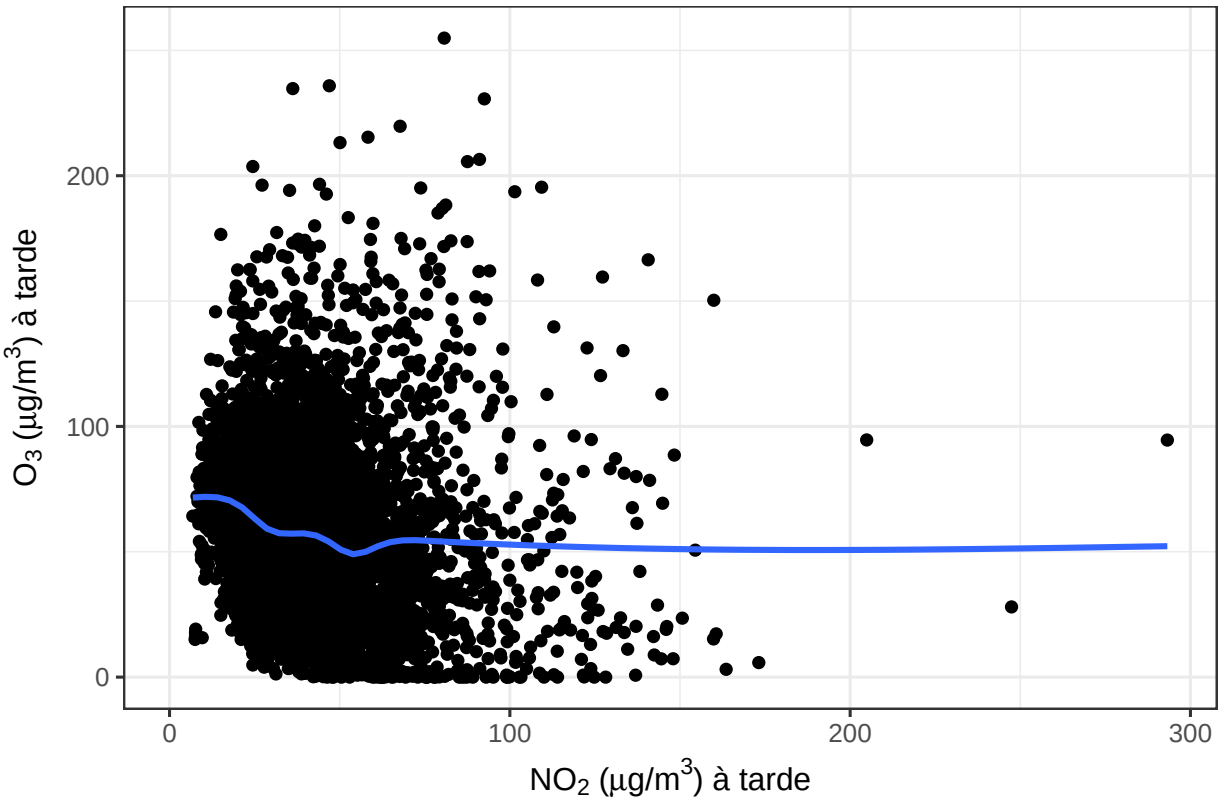

Figura 2.7: *Gráfico de dispersão da concentração de ozônio contra a concentração de dióxido de nitrogênio medidas das 12 às 16 horas na estação de monitoramento Dom Pedro II, em São Paulo, de 2008 a 2011.*

sobre a associação, que devem ser estudados com mais atenção, eliminando primeiro o possível efeito de outras variáveis.

2.1.3 Gráficos de distribuição

Muitas vezes queremos observar a distribuição amostral de uma variável. Um gráfico muito comum nesses casos é o histograma. Na Figura 2.9, apresentamos o histograma da concentração diária média medida de ozônio das 12 às 16 horas, em São Paulo, de 2008 a 2013. Podemos observar que a distribuição amostral é levemente assimétrica à direita, sendo que a maioria dos dias apresenta concentração de ozônio entre 25 e 75 μg/m^3.

Quando estamos interessados em, além de observar a distribuição amostral de uma variável, compará-la entre os níveis de uma segunda variável, os chamados *ridges graphs* são uma boa alternativa. Podemos observar na Figura 2.10 que as máximas de ozônio ocorrem nos meses mais quentes, sendo esses os períodos também de maior variação, provavelmente devido ao efeito conjunto da temperatura e da chuva.

2.2 Componentes temporais

Algumas classes de modelos utilizados para o ajuste de séries temporais assumem que a série estudada é *estacionária*, isto é, mantem um mesmo comportamento ao longo do tempo[3]. Além de simplificar o processo de estimação, a estacionariedade garante que algumas quantidades sejam

[3]Em termos probabilísticos, isso significa que a sua distribuição de probabilidades não muda no tempo.

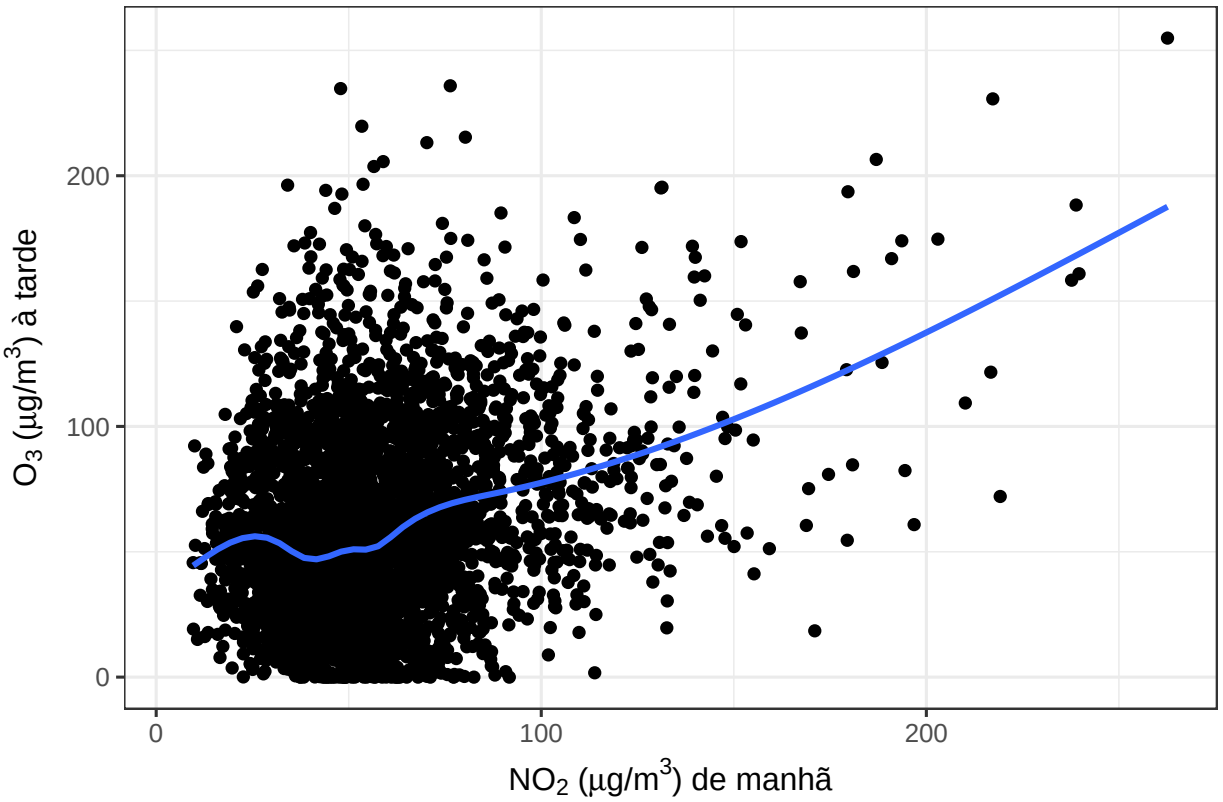

Figura 2.8: *Gráfico de dispersão da concentração de ozônio, medida das 12 às 16 horas, contra a concentração de óxido de nitrogênio, medida das 7 às 11 horas, na estação de monitoramento Dom Pedro II, em São Paulo, de 2008 a 2011.*

interpretáveis. No entanto, a não-estacionariedade é muito comum no mundo real, o que gera a necessidade de identificá-la e, se necessário, removê-la antes do ajuste.

Dada uma série Y_t, a definição de estacionariedade[4] depende da sua média, $\mu_t = E(Y_t)$, e da sua função de autocovariância, $\gamma(s,t) = E[(Y_s - \mu_s)(Y_t - \mu_t)]$, que mede a dependência linear entre duas observações da série (Shumway e Stoffer, 2006). Dizemos que Y_t é estacionária se

i. a média μ_t é constante e não depende de t; e

ii. a função de covariância $\gamma(s,t)$ depende de t e s apenas pela distância $h = t - s$.

Dessa forma, se a série Y_t for estacionária, seu valor médio pode ser representado por $\mu_t = \mu$, para todo $t \geq 0$, e a função de autocovariância $\gamma(s,t)$ pode ser simplificada para $\gamma(h)$. Neste caso, a autocovariância é interpretada como a dependência linear entre observações separadas por h unidades de tempo, independentemente do período em que foram observadas. Sem a estacionariedade, essa função não tem interpretação prática.

A tendência e a sazonalidade são componentes temporais que geram não estacionariedade. A tendência ocorre quando μ_t aumenta ou diminui ao longo do tempo. Se μ_t assume um padrão cíclico, em um intervalo fixo de tempo, dizemos que a série é sazonal. Se a classe de modelos utilizada supõe

[4] Formalmente, a definição de estacionariedade se divide em *fraca* e *forte*. A estacionariedade forte faz restrições sobre as distribuições multivariadas de todos os subconjuntos de observações da série, sendo uma suposição muito limitante para a maioria das aplicações (Shumway e Stoffer, 2006). A definição utilizada neste trabalho corresponde à estacionariedade fraca.

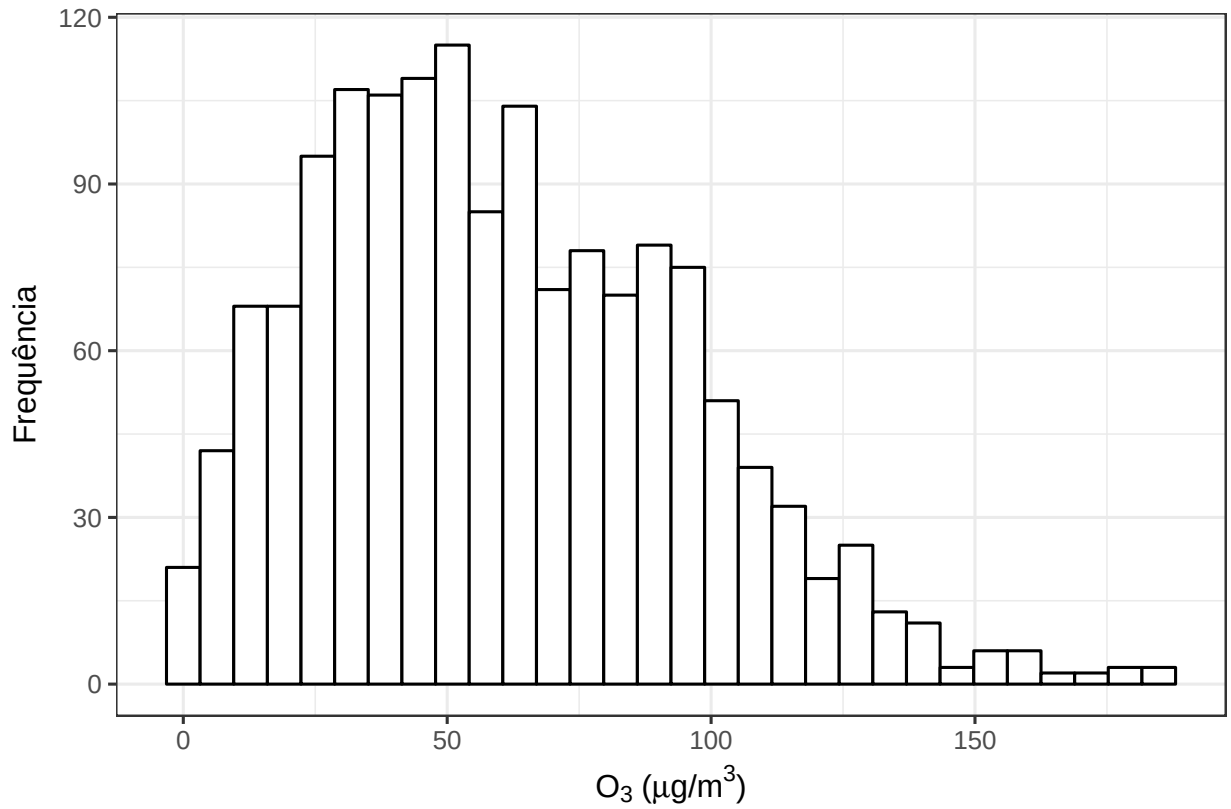

Figura 2.9: *Histograma da concentração diária média de ozônio medida das 12 às 16 horas na estação de monitoramento Dom Pedro II, na cidade de São Paulo, de 2008 a 2013.*

estacionariedade, a melhor estratégia para o ajuste de uma série não estacionária é aplicar alguma transformação na variável original. Se a classe de modelos não supõe estacionariedade, a tendência e a sazonalidade podem ser ajustadas pelo modelo. Discutiremos o primeiro caso nas seções a seguir e o segundo caso no Capítulo 3.

2.2.1 Tendência

A tendência de uma série pode ser eliminada pela utilização da *série de diferenças*. A diferença de primeira ordem é definida como

$$\Delta Y_t = Y_t - Y_{t-1}, \quad t = 1, 2, \ldots.$$

Ela é utilizada para eliminar uma tendência linear de uma série. A ordem da diferença está associada ao grau da tendência. No caso de uma tendência quadrática, por exemplo, podemos utilizar a diferenciação de segunda ordem

$$\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}, \quad t = 1, 2, \ldots.$$

No caso geral, definimos a diferenciação de ordem n como

$$\Delta^n Y_t = \Delta^{n-1} Y_t - \Delta^{n-1} Y_{t-1}, \quad t = 1, 2, \ldots. \tag{2.1}$$

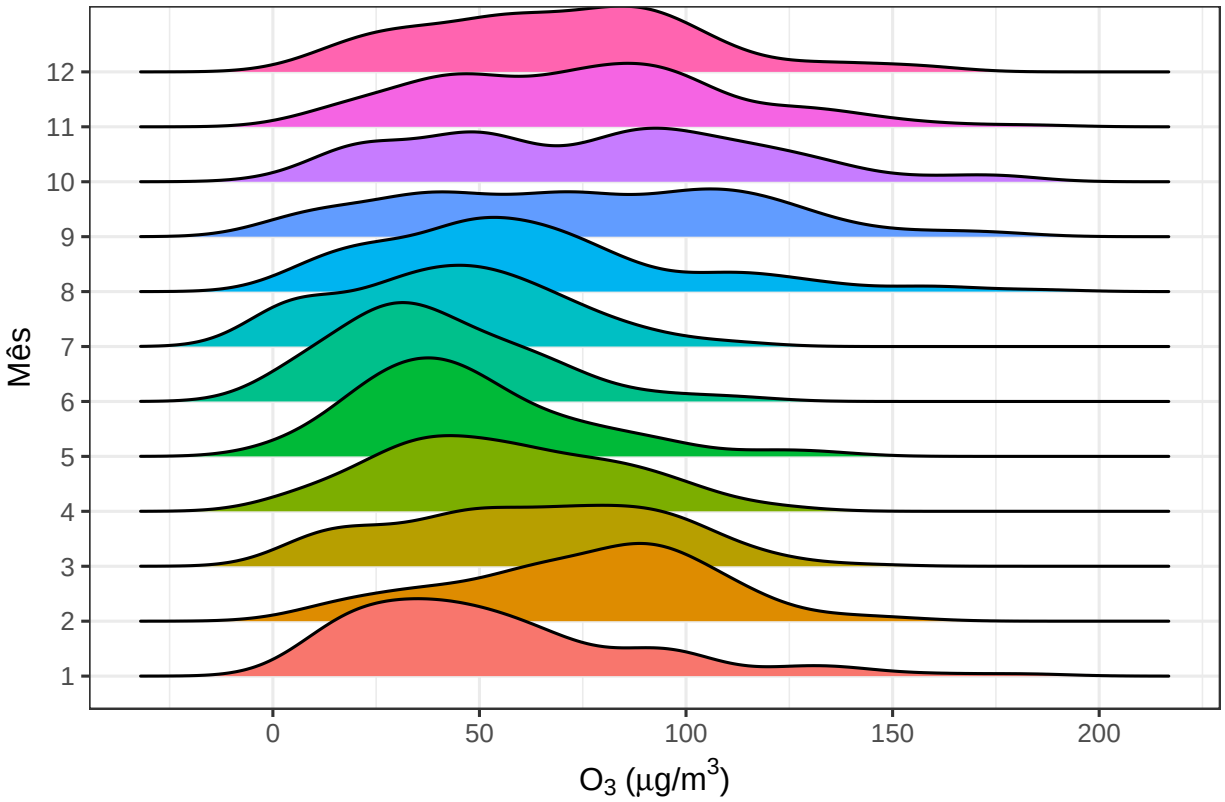

Figura 2.10: *Distribuição por mês da concentração diária média de ozônio medida das 12 às 16 horas na estação de monitoramento Dom Pedro II, na cidade de São Paulo, de 2008 a 2013.*

Na prática, dificilmente encontramos séries com tendência quadrática ou de grau mais elevado, então a diferença de primeiro grau é geralmente suficiente para alcançar a estacionariedade.

Como exemplo, observe a Figura 2.11. No painel (a), temos a série da concentração diária média de ozônio, em que podemos observar uma leve tendência linear positiva, isto é, a concentração média parece crescer com o tempo. No painel (b), apresentamos o gráfico da série de diferenças (primeira ordem). Podemos observar que a série já não apresenta qualquer tendência linear.

Quando o modelo ajustado não assume estacionariedade, podemos incorporar um termo de tendência diretamente no modelo. Discutiremos esse tópico no Capítulo 3.

2.2.2 Sazonalidade

Em geral, o gráfico da série é suficiente para a identificação de sazonalidade. No entanto, em alguns casos, outras variáveis podem mascarar o efeito sazonal, sendo difícil identificar esse componente apenas observando o gráfico. Assim, é sempre recomendável a construção de um *periodograma* para auxiliar a identificação da sazonalidade.

Toda série temporal pode ser decomposta em uma soma de ondas senoidais, com frequências e amplitudes diferentes (Shumway e Stoffer, 2006). Para um conjunto de ondas de frequências diferentes e fixadas a priori, podemos calcular quais são as amplitudes de cada uma dessas ondas para que a soma delas gere a série original. Podemos então definir uma medida de associação linear entre a série original e cada uma das ondas senoidais. Essa medida, chamada de *densidade espectral*, é proporcional à amplitude calculada para cada onda. Assim, quanto maior a densidade espectral associada a uma determinada frequência, maior será a importância dessa frequência para explicar a

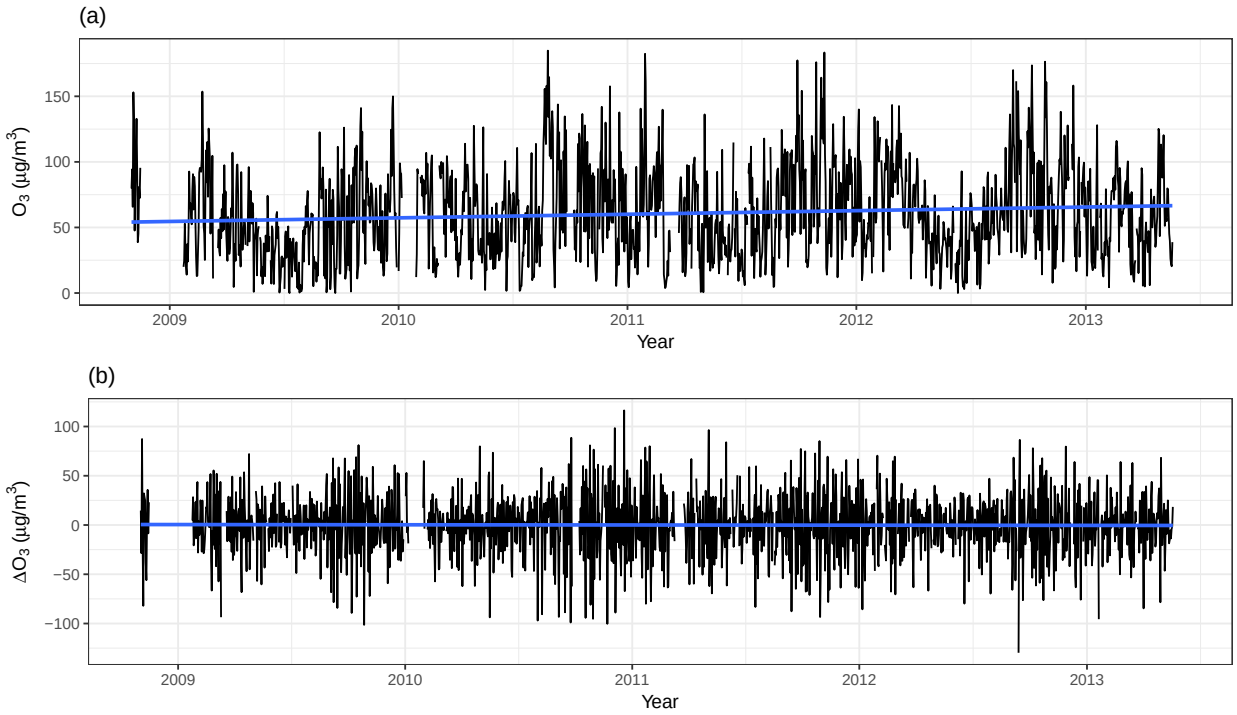

Figura 2.11: *Série da concentração de ozônio diária média, medida na cidade de São Paulo (estação de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2011, das 12h às 16h.*

periodicidade da série. O periodograma é justamente um gráfico da densidade espectral em função das frequências.

Na Figura 2.12, apresentamos o periodograma da série horária de ozônio da cidade de São Paulo de 2008 a 2013. Podemos observar que o período[5] mais importante para explicar a periodicidade da série corresponde a um dia, isto é, o periodograma aponta sazonalidade diária, o que é esperado se observarmos a Figura 2.2.

Existem na literatura técnicas para remover o componente sazonal de uma série (Morettin e Toloi, 2004), mas esse tópico não será abordado aqui. Nosso foco será ajustar modelos que contemplem o componente sazonal, como veremos nos próximo capítulos.

Para mais informações sobre estacionariedade, recomendamos a leitura do primeiro capítulo de Shumway e Stoffer (2006).

2.2.3 Autocorrelação

É natural supor a existência de algum grau de associação entre as observações de uma série temporal coletadas em instantes próximos. Por exemplo, considere a concentração de um poluente medida às nove da manhã em uma certa localidade. Se o valor observado foi alto, as concentrações às oito e às dez da manhã provavelmente também foram altas. Essa informação extraída de Y_t sobre o valor das observações anteriores, $Y_{t-1}, Y_{t-2}, \ldots$, ou das seguintes, $Y_{t+1}, Y_{t+2}, \ldots$, é chamada de *autocorrelação* ou, neste contexto, *correlação temporal*.

Dependendo da forma como as observações estão associadas, podemos definir diferentes tipos de correlação. Uma das medidas mais simples e mais utilizadas na prática se chama *correlação linear*.

[5] O período é o inverso da frequência.

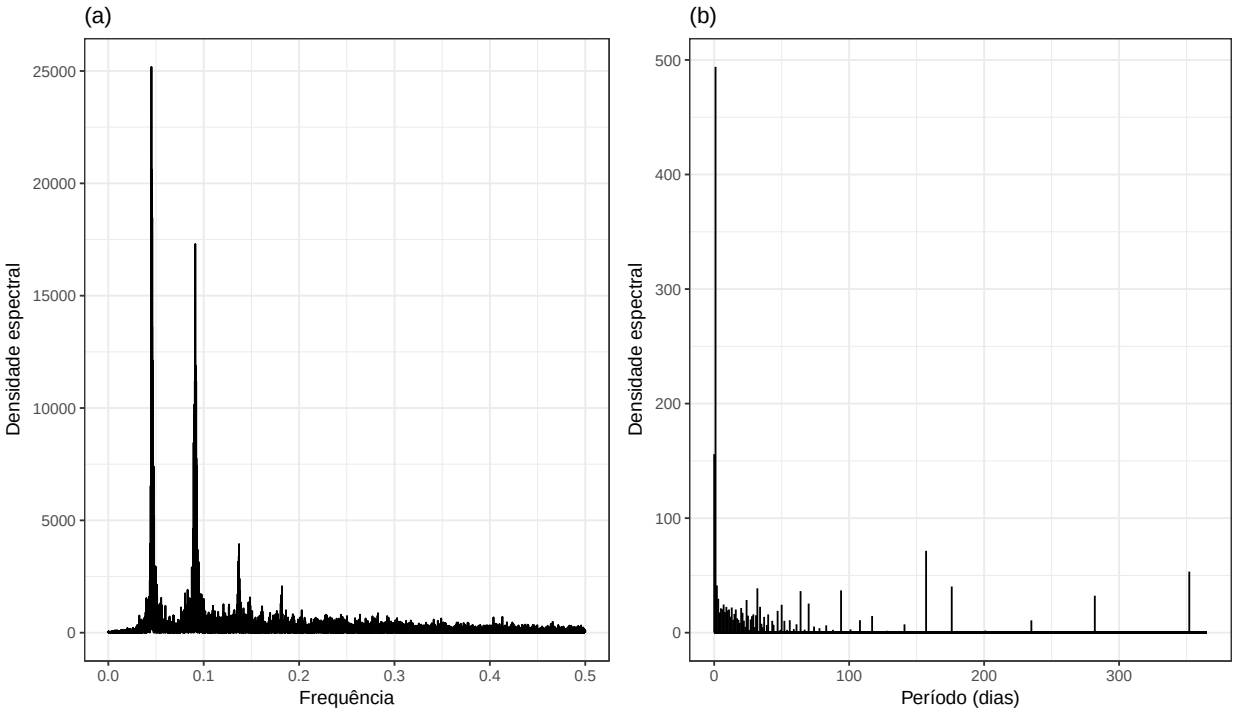

Figura 2.12: *Periodogramas para a concentração horária de ozônio medida na cidade de São Paulo (estação de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2013. Dados disponibilizados por Salvo e Geiger (2014). No painel (a), apresentamos a densidade espectral contra a frequência. No painel (b), resumimos a densidade espectral por período, apresentado em dias.*

Ela supõe que a relação entre as observações pode ser descrita por uma função linear, ou seja, invariante ao valor das observações[6]. Quando outro tipo de relação não for especificada, essa será a definição utilizada neste trabalho para descrevermos a correlação temporal entre as observações.

Embora a função de autocovariância seja útil para medir a associação linear entre observações de uma série, ela não nos fornece a magnitude dessa relação, pois seus valores dependem da grandeza da variável sob estudo. Para contornar esse problema, podemos padronizar a função de autocovariância, restringindo-a a um intervalo fixo. Definimos então a *função de autocorrelação* dada por

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}. \tag{2.2}$$

Não é difícil mostrar[7] que essa medida varia no intervalo $[-1,1]$, com os extremos representando uma correlação perfeita entre as observações Y_t e Y_s. A função de autocorrelação mede a previsibilidade da série no instante t, a partir apenas do valor da variável no instante s. Se Y_t pode ser perfeitamente predita por Y_s por meio de uma função linear, então a autocorrelação será 1, se a associação for positiva, ou -1, se a associação for negativa.

Repare que os termos $\gamma(t,t)$ e $\gamma(s,s)$ em (2.2) representam, respectivamente, a variância de Y_t e Y_s. Se a série é estacionária, sua variância é constante, isto é, $\gamma(t,t) = \gamma(s,s)$, e a função de autocorreção dependerá apenas da diferença $h = s-t$. Neste caso, o gráfico da função de autocorrelação em função de h se torna uma ferramenta descritiva importante para investigarmos se existem relações lineares substanciais entre a série e suas observações defasadas. Sem a suposição de estacionariedade,

[6] Para mais detalhes sobre a interpretação de linearidade, consulte a Seção 3.1.5.
[7] Utilizando a desigualdade de Cauchy-Schwarz (Nicholson, 2001)

o gráfico de autocorrelação pode ser utilizado para detectar tendência e sazonalidade.

A função de autocorrelação pode ser estimada pela função de autocorrelação amostral

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)},$$

sendo

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(y_{t+h} - \bar{y})(y_t - \bar{y}) \tag{2.3}$$

a função de autocovariância amostral, y_t o valor observado no instante t e $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ a média amostral.

Na Figura 2.13, apresentamos a função de autocorreção da concentração de ozônio medida na estação Dom Pedro II. Podemos observar que a autocorrelação é sempre positiva e não decai para o zero, indicando que a série apresenta tendência. Se a série fosse estacionária, esperaríamos que apenas observações próximas fossem correlacionadas, e então a função de autocorrelação convergiria rapidamente para zero conforme aumentássemos o valor de h.

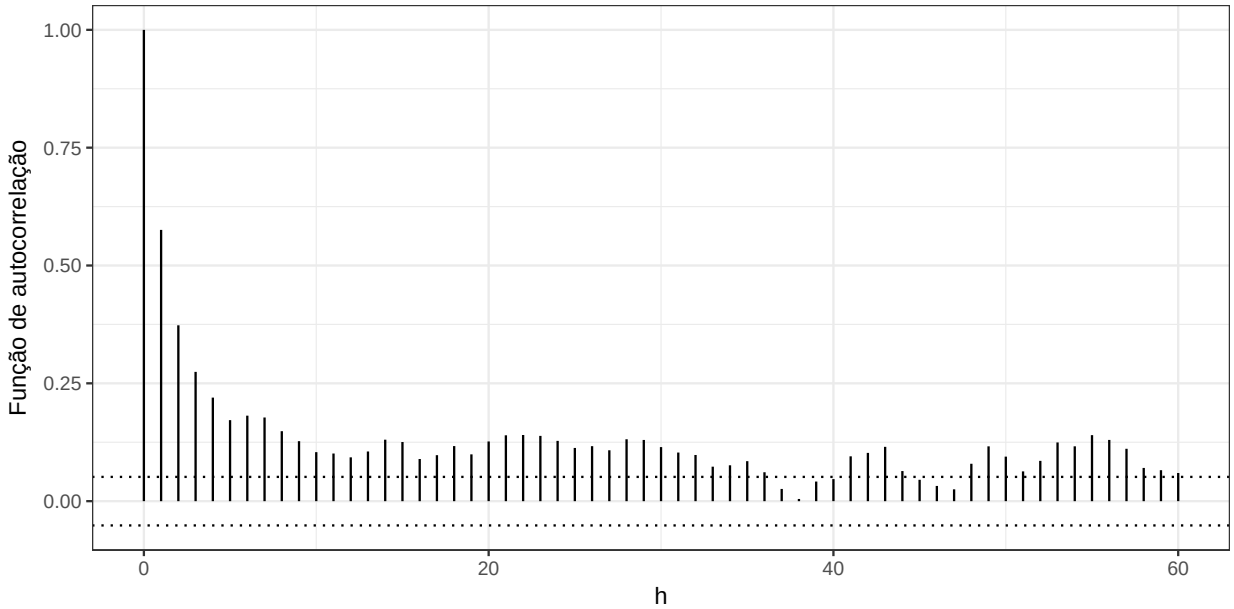

Figura 2.13: *Função de autocorreção da concentração de ozônio diária média, medida na cidade de São Paulo (estação de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2011, das 12h às 16h. Dados disponibilizados por Salvo e Geiger (2014). As linhas pontilhadas representam os limites $\pm 2/\sqrt{n}$, sendo n o tamanho da amostra. Valores fora desse intervalo de confiança (95%) podem ser considerados significantemente diferentes de zero.*

Note que se as observações Y_t e Y_{t-1} são correlacionadas, e da mesma forma as observações Y_{t-1} e Y_{t-2}, parte da correlação entre Y_t e Y_{t-2} pode ser explicada por Y_{t-1}. Como a função de autocorrelação nos dá a correlação total entre Y_t e Y_{t-2}, independentemente do fato de parte dela poder ser explicada por Y_{t-1}, se quisermos encontrar apenas a variabilidade explicada por Y_{t-2} precisamos utilizar a *função de autocorrelação parcial*. No caso geral, essa função mede a correlação entre as observações Y_t e Y_{t-m}, controlando pelas observações intermediárias $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-m+1}$.

Na Figura 2.14, apresentamos a função de autocorreção parcial da concentração de ozônio, como

no exemplo anterior. Podemos observar agora que a maioria das defasagens são não significativas. Mesmo assim, ainda encontramos algumas defasagens altas significativos, indicando que a série realmente não é estacionária.

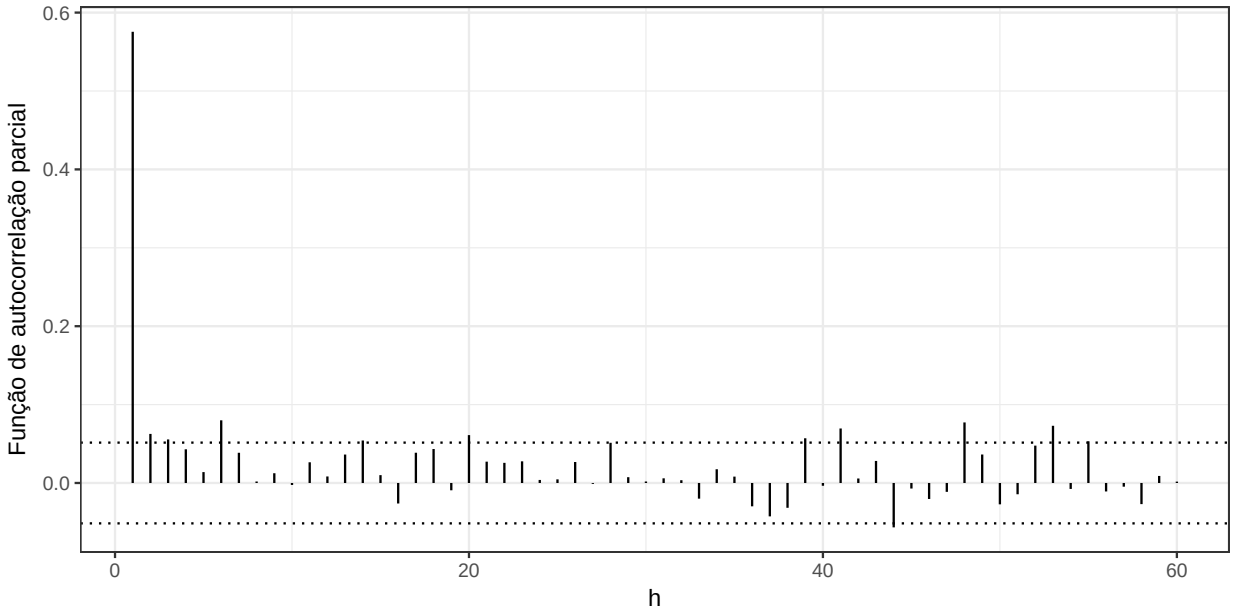

Figura 2.14: *Função de autocorreção parcial da concentração de ozônio diária média, medida na cidade de São Paulo (estação de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2011, das 12h às 16h. As linhas pontilhadas representam os limites $\pm 2/\sqrt{n}$, sendo n o tamanho da amostra. Valores fora desse intervalo de confiança (95%) podem ser considerados significantemente diferentes de zero.*

Até agora discutimos como avaliar a correlação entre observações defasadas de uma mesma série. A seguir, vamos discutir como avaliar a associação entre observações de duas ou mais séries.

2.2.4 Função de correlação cruzada

Muitas vezes, queremos avaliar a previsibilidade de uma determinada série Y_t a partir de outra série, digamos X_s. Nesse caso, definimos a *função de correlação cruzada*

$$\rho_{XY}(s,t) = \frac{\gamma_{XY}(s,t)}{\sqrt{\gamma_X(s,s)\gamma_Y(t,t)}} \tag{2.4}$$

sendo,

$$\gamma_{XY} = E\left[(X_s - \mu_s)(Y_t - \mu_t)\right]$$

a função de covariância cruzada. Se as séries Y_t e X_s forem estacionárias, a expressão (2.4) se reduz a

$$\rho_{XY}(h) = \frac{\gamma_{XY}(h)}{\sqrt{\gamma_X(0,0)\gamma_Y(0,0)}}. \tag{2.5}$$

Essa expressão nos dá a relação entre Y_t e X_{t+h}, para todo $t \geq 0$. Assim, valores positivos de h revelam o quanto Y_t antecipa X_{t+h} e valores negativos de h o quanto X_{t+h} antecipa Y_t. Repare que $\rho_{XY}(h) = \rho_{YX}(-h)$.

A função de correlação cruzada pode ser estimada pela *função de correlação cruzada amostral* definida por

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}},$$

sendo $\hat{\gamma}(h)$ como definido em (2.3) e

$$\hat{\gamma}_{xy}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y})$$

a função de covariância cruzada amostral.

Em estudos de poluição do ar, é muito comum a inclusão de variáveis defasadas na análise. Essas variáveis representam fenômenos que antecipam a formação de um poluente ou a ocorrência de doenças. Uma chuva no período da manhã, por exemplo, além de alterar o trânsito, pode diminuir a concentração de poluentes no começo da tarde. Altos níveis de poluentes em um determinado dia, podem aumentar o número de internações por problemas respiratórios dias ou até semanas depois.

A identificação de quais variáveis defasadas devem entrar na análise pode ser uma tarefa difícil, principalmente quando existe muita incerteza sobre o processo de geração do fenômeno sob estudo. A função de correlação cruzada é uma boa alternativa neste caso. Com ela, podemos avaliar quais são os valores da defasagem h que geram maior correlação entre as séries e utilizá-los para definir as variáveis defasadas.

A Figura 2.15 apresenta a função de correlação cruzada do ozônio em função da temperatura na estação Dom Pedro II. Ambas as medidas são horárias. Observamos que a maior correlação (após a defasagem zero) é na defasagem -1, isto é, a concentração de ozônio parece ser altamente associada com a temperatura medida uma hora antes. Assim, a temperatura no instante $t-1$ é uma boa candidata para ser incluída no modelo.

Em alguns casos, quando o fenômeno associado não varia muito no tempo, podemos considerar a média de um certo intervalo como variável defasada. Pela Figura 2.15, observamos uma certa correlação entre o ozônio e a temperatura nas defasagens de -5 a -8. Assim, a média da temperatura medida entre $t-8$ e $t-5$ também poderia ser incluída no modelo.

As técnicas abordadas até aqui podem ser utilizadas para obter um conhecimento inicial sobre o fenômeno estudado, auxiliando-nos a escolher a melhor estratégia de modelagem. No próximo capítulo, apresentaremos os principais modelos utilizados em análises envolvendo poluição do ar.

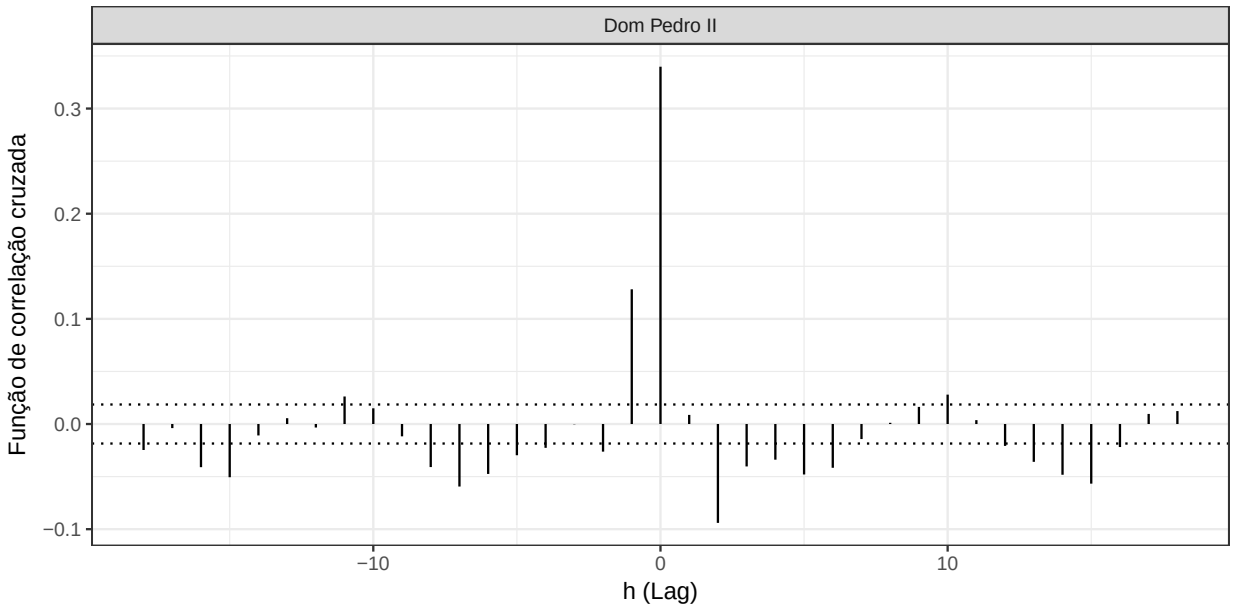

Figura 2.15: *Função de correlação cruzada do ozônio em função da temperatura na estação Dom Pedro II (São Paulo) no período de outubro de 2009 a junho de 2011. As linhas pontilhadas representam os limites* $\pm 2/\sqrt{n}$*, sendo* n *o tamanho da amostra. Valores fora desse intervalo de confiança (95%) podem ser considerados significantemente diferentes de zero.*

Capítulo 3

Estratégias usuais de modelagem

> Data will often point with almost
> equal emphasis on several possible
> models, and it is important that the
> statistician recognize and accept this.
> — McCullah and Nelder (1989)

O grande objetivo de uma análise estatística é usar um conjunto de dados para gerar conhecimento sobre um fenômeno de interesse. Podemos pensar nesse fenômeno como um mecanismo da natureza, desconhecido e complexo, no qual um conjunto de variáveis explicativas $\mathbf{X}$ são transformadas em uma variável resposta Y[1] (Figura 3.1). Os dados são o resultado desse processo (Breiman, 2001).

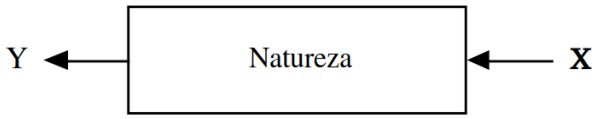

Figura 3.1: *Esquematização do mecanismo gerador dos dados.*

No contexto da modelagem estatística supervisionada[2] (Hastie *et al.*, 2008), dada a variável resposta Y e o vetor de variáveis explicativas $\mathbf{X} = (X_1, \ldots, X_p)$, queremos encontrar funções f's tais que

$$Y \approx f(\mathbf{X}), \tag{3.1}$$

isto é, queremos uma função $f(\cdot)$ que descreva o mecanismo gerador dos dados da forma mais precisa possível. A partir dessa função, poderíamos tanto fazer predições — descobrir qual é o novo valor de Y para novas observações $\mathbf{X}$ — quanto inferência — investigar como as variáveis $\mathbf{X}$ e Y estão relacionadas.

A expressão (3.1) representa diversas classes de modelos, a depender da escolha de $f(\cdot)$. De uma forma geral, modelos estatísticos são simplificações da realidade e, por isso, estão sujeitos a erros.

[1] Também podemos ter o caso multivariado, em que são geradas um conjunto de variáveis respostas $\mathbf{Y}$.

[2] No qual uma variável resposta *supervisiona* a estimação dos parâmetros do modelo. Na prática, são os casos em que temos acesso a uma amostra da variável resposta.

Quando modelamos a série de um poluente, por exemplo, estamos supondo que a sua concentração ao longo do tempo pode ser aproximada por uma função matemática. Neste caso, o erro do modelo quantifica o quanto a nossa função se afasta do verdadeiro mecanismo gerador do poluente. Parte desse erro é irredutível e se deve a impedimentos práticos, como erros de medida, variáveis que não podem ser observadas e desconhecimento de outros fatores que influenciam o fenômeno. No entanto, o erro total pode ser minimizado pela escolha adequada do modelo utilizado, o que torna essencial o desenvolvimento de estratégias de modelagem que contemplem as particularidades de cada estudo. Assim, podemos reescrever (3.1), já no contexto de séries temporais, como

$$Y_t = f(\mathbf{X}_t) + \epsilon_t, \tag{3.2}$$

sendo ϵ_t um erro aleatório, isto é, um componente que representa toda a informação de Y_t que não pode ser explicada pelos preditores $\mathbf{X}_t$. Apesar de a expressão (3.1) ser mais intuitiva, (3.2) é mais conveniente para a formulação dos modelos estatísticos.

Na prática, há duas abordagens bastante utilizadas na especificação da função $f(\cdot)$. A primeira consiste em supor um modelo probabilístico para o fenômeno sob estudo, de tal forma que $f(\cdot)$ seja uma função dos parâmetros de alguma distribuição conhecida, que podem ser estimados a partir dos dados. Essa estratégia geralmente produz modelos interpretáveis, que trazem informação sobre a relação da variável resposta e os preditores, e por isso é preferível quando o interesse é fazer inferência. A segunda abordagem é mais flexível e permite que os próprios dados definam uma estrutura para $f(\cdot)$. Essa estratégia dificilmente gera modelos interpretáveis, pois a complexidade do mecanismo gerador dos dados é refletida na função resultante. Por outro lado, a maior flexibilidade leva a uma maior precisão desses modelos, sendo muito utilizados para predição. A abordagem escolhida deve atender aos objetivos do estudo. Nesta tese, exploraremos exemplos de ambos os casos.

Neste capítulo, introduziremos os principais modelos interpretáveis utilizados na literatura para análise de dados de poluição do ar, como o modelo de regressão linear, modelos aditivos e modelos para séries temporais. No Capítulo 4, discutiremos técnicas focadas em predição, como validação cruzada e regularização, mas que muitas vezes também podem ser utilizadas para inferência.

3.1 Regressão linear

O modelo de regressão linear corresponde à aproximação (3.1) mais simples e bem estabelecida dentro da modelagem estatística. Mesmo com a disponibilidade de modelos mais flexíveis, essa classe de modelos ainda é bastante utilizada hoje em dia, principalmente por se ajustar bem a diversos problemas reais, facilidade de interpretação dos resultados e estar disponível nos principais programas estatísticos.

Em estudos de poluição do ar, modelos de regressão linear podem ser ajustados para investigar a relação entre variáveis explicativas e uma variável resposta, seja a concentração de poluentes ou dados epidemiológicos. Saldiva *et al.* (1995), por exemplo, utilizou esses modelos para estudar o efeito de alguns poluentes nas taxas de mortalidade de idosos, controlando por condições climáticas e sazonais. Já Salvo *et al.* (2017) utilizou para associar os níveis de material particulado com a proporção de carros a álcool e gasolina.

Apesar da sua popularidade, a complexidade presente nos estudos de poluição atmosférica, como

relações não-lineares e autocorrelação, pode desqualificar o modelo de regressão linear como a opção mais adequada para o ajuste dos dados. Pela sua facilidade de implementação e interpretação, ele é uma boa ferramenta para uma análise preliminar.

Nas próximas seções, especificaremos o modelo de regressão linear, discutiremos as suas restrições e apresentaremos as maneiras mais utilizadas para tratar séries com tendência, sazonalidade e autocorrelação.

3.1.1 Especificação do modelo

Seja Y_t a variável resposta, $\mathbf{X}_t = (X_{1t}, ..., X_{pt})$ um vetor de variáveis explicativas cuja associação com Y_t estamos interessados em avaliar e $t = 1, \ldots, n$ a ordem na qual essas variáveis foram medidas. Aqui, não faremos suposições sobre a natureza dos preditores $\mathbf{X}_t$, isto é, essas variáveis podem ser fixas ou aleatórias, qualitativas ou quantitativas. Dado os vetores de parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$, o modelo de regressão linear pode ser definido por

$$Y_t = \alpha + \beta_1 X_{1t} + ... + \beta_p X_{pt} + \epsilon_t, \quad t = 1, ..., n. \tag{3.3}$$

Em geral, supomos que os erros $(\epsilon_1, ..., \epsilon_n)$ tenham média zero, variância constante (homoscedasticidade) e sejam não-correlacionados[3]. Além disso, a especificação (3.3) impõe que a relação entre a resposta Y_t e os preditores $\mathbf{X}_t$ seja linear e aditiva.

A suposição de linearidade estabelece que a variação esperada em Y_t causada pelo acréscimo de uma unidade em X_{it}, mantidos fixados os outros preditores, é constante e não depende do valor de X_{it}. A interpretação dos coeficientes será discutida com mais detalhes na Seção 3.1.5 e nas aplicações dos Capítulos 5 e 6. Conceitos mais gerais podem ser encontrados em Hastie *et al.* (2008) e James *et al.* (2013).

A suposição de aditividade estabelece que a variação esperada em Y_t causada por uma mudança no preditor X_{it} independe do valor (fixado) dos outros preditores. Essa suposição pode ser relaxada com a introdução de termos de interação (ver Seção 3.3 de James *et al.* (2013)), que abordaremos na Seção 3.1.6.

Na prática, os coeficientes $\beta_1, ..., \beta_p$ são desconhecidos e precisam ser estimados. O procedimento de estimação mais utilizado é o método de mínimos quadrados (Hastie *et al.*, 2008). Outro método bastante utilizado é a estimação por máxima verossimilhança (Casella e Berger, 2001). Sob a suposição de normalidade, as duas abordagens são equivalentes.

Como o instante em que as observações omissas ocorrem não é relevante no processo de estimação, o modelo linear é uma alternativa para avaliar a associação de séries com "buracos" ou grandes períodos sem informação, apesar de a identificação da estrutura de tendência e sazonalidade ser mais difícil em dados com essa característica.

A adequação do modelo é avaliada a partir de medidas de qualidade de ajuste, como o R^2 e o erro quadrático médio, e da *análise de resíduos*. A partir da expressão (3.3), para $t = 1, ..., n$, podemos definir os resíduos como

[3] A suposição de distribuição Normal também é feita em alguns casos. Essa suposição é relevante na construção de intervalos de confiança e testes de hipóteses para os coeficientes do modelo. No entanto, para amostras grandes, característica comum em estudos de poluição do ar, existem resultados assintóticos (Casella e Berger, 2001) que garantem a validade desses procedimentos.

$$r_t = Y_t - \widehat{Y}_t, \tag{3.4}$$

em que $\hat{Y}_t$ representa o valor predito de Y_t com base nas estimativas dos coeficientes do modelo. Os resíduos medem o quanto os valores preditos se afastam dos valores observados, sendo muito úteis para avaliar a qualidade do ajuste e a violação das suposições do modelo. Esse tópico será discutido com mais detalhes na Seção 3.1.7.

No R, os modelos de regressão linear podem ser ajustados via mínimos quadrados com a função `lm()` do pacote `stats` ou utilizando a função `train` do pacote `caret` com `method = "lm"`. O pacote `caret` traz uma abordagem padronizada para o ajuste de modelos estatísticos.

A seguir, abordaremos como modelar tendência e sazonalidade utilizando o modelo de regressão linear.

3.1.2 Incorporando tendência e sazonalidade

Séries de poluição do ar não costumam ser estacionárias. Como vimos nos exemplos do Capítulo 2, é comum encontrarmos tendência (positivas ou negativas) e diversos tipos de sazonalidade (diária, semanal, anual etc). Fatores como crescimento populacional, industrialização, aumento da frota de automóveis, leis de regulamentação de combustíveis, entre outros, podem gerar mudanças a longo prazo na concentração de poluentes, alterando o comportamento da série, e muitas vezes não temos informação disponível para incorporá-los no modelo.

Como o modelo de regressão linear não faz suposições sobre a estacionariedade da variável resposta, podemos modelar a tendência e a sazonalidade da série incluindo preditores que controlam esses componentes[4]. A inclusão desses termos é interessante principalmente nos casos em que não estamos interessados em estudar a evolução da série, mas sim o efeito de preditores na variável resposta, independentemente desses componentes.

Para acrescentar um termo de tendência linear ao modelo (3.3), podemos especificar $X_{1t} = t$, $t = 1, ..., n$. Assim, um coeficiente β_1 positivo em (3.3) indica que Y cresce linearmente com o tempo, enquanto um coeficiente negativo indica que Y decresce linearmente com o tempo. Podemos definir outras formas para a tendência, como quadrática, $X_{1t} = t^2$, ou logarítmica, $X_{1t} = \log(t)$. A Figura 3.2 mostra um exemplo de séries com tendências linear e quadrática.

Note que, se modelarmos a tendência dessa maneira, estamos impondo a mesma função ao longo de todo período observado. Em alguns casos, a tendência pode ser diferente em certos intervalos de tempo (Figura 3.3). Uma alternativa seria definir um termo de tendência para cada intervalo, por exemplo:

$$X_{1t} = \begin{cases} t, & \text{se } t \text{ pertence ao conjunto } \{1, 2, ..., m\}; \text{ e} \\ 0, & \text{em caso contrário.} \end{cases}$$

e

$$X_{2t} = \begin{cases} t - m, & \text{se } t \text{ pertence ao conjunto } \{m+1, m+2, ..., m+n\}; \text{ e} \\ 0, & \text{em caso contrário.} \end{cases}$$

A sazonalidade também pode ser controlada por meio de variáveis explicativas. A sua presença

[4] Em vez de transformar a série original, como discutido na Seção 2.2. As vantagens de se incluir um termo de tendência ao modelo, em vez de se transformar Y_t, são: (1) poder interpretar os coeficientes do modelo em função da variável original e (2) estimar a tendência da série.

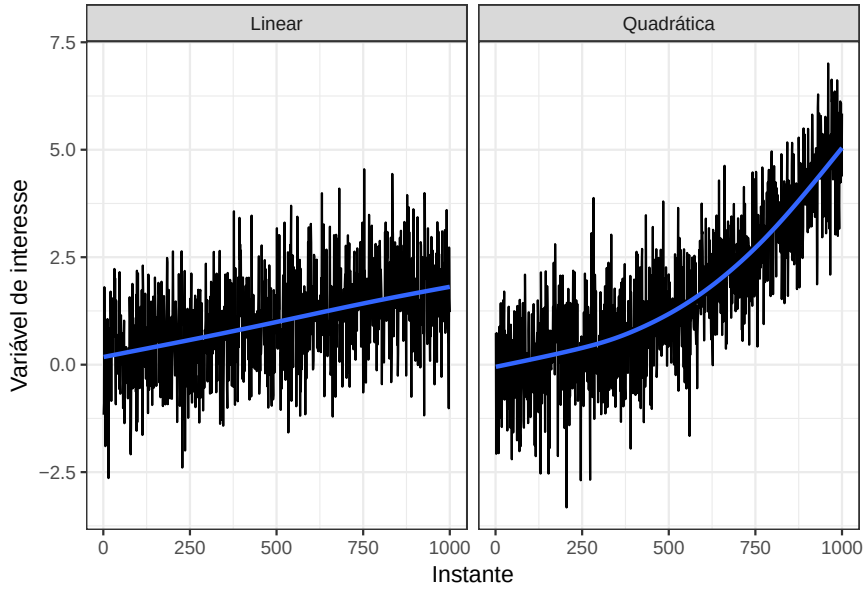

Figura 3.2: *Exemplos de séries com tendência linear e quadrática, ambas positivas.*

indica que a média da variável resposta está associada a efeitos periódicos, ligados a intervalos de tempo, como dias, semanas, meses, estações do ano, temporadas de chuva etc. Os níveis de ozônio, por exemplo, crescem no verão e diminuem no inverno; o número de problemas respiratórios tende a aumentar nos meses mais secos; e a concentração de diversos poluentes varia nos fins de semana, devido à menor intensidade de tráfego.

De uma maneira geral, podemos classificar a sazonalidade como *determinística* — o padrão é constante ao longo do tempo — ou *estocástica* — o padrão muda ao longo do tempo. É possível controlar a sazonalidade determinística no modelo (3.3) a partir de variáveis indicadoras. Se, por exemplo, acreditamos que há um efeito sazonal de mês, podemos adicionar ao modelo 11 variáveis indicadoras X_{it}, $i = 1, ..., 11$ tais que

$$X_{it} = \begin{cases} 1, & \text{se a observação } t \text{ pertence ao } i\text{-ésimo mês do ano; e} \\ 0, & \text{caso contrário.} \end{cases} \tag{3.5}$$

Com essa formulação, o mês de dezembro será tomado como referência, isto é, a interpretação dos coeficientes correspondentes aos meses será feita sempre em relação ao mês de dezembro.

A inclusão de variáveis indicadoras também pode ser feita para controlar o efeito de variáveis que não estão disponíveis na amostra. Belusic *et al.* (2015), por exemplo, utilizaram variáveis indicadoras para a hora do dia com o objetivo de controlar o efeito do trânsito no monitoramento de diversos poluentes na cidade de Zagreb, na Croácia. Dessa forma, cada coeficiente explicará as condições específicas da hora do dia a que ele se refere. Uma desvantagem dessa estratégia é não podermos avaliar se o coeficiente de fato está capturando o efeito do trânsito ou qualquer outra variável associada com a concentração dos poluentes que também varia a cada hora.

Para mais informações sobre a utilização de variáveis indicadoras em modelos de regressão, consultar a Seção 3.3.1 de James *et al.* (2013).

Se a sazonalidade for estocástica, procedimentos um pouco mais sofisticados serão necessários para controlá-la. Não trataremos desse tópico neste trabalho. Mais informações podem ser encontradas em Shumway e Stoffer (2006).

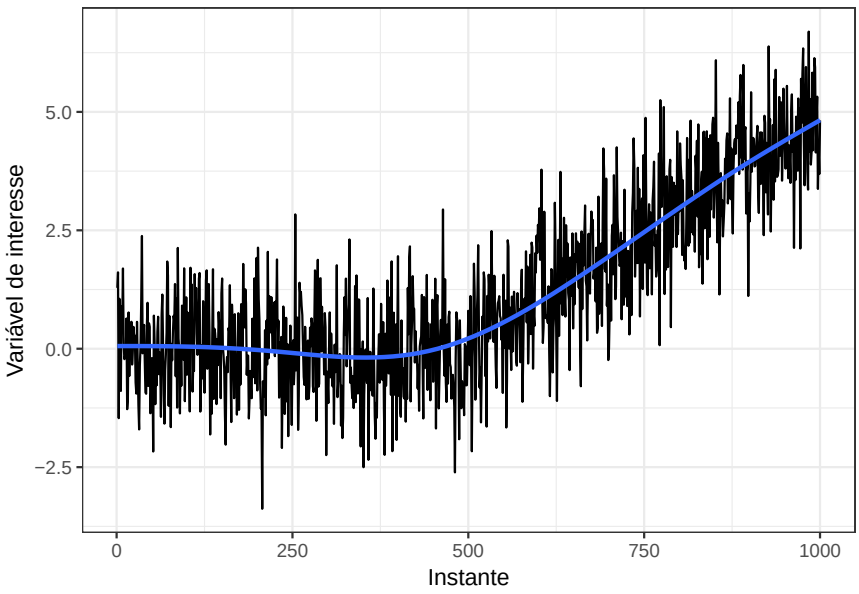

Figura 3.3: *Exemplos de uma série com tendência não-constante.*

A seguir, discutiremos como contornar as suposições de erros não-correlacionados, homoscedasticidade, linearidade e aditividade utilizando o modelo de regressão linear.

3.1.3 Tratando erros correlacionados

O modelo de regressão linear supõe que os erros $(\epsilon_1, \ldots, \epsilon_n)$ sejam não-correlacionados. Em estudos de poluição do ar, essa suposição é, em geral, inadequada. Como discutimos na Seção 2.2.3, é natural que observações de uma série temporal sejam autocorrelacionadas. A formação de gases na atmosfera, por exemplo, é um processo contínuo ao longo do tempo, sendo que as concentrações medidas no instante t podem estar fortemente associadas aos níveis observados nas últimas horas ou mesmo nos últimos dias.

Uma forma de avaliar a violação dessa suposição é construir o gráfico dos resíduos do modelo em função do tempo. A presença de padrões na sequência de pontos, isto é, resíduos adjacentes com valores próximos, é um indício de correlação. Na Figura 3.4, apresentamos os resíduos de um modelo de regressão linear ajustado em dados auto-correlacionados e em dados não-correlacionados. Para o primeiro caso, observe que os pontos adjacentes tendem a permanecer em um mesmo lado da reta $y = 0$. Na ausência de correlação, temos uma sequência aleatória de valores positivos e negativos.

Se as observações são muito correlacionadas, os erros-padrão estimados pelo modelo de regressão linear tenderão a subestimar os verdadeiros erros, o que comprometeria a inferência, já que os valores p associados seriam menores do que deveriam ser. Nesses casos, outras estratégias de modelagem devem ser adotadas.

Outro tipo de correlação muito comum é a causada por observações que pertencem a um mesmo grupo. Indivíduos de uma mesma família, por exemplo, compartilham a mesma genética e tendem apresentar respostas correlacionadas em estudos epidemiológicos. A localidade também configura formação de grupos, já que pessoas que moram numa mesma região geralmente estão expostas às mesmas condições ambientais. Observações realizadas em diferentes localizações, mas no mesmo instante também podem apresentar correlação. Salvo *et al.* (2017) utilizaram concentrações horárias de alguns poluentes em diversas estações de monitoramento em São Paulo, e é natural supor que

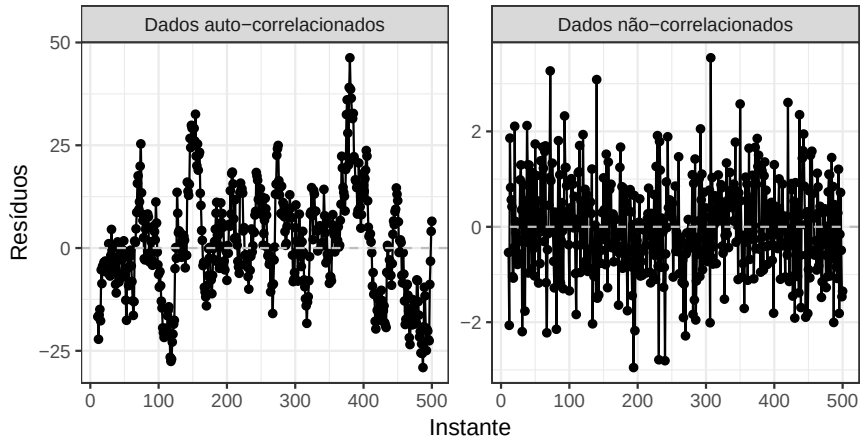

Figura 3.4: *Comparação entre os gráficos dos resíduos de um modelo linear contra o tempo para dados auto-correlacionados e dados não correlacionados.*

as medidas feitas na mesma hora ou no mesmo dia estão correlacionadas.

A depender dos objetivos do estudo, agregar os dados pode ser boa uma alternativa para reduzir o efeito da correlação. Se estamos trabalhando com uma série horária e não temos o objetivo de investir a relação entre as variáveis ao longo do dia, podemos simplificar o problema utilizando a série de médias diárias (veja o exemplo discutido na Seção 5.2). Assim, eliminamos a correlação gerada pelas medidas realizadas dentro do mesmo dia.

Para reduzir o efeito da correlação na estimação da variabilidade dos coeficientes, Salvo *et al.* (2017) utilizaram métodos robustos para o cálculo do erro-padrão. Os chamados *clustered standard errors* (Cameron e Miller, 2015) são obtidos a partir de uma especificação da matriz de variâncias e covariâncias que contempla a correlação entre indivíduos de um mesmo grupo. Essa técnica tem como vantagem a necessidade de especificar um modelo para os dados agrupados, mas faz a suposição que o número de grupos tende ao infinito.

Outra alternativa consiste na utilização de modelos que permitem a especificação das observações correlacionadas, como os modelos mistos (Demidenko, 2013; McCulloch e Searle, 2001).

3.1.4 Contornado a suposição de homoscedasticidade

Assim como a média, a variância de Y também pode mudar segundo algum preditor ou o próprio tempo, violando a suposição de homoscedasticidade do modelo de regressão linear. Nesses casos, precisamos escolher entre utilizar modelos mais flexíveis, que contemplem variância não-constante, ou aplicar transformações que estabilizem a variância das informações.

O gráfico dos resíduos em função dos valores preditos é uma boa ferramenta para identificar heteroscedasticidade. Como podemos observar na Figura 3.5, nuvens de pontos em forma de funil são indícios de observações heteroscedásticas: a variância é maior para valores preditos menores.

Uma maneira de estabilizar a variância das observações é transformar a variável Y usando funções côncavas, como $\log Y$ e $\sqrt{Y}$. Uma outra alternativa consiste em ponderar as observações com pesos proporcionais ao inverso de sua variância, mas essa técnica se limita aos casos em que a variabilidade pode ser estimada com precisão.

Os modelos lineares generalizados duplos (Paula, 2013) e os modelos mistos (Demidenko, 2013; McCulloch e Searle, 2001) são alternativas aos modelos de regressão linear que modelam também

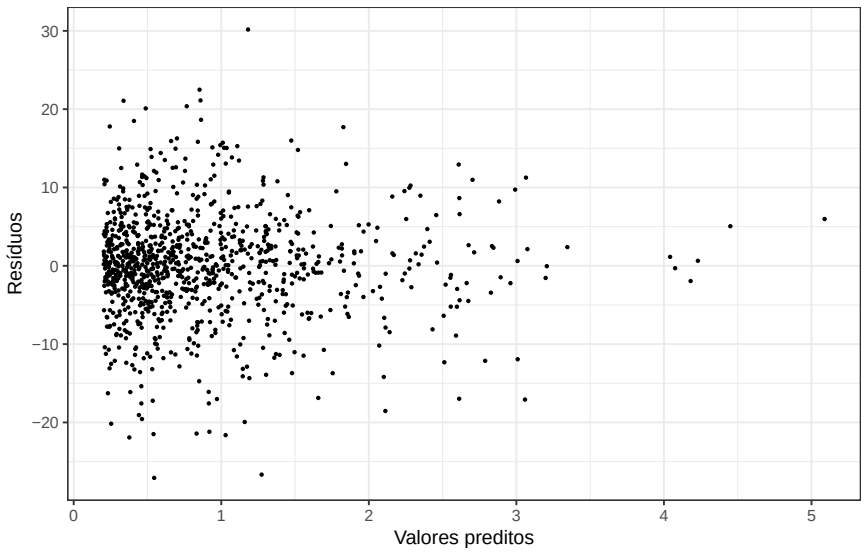

Figura 3.5: *Gráfico dos resíduos contra os valores preditos. Exemplo de nuvem de pontos em forma de funil, indicando heteroscedasticidade.*

a variância das observações.

3.1.5 Contornado a suposição de linearidade

Para entendermos melhor a suposição de linearidade, vamos considerar o modelo de regressão linear mais simples, com apenas um preditor:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \quad t = 1, \ldots, n. \tag{3.6}$$

Ao estimarmos os parâmetros β_0 e β_1 (pelo método de mínimos quadrados, por exemplo), obtemos a seguinte reta de regressão

$$\widehat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t, \quad t = 1, \ldots, n, \tag{3.7}$$

sendo $\widehat{Y}_t$ o valor de Y_t predito pelo modelo e $\hat{\beta}_0$ e $\hat{\beta}_1$ as estimativas de β_0 e β_1 respectivamente. Note que (3.7) representa a equação de uma reta com intercepto $\hat{\beta}_0$ e coeficiente angular $\hat{\beta}_1$. Isso significa que essa reta cruza o eixo y no ponto $\hat{\beta}_0$ e, se variamos o valor de X_t em uma unidade, $\widehat{Y}_t$ vai variar $\hat{\beta}_1$ unidades, não importa qual seja o valor de X_t. Essa associação entre $\widehat{Y}_t$ e X_t (ou Y_t e X_t) é dita ser *linear* com respeito aos parâmetros e está ilustrada na Figura 3.6, para $\hat{\beta}_0$ igual a 0 e $\hat{\beta}_1$ igual a 10. Quando temos mais de um preditor, como no modelo (3.3), a interpretação é análoga para cada par $(\widehat{Y}_t, X_{it})$, mantendo-se as outras variáveis fixadas.

Repare que a suposição de linearidade é mais forte do que apenas monotonicidade. Enquanto esta última restringe que a associação entre as variáveis seja sempre crescente ou decrescente, a primeira também restringe o quanto a variável resposta varia quando o preditor aumenta ou diminui em uma unidade. Essa diferença é importante, pois muitas vezes utilizamos modelos lineares na tentativa de explicar relações que são apenas monotônicas, o que pode levar a estimativas pouco confiáveis e conclusões equivocadas. Uma discussão mais detalhada sobre esse problema pode ser encontrada em Achen (2005).

Os resíduos, definidos pela expressão (3.4), podem ser utilizados para avaliar se a suposição

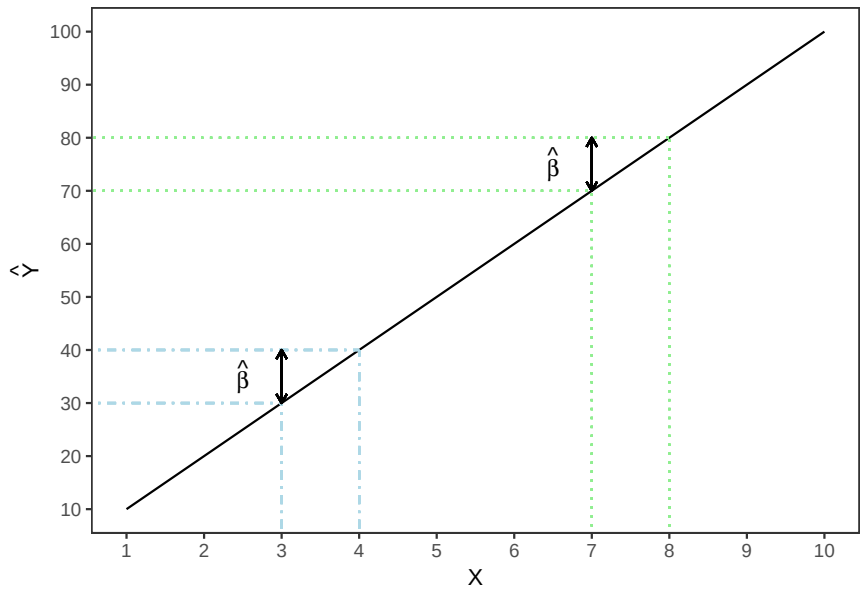

Figura 3.6: *A estimativa $\hat{\beta}$ representa a variação em Y quando acrescemos X em uma unidade, não importando o valor de X.*

de linearidade é razoável. A ideia consiste em construir o gráfico dos resíduos contra os valores preditos e verificar se a nuvem de pontos apresenta algum padrão. Nuvens em forma de "U", por exemplo, mostram que o modelo não está bem ajustado para valores extremos de Y, indicando não-linearidade (veja Figura 3.7).

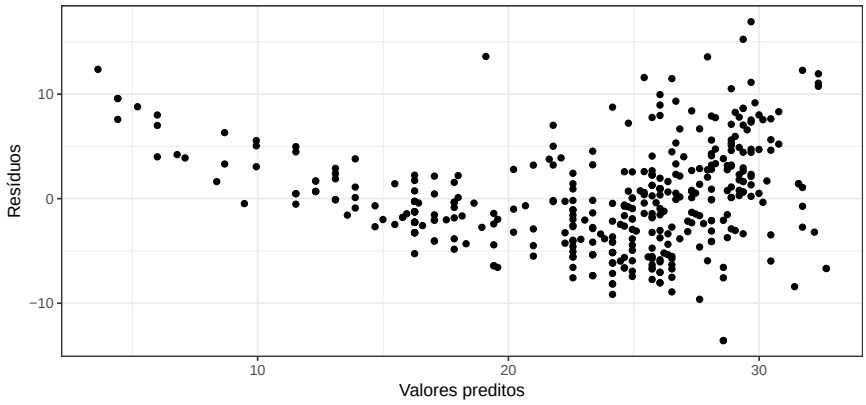

Figura 3.7: *Gráfico dos resíduos contra os valores preditos, um exemplo de nuvem de pontos em forma de "U", indicando não-linearidade.*

Uma maneira simples de contornar esse problema é ajustar modelos da forma

$$Y_t = \beta_0 + \beta_1 T(X_t) + \epsilon_t, \quad t = 1, \ldots, n, \tag{3.8}$$

em que $T(\cdot)$ representa uma função "linearizadora". As escolhas mais comuns para $T(X)$ são $\log X$ e $\sqrt{X}$. Observe que, embora a relação entre Y e X em (3.8) não seja mais linear, o modelo continua sendo linear nos parâmetros. Um ponto negativo nessa abordagem é a perda de interpretabilidade do modelo, já que os parâmetros estarão associados agora à $T(X)$ e não mais a X.

Modelos polinomiais (James *et al.*, 2013) também podem ser utilizados para contornar a não-linearidade. Dado um único preditor X, um modelo polinomial pode ser especificado como

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \cdots + \beta_p X_t^p + \epsilon_t, \quad t = 1, \ldots, n.$$

Essa classe de modelos é bem flexível e permite ajustar associações complexas entre as variáveis X e Y, sendo uma boa alternativa para predição, mas pouco utilizados para inferência devido à falta de interpretação.

Mais detalhes sobre linearidade e outras alternativas podem ser encontradas em Hastie *et al.* (2008) e James *et al.* (2013).

3.1.6 Contornado a suposição de aditividade

Pela suposição de aditividade, os termos do modelo (3.3) são sempre somados, permitindo que cada coeficiente possa ser interpretado independentemente dos demais se os mantivermos fixados.

Na prática, o efeito de uma variável explicativa X_1 em Y pode depender do nível de um outro preditor X_2. O efeito da poluição do ar (X_1) em crises respiratórias (Y), por exemplo, é muito mais acentuado em certas condições climáticas, como dias de baixa umidade (X_2). Essa relação entre X_1 e X_2 na variabilidade de Y é chamada de *interação*.

Gráficos de perfis (Singer *et al.*, 2012) podem ser utilizados para identificar interação entre variáveis. Esses gráficos exigem que pelo menos um dos preditores seja categórico. Se ambas variáveis forem quantitativas, uma delas pode ser categorizada para a construção dos gráficos de perfis.

A interação de duas variáveis pode ser contemplada pelo modelo de regressão linear acrescentando-se termos da forma $X_1 \times X_2$. Interações de três ou mais variáveis também podem ser incluídas, mas dificilmente tem interpretação prática.

Termos de interação bastante utilizados em estudos de poluição do ar são aqueles entre as variáveis meteorológicas. Em geral, além de controlarmos o efeito marginal da temperatura, umidade, precipitação, radiação, vento etc., precisamos também incluir o efeito conjunto dessas variáveis.

3.1.7 Avaliando a qualidade do ajuste

Como discutido na introdução deste capítulo, nossos modelos sempre estarão sujeitos a erros. Assim, além de verificarmos se o modelo escolhido viola as suposições pré-estabelecidas, também precisamos avaliar a magnitude do erro que estamos cometendo ao utilizarmos suas estimativas para descrever o evento sob estudo. Para modelos de regressão linear, isso pode ser feito a partir da raiz do erro quadrático médio (RMSE[5]) e do coeficiente de determinação (R^2).

A raiz do erro quadrático médio é uma estimativa do desvio-padrão de ϵ, uma medida do quanto, em média, a resposta Y se desvia da verdadeira reta de regressão[6]. Valores baixos de RMSE significam que $\hat{Y}_t \approx Y_t$, para $t = 1, \ldots, n$, sugerindo que o modelo está bem ajustado aos dados. Como essa medida depende da magnitude da variável resposta, não existem pontos de corte para definir o que é um RMSE pequeno.

[5]Sigla para o termo em inglês *root mean square error*. Utilizaremos aqui a sigla em inglês porque ela é bastante comum na literatura e nos programas estatísticos.

[6]No caso do modelo de regressão linear simples, por exemplo, a verdadeira reta de regressão é dada por $Y = \tilde{\beta}_0 + \tilde{\beta}_1 X$, em que $\tilde{\beta}_0$ e $\tilde{\beta}_1$ representam os verdadeiros valores de β_0 e β_1. Na prática, $\tilde{\beta}_0$ e $\tilde{\beta}_1$ são desconhecidos e substituídos por valores estimados, como apresentado em (3.7).

O coeficiente de determinação é uma medida da proporção da variância de Y explicada pelos preditores incluídos no modelo. Esse coeficiente varia entre 0 e 1 e, ao contrário do RMSE, não depende da escala de Y. Valores próximos de 1 apontam que uma porção considerável da variabilidade está sendo explicada, indicando que o modelo se ajusta bem aos dados. Na prática, valores de R^2 maiores que 0.7 são considerados altos.

Valores altos de RMSE ou baixos de R^2 sugerem problemas com o modelo. Não-linearidade e omissão de preditores importantes são os mais comuns. No primeiro caso, a principal estratégia é transformar os preditores cuja associação com Y suspeitamos ser não-linear, assim como discutido na Seção 3.1.5. A solução para o segundo caso é obter mais informação sobre o fenômeno sob analise e incluir novos preditores ao modelo. Essa é uma tarefa complicada, pois dificilmente temos acesso a novas variáveis explicativas, e geralmente demonstra uma falha no delineamento do estudo.

Ao se avaliar o RMSE e R^2, um cuidado muito importante deve ser tomado. Acrescentar mais preditores ao modelo sempre irá diminuir o RMSE e aumentar o R^2, o que torna a estratégia de escolher o modelo com menor RMSE ou menor R^2 problemática. O excesso de parâmetros pode gerar *sobreajuste* (ou *overfitting*, em inglês), que acontece quando o modelo passa a explicar padrões que não são generalizáveis para a população. Um modelo sobreajustado captura a variação gerada pelos erros aleatórios, que, por construção, não pode ser explicada pelos preditores. Sendo assim, o modelo será ótimo para representar a amostra, mas, em geral, péssimo para ser estendido para um contexto mais amplo.

Uma maneira de evitar esse problema no modelo de regressão linear é utilizar versões do RMSE e do R^2 penalizadas pelo número de parâmetros, conhecidas como RMSE ajustado e R^2 ajustado. Os valores dessas medidas diminuem quando acrescentamos variáveis que não colaboram muito para explicar a variabilidade de Y, o que nos permite controlar o balanço (*trade off*) existente entre um modelo mal ajustado e um modelo sobreajustado. Discutiremos o conceito de sobreajuste com mais detalhes no Capítulo 4.

Na prática, o coeficiente R^2 é muito mais utilizado que o RMSE para a avaliação do ajuste de modelos de regressão linear. Com objetivo de explicar a variabilidade da concentração de ozônio na cidade de São Paulo, Salvo e Geiger (2014), por exemplo, ajustaram sete modelos lineares com diferentes preditores para controlar os efeitos meteorológicos e de tráfego e escolheram aquele com maior R^2 como o modelo final.

Em alguns casos, a complexidade do fenômeno sob estudo demandará modelos mais flexíveis que o modelo de regressão linear. A seguir, discutiremos os modelos lineares generalizados, uma ampla classe de modelos que permite a utilização de distribuições interessantes para o ajuste de diversos casos práticos, e os modelos aditivos generalizados, que relaxa a suposição de linearidade entre a variável resposta e os preditores.

3.2 Modelos lineares generalizados

É muito comum na modelagem estatística assumirmos um modelo probabilístico para os dados. O que de fato estamos fazendo é supor que as observações no mundo real estão distribuídas conforme uma distribuição de probabilidades, cujos parâmetros podem ser relacionados com os coeficientes do modelo e estimados, por exemplo, por máxima verossimilhança (Casella e Berger, 2001).

Para o modelo de regressão linear discutido na última seção, podemos utilizar o método de

mínimos quadrados para estimação e, para grandes amostras, existem resultados assintóticos que garantem as propriedades necessárias para a construção de intervalos de confiança e testes de hipóteses para as estimativas. Quando trabalhamos com amostras pequenas, não podemos garantir a validade dos resultados assintóticos, e então precisamos supor que a variável resposta é normalmente distribuída para a construção dos intervalos e testes. Embora muito utilizada na prática, a distribuição Normal pode ser restritiva na prática, pois ela assume que as observações variam na reta real (valores positivos e negativos) e são simetricamente distribuídas em torno da média.

A concentração de poluentes é uma medida positiva, em geral assimétrica e heteroscedástica. Quando estamos trabalhando com dados epidemiológicos, o número de casos de doenças ou mortalidade é uma medida de contagem, isto é, assume apenas valores não-negativos inteiros. Se queremos aplicar modelos que fazem suposições sobre a distribuição de probabilidade das observações, é importante que possamos escolher distribuições compatíveis com a natureza dos dados. Nesses casos, as distribuições Gama e Poisson seriam, respectivamente, boas alternativas para a modelagem de concentração de poluentes e dados epidemiológicos de contagem.

Nesse sentido, os modelos lineares generalizados, introduzidos por Nelder e Wedderburn (1972), são uma generalização do modelo de regressão linear que permitem a utilização de distribuições para dados assimétricos (Gama, Normal inversa, Log-normal), dados de contagem (Poisson, Binomial negativa), dados binários (Binomial), entre outros. Nas próximas seções, discutiremos como utilizar essa classe de modelos para o ajuste de dados de poluição do ar.

3.2.1 Especificação do modelo

Sejam Y_t e $\mathbf{X}_t$ definidos como na Seção 3.1.1. O modelo linear generalizado pode ser definido como

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

$$g(\mu_t) = \alpha + \beta_1 X_{1t} + ... + \beta_p X_{pt}, \quad t = 1, \ldots, n, \tag{3.9}$$

sendo[7] $\mathcal{D}$ uma distribuição pertencente à família exponencial[8], $g(\cdot)$ uma função de ligação, μ_t um parâmetro de posição e ϕ um parâmetro de precisão[9].

Os parâmetros deste modelo podem ser estimados por máxima verossimilhança. Os cálculos envolvem o uso de procedimentos iterativos, como Newton-Raphson e escore de Fisher (Dobson, 1990). Distribuições que têm um parâmetro de precisão permitem a modelagem conjunta de μ e ϕ. Essa abordagem é conhecida como *modelo linear generalizado duplo* e flexibiliza a suposição de homoscedasticidade feita em (3.9). Mais informações sobre esses modelos podem ser encontradas em Paula (2013).

A especificação dos termos de tendência e sazonalidade para modelos lineares generalizados pode ser feita da mesma forma que no modelo linear (ver Seção 3.1.2). A utilização de resíduos

[7] A notação $Y_t|\mathbf{X}_t \overset{ind}{\sim} \mathcal{D}(\mu_t, \phi)$ significa que, conhecido os valores dos preditores $\mathbf{X_t}$, as variáveis $Y_1, \ldots, Y_n$ são independentes e seguem a distribuição $\mathcal{D}$, governada pelos parâmetros μ_t e ϕ.

[8] A família exponencial corresponde a uma classe de distribuições de probabilidade que, sob certas condições de regularidade, apresentam algumas características em comum. Essas características permitem que o mesmo *framework* de estimação possa ser utilizado para qualquer uma das distribuições dentro dessa família. Para mais informações, consulte Paula (2013).

[9] Se ϕ e um parâmetro de precisão, ϕ^{-1} é um parâmetro de dispersão. Algumas distribuições não têm um parâmetro de precisão. Nas distribuições Binomial e Poisson, por exemplo, $\phi = 1$ e a precisão é uma função da média μ.

para avaliar a qualidade do ajuste também pode ser conduzida de forma análoga à apresentada nas seções anteriores. Os resíduos mais utilizados em modelos lineares generalizados são definidos a partir da *função desvio*. Uma técnica muito utilizada é a construção de gráficos envelope para investigar a adequação da distribuição escolhida para os dados. Para mais informações sobre a análise de resíduos de modelos lineares generalizados, consulte Paula (2013).

Os modelos com distribuição Gama, Normal inversa e Log normal são boas alternativas para ajustar dados positivos assimétricos, sendo, em geral, mais adequados para concentrações de poluentes do que a distribuição Normal. Discutiremos os dois primeiros na Seção 3.2.2.

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. Discutiremos esses modelos com mais detalhes na Seção 3.2.3.

3.2.2 Modelos para dados positivos assimétricos

A distribuição Gama costuma ser a principal alternativa para o ajuste de dados positivos assimétricos. Se $Y \sim \text{Gama}(\mu, \phi)$, sendo $\mu > 0$ a média de Y, $\phi > 0$ um parâmetro de precisão, a sua função densidade de probabilidade está representada na Figura 3.8 para $\mu = 1$ e diversos valores de ϕ. Podemos observar que, à medida que ϕ aumenta, a distribuição Gama se torna mais simétrica em torno da média. Conforme ϕ tende para infinito, Y se aproxima da distribuição Normal de média μ e variância $\mu^2\phi^{-1}$, o que torna a distribuição Gama atrativa para a modelagem tanto de observações assimétricas quanto de observações simétricas cuja dispersão varia em função da média ao quadrado.

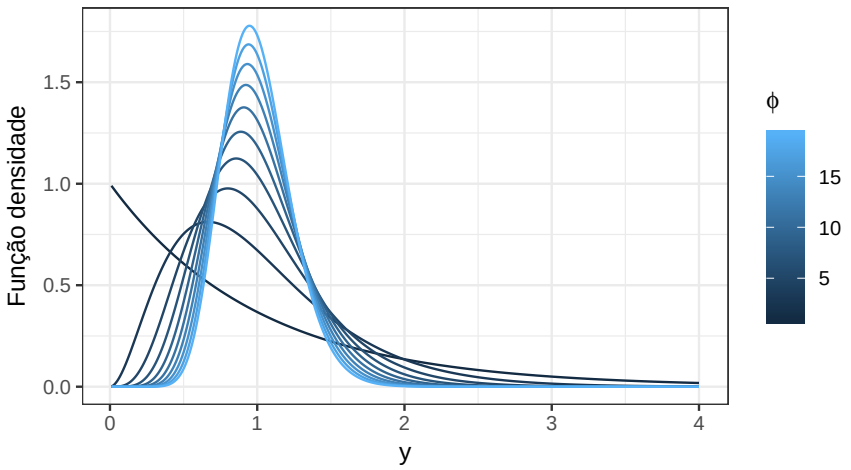

Figura 3.8: *Função densidade da distribuição Gama com $\mu = 1$ e diversos valores de ϕ. Conforme ϕ aumenta, a distribuição se torna menos assimétrica, centralizando-se ao redor da média.*

Uma alternativa para a distribuição Gama é a Normal inversa. Considere agora $Y \sim \text{NI}(\mu, \phi)$, novamente sendo $\mu > 0$ a média de Y e $\phi > 0$ um parâmetro de precisão. Podemos ver pela Figura 3.9 que, para $\mu = 1$, a simetria da distribuição diminui conforme ϕ aumenta. Mais precisamente, Y

se aproxima de uma distribuição Normal com média μ e variância $\mu^3\phi^{-1}$. A Normal inversa é apropriada para modelar tanto observações assimétricas quanto observações simétricas cuja dispersão varia em função da média ao cubo.

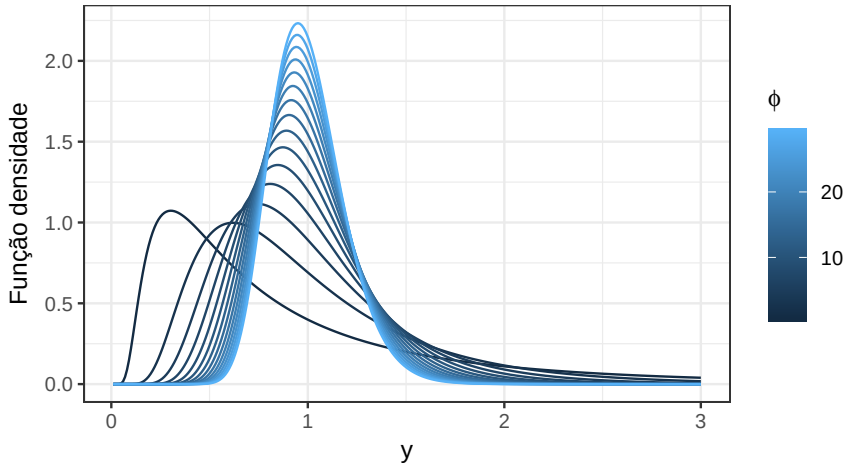

Figura 3.9: *Função densidade da distribuição Normal inversa com $\mu = 1$ e diversos valores de ϕ. Conforme ϕ aumenta, a distribuição se torna menos assimétrica, centralizando-se ao redor da média.*

As funções de ligação mais utilizadas em ambos os modelos são a identidade ($g(\mu) = \mu$), a logarítmica ($g(\mu) = \log(\mu)$) e a recíproca ($g(\mu) = 1/\mu$). Gráficos de resíduos podem ser feitos para avaliar a adequabilidade da distribuição e da função de ligação escolhidas. Para mais informações sobre análise de diagnóstico para modelos lineares generalizados, consultar Williams (1987) e Paula (2013) .

No R, os modelos Gama e Normal inversa podem ser ajustados com a função `glm()` do pacote `stats`, utilizando os argumentos `family = Gamma` e `family = inverse.gaussian`, respectivamente. No pacote `caret`, modelos lineares generalizados podem ser ajustados utilizando a função `train()` com `method="glm"`.

Outras distribuições da família exponencial também podem ser utilizadas para a análise de dados positivos assimétricos, como a Weibull, a Pareto e a Log-Normal (Wood, 2006). Fora do contexto de modelos lineares generalizados, a distribuição de Birnbaum-Saunders generalizada (GBS) é outra alternativa para o ajuste de dados positivos assimétricos. Leiva *et al.* (2008), por exemplo, utilizaram o modelo GBS para ajustar concentrações horárias de dióxido de enxofre em Santiago, no Chile, mostrando que essa distribuição se ajustava melhor aos dados do que a Log-Normal. Para mais informações sobre a distribuição de Birnbaum-Saunders, consulte Barros *et al.* (2009) e Leiva (2015).

3.2.3 Modelos para dados de contagem

Em algumas situações, o objetivo do estudo de poluição do ar não está em descrever as séries de poluentes, mas sim utilizá-las para explicar eventos epidemiológicos, como, por exemplo, a morbidade ou mortalidade causada por doenças respiratórias. A variável resposta nesses estudos é, em geral, uma contagem, isto é, assume valores inteiros positivos que representam o número de casos da doença ou de mortes em cada instante observado.

Schwartz e Dockery (1992), por exemplo, utilizaram o modelo de Poisson para avaliar a relação entre a concentração de material particulado e o número de mortes no dia seguinte, sugerindo

uma associação positiva entre as variáveis. Conceição *et al.* (2001b) também utilizando o modelo Poisson, estudaram a relação entre a concentração de alguns poluentes e marcadores de mortalidade em idosos na cidade de São Paulo, controlando por variáveis meteorológicas. Os autores observaram uma associação positiva entre mortalidade e níveis de CO, SO_2 e, em menor escala, PM10. Já Saldiva *et al.* (1995) discutiram a utilização de um modelo Poisson para modelar a associação entre concentração de diversos poluentes e a mortalidade em idosos, mas optaram pelo ajuste de um modelo gaussiano, justificando que a aproximação pela distribuição Normal era válida pois a média diária de mortes era suficiente alta (62 eventos por dia).

Se a variável resposta Y, segue uma Poisson com parâmetro λ, simbolicamente $Y \sim \text{Poisson}(\lambda)$, o modelo assume que o evento sob estudo ocorre com taxa λ dentro de um intervalo de tempo fixado[10]. Essa taxa representa o valor médio[11] de casos observados no intervalo e, na prática, queremos explicá-la a partir de séries de poluentes, controlando por variáveis climáticas. Dessa forma, para o modelo Poisson, temos $\mu_t = \lambda_t$ em (3.9). A função de ligação mais utilizada nesse contexto é a logarítmica.

Na distribuição Poisson, a média é igual à variância, isto é, $E(Y) = VAR(Y) = \lambda$. Isso gera uma restrição importante no modelo Poisson, deixando-o inadequado para o ajuste de dados com sobredispersão, observações com a variância maior do que a média[12]. Uma alternativa nesse caso é a utilização de modelos com resposta Binomial Negativa.

Se $Y \sim \text{BN}(\mu, \phi)$, temos que $E(Y) = \mu$ e $VAR(Y) = \mu + \mu^2/\phi$, com $\mu \geq 0$ e $\phi > 0$, o que faz a distribuição Binomial Negativa adequada para dados com variância maior do que a média.

Modelos de contagem geralmente são utilizados para obter uma estimativa do risco de mortalidade associado a cada poluente, isto é, qual a variação esperada na taxa de mortalidade se aumentássemos (ou diminuíssemos) a concentração de um poluente em m unidades. Essa quantidade, conhecida como dose-resposta, concentração-resposta ou exposição-resposta, é muito importante para a implementação de medidas para a redução da poluição do ar pois quantifica de forma objetiva o efeito dos poluentes na saúde pública. Outra métrica bastante utilizada são as *funções de impacto na saúde*, discutidas brevemente na Seção 3.4.5.

No R, o modelo Poisson pode ser ajustado com a função `glm()` do pacote `stats`, utilizando o argumento `family = poisson`, enquanto o modelo com resposta Binomial Negativa utilizando a função `glm.nb()` do pacote `MASS`. Utilizando o pacote `caret`, esses modelos podem ser ajustados utilizando a função `train()` com `method="glm"`.

3.3 Modelos aditivos generalizados

Os modelos lineares têm um papel muito importante na análise de dados, provendo técnicas de inferência e predição computacionalmente simples e de fácil interpretação. Contudo, em problemas reais, a relação entre a variável resposta e os preditores pode não ser linear, tornando os modelos lineares muito restritivos. No estudo de poluentes atmosféricos, por exemplo, o aspecto temporal dos dados gera efeitos sazonais cuja relação com a variável reposta é muito melhor representada por curvas senoidais do que por retas.

[10] Esse intervalo de tempo se refere à frequência com que os dados são coletados, isto é, se as séries são diárias, semanais, mensais, anuais etc.

[11] A distribuição de Poisson atribui maiores probabilidades aos valores próximos à média λ.

[12] Para o modelo Poisson, $\phi = 1$.

Os modelos aditivos generalizados (Hastie *et al.*, 2008) são um método integrado, automático e flexível para identificar e caracterizar relações não-lineares entre as variáveis. Ao contrário das estratégias discutidas na Seção 3.1.5, como transformação da variável e regressão polinomial, os modelos aditivos generalizados não são lineares nos parâmetros, permitindo a estimação de funções não-lineares entre os preditores e a resposta. Belusic *et al.* (2015), por exemplo, utilizaram essa classe de modelos para avaliar quais as variáveis mais importantes para descrever a série de diversos poluentes em Zagreb, Croácia. O modelo ajustado apontou que as variáveis meteorológicas explicavam a maior proporção da variabilidade dos poluentes.

Neste seção, vamos discutir o ajuste e interpretação dos modelos aditivos generalizados no contexto de estudos de poluição do ar.

3.3.1 Especificação do modelo

O modelo aditivo generalizado é 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 (Seção 3.1.6). Como nas seções anteriores, sejam Y_t e $\mathbf{X}_t$ séries temporais que representam, respectivamente, a variável resposta e as variáveis preditoras, com $t = 1, ..., n$. O modelo aditivo generalizado pode ser escrito como

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

$$g(\mu_t) = \beta_0 + f_1(X_{1t}) + \cdots + f_p(X_{pt}), \tag{3.10}$$

sendo $\mathcal{D}$ uma distribuição pertencente à família exponencial e f_i, $i = 1, ..., p$, funções possivelmente não-lineares. 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 sobre como as funções $f_1, ..., f_p$ devem ser representadas, incluindo o uso de *splines* naturais, *splines* suavizados e regressão local (Hastie e Tibshirani, 1990). Outro ponto importante diz respeito à suavidade dessas funções, controlada por *parâmetros de alisamento*, que devem ser determinados a priori[13]. Curvas muito suaves podem ser muito restritivas, enquanto curvas muito *rugosas* podem sobreajustar os dados (*overfitting*). Discutiremos esse tema com mais detalhes na Seção 3.3.2.

O procedimento de estimação no contexto de modelos aditivos generalizados depende da forma escolhida para as funções $f_1, ..., f_p$. A utilização de *splines* naturais, por exemplo, permite a aplicação direta de mínimos quadrados, graças à sua construção a partir de *funções base* (ver Seção 3.3.2). Já para *splines* penalizados, o processo de estimação envolve algoritmos um pouco mais complexos, como *backfitting* (Breiman e Friedman, 1985). Para mais informações sobre a estimação dos parâmetros dos modelos lineares generalizados, consulte Hastie e Tibshirani (1990) e Hastie *et al.* (2008).

A seguir, introduziremos os conceitos de *splines* e regressão local, e apresentaremos os principais aspectos em torno do ajuste dessas técnicas.

[13] Uma maneira de determinar valores para esses parâmetros é utilizar validação cruzada, que será discutida no Capítulo 4.

3.3.2 Splines e regressão local

Para introduzir o conceito de *splines* e regressão local, vamos considerar novamente o modelo mais simples, com apenas uma variável explicativa

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \quad t = 1, ..., t. \tag{3.11}$$

Uma das principais ideias por trás dos modelos aditivos generalizados está na utilização de *funções bases*. Essa abordagem considera uma família de transformações $b_1(X), b_2(X), ..., b_k(X)$, fixadas e conhecidas, no lugar de X em (3.11). Assim, o modelo (3.11) passa a ser

$$Y_t = \beta_0 + \beta_1 b_1(X_t) + \beta_2 b_2(X_t) + \cdots + \beta_k b_k(X_t) + \epsilon_t, \quad t = 1, ..., t, \tag{3.12}$$

que pode assumir diversas classes de associações não-lineares entre X e Y. Note que o modelo polinomial apresentado na Seção 3.1.5 é um caso particular de (3.12), com $b_j(X_t) = X_t^j$, $j = 1, \ldots, k$.

Como uma tentativa para aumentar a flexibilidade da curva ajustada, podemos segmentar X e ajustar diferentes polinômios de grau d em cada um dos intervalos[14]. Cada ponto de segmentação é chamado de *nó*, e uma segmentação com k nós gera $k+1$ polinômios. Na Figura 3.10 apresentamos um exemplo com polinômios de terceiro grau e 4 nós. Nesse exemplo, a expressão (3.12) tem a forma

$$Y_t = \begin{cases} \beta_{01} + \beta_{11} X_t + \beta_{21} X_t^2 + \beta_{31} X_t^3 + \epsilon_t, & \text{se } X_t \leq -0.5, \\ \beta_{02} + \beta_{12} X_t + \beta_{22} X_t^2 + \beta_{32} X_t^3 + \epsilon_t, & \text{se } -0.5 < X_t \leq 0, \\ \beta_{02} + \beta_{13} X_t + \beta_{23} X_t^2 + \beta_{33} X_t^3 + \epsilon_t, & \text{se } 0 < X_t \leq 0.5, \\ \beta_{02} + \beta_{14} X_t + \beta_{24} X_t^2 + \beta_{34} X_t^3 + \epsilon_t, & \text{se } 0.5 < X_t \leq 1, \\ \beta_{05} + \beta_{15} X_t + \beta_{25} X_t^2 + \beta_{35} X_t^3 + \epsilon_t, & \text{se } X_t > 1, \end{cases}$$

sendo que as funções base $b_1(X), b_2(X), ..., b_k(X)$ nesse caso são construídas com a ajuda de funções indicadoras. Esse modelo é conhecido como modelo polinomial cúbico segmentado.

Repare que a curva formada pela junção de cada um dos polinômios na Figura 3.10 não é contínua, isto é, apresenta saltos nos nós. Essa característica não é desejável para um modelo ajustado, já que essas descontinuidades não são interpretáveis. Para contornar esse problema, vamos definir um *spline* de grau d como um polinômio segmentado de grau d com as $d-1$ primeiras derivadas contínuas em cada nó. Essa restrição garante a continuidade e suavidade (ausência de vértices) da curva obtida.

Utilizando a representação por bases (3.12), um *spline* cúbico com k nós pode ser modelado por

$$Y_t = \beta_0 + \beta_1 b_1(X_t) + \beta_2 b_2(X_t) + ... + \beta_{k+3} b_{k+3}(X_t) + \epsilon_t, \quad t = 1, ..., t,$$

para uma escolha apropriada de funções $b_1(X), b_2(X), ..., b_{k+3}(X)$. Usualmente, essas funções envolvem três termos polinomiais — X, X^2 e X^3, mais precisamente — e k termos $h(X, c_1), \ldots, h(X, c_k)$ da forma

$$h(X, c_j) = (x - c_j)_+^3 = \begin{cases} (x - c_j)^3, & \text{se } x < c_j, \\ 0, & \text{em caso contrário,} \end{cases}$$

[14] Em contrapartida ao modelo polinomial, que ajusta um único polinômio sobre todo o intervalo de variação de X.

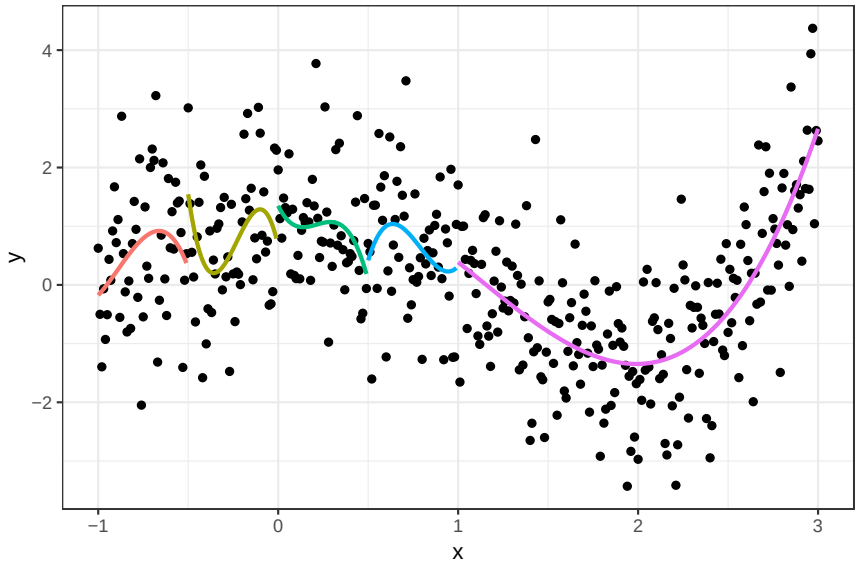

Figura 3.10: *Polinômios de terceiro grau ajustados em cada segmentação da variável X. Os nós são os pontos $x = -0.5$, $x = 0$, $x = 0.5$ e $x = 1$.*

sendo $c_1, \ldots, c_k$ os k nós. Assim, incluindo o termo β_0, o ajuste de um *spline* cúbico com k nós envolve a estimação de $k+4$ parâmetros e, portanto, utiliza $k+4$ graus de liberdade. Mais detalhes sobre a construção dessas restrições podem ser encontrados em Hastie *et al.* (2008) e James *et al.* (2013).

Além das restrições sobre as derivadas, podemos adicionar *restrições de fronteira*, exigindo que a função seja linear na região de X abaixo do menor nó e acima do maior nó. Essas restrições diminuem a variância nos extremos do preditor, produzindo estimativas mais estáveis. Um *spline* cúbico com restrições de fronteira é chamado de *spline natural*.

No ajuste de *splines* cúbicos ou naturais, o número de nós determina o grau de suavidade da curva, e a sua escolha pode ser feita por *validação cruzada* (James *et al.*, 2013). De uma forma geral, a maior parte dos nós é posicionada nas regiões do preditor com mais informação, isto é, mais observações. Por pragmatismo, para modelos com mais de uma variável explicativa, costuma-se adotar o mesmo número de nós para todos os preditores.

Os *splines suavizados* constituem uma classe de funções suavizadoras que não utilizam a abordagem por funções bases. De maneira resumida, um *spline* suavizado é uma função f que minimiza a seguinte expressão

$$\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda \int f''(u)^2 du. \tag{3.13}$$

O primeiro termo dessa expressão garante que f se ajustará bem aos dados, enquanto o segundo penaliza a sua variabilidade, isto é, controla o quanto f será suave. A suavidade é regulada pelo parâmetro λ, sendo que f se torna mais suave conforme λ cresce. A escolha desse parâmetro é geralmente feita por validação cruzada.

Uma outra forma para ajustar funções não-lineares entre X e Y é a regressão local. Essencialmente, essa técnica consiste em ajustar modelos de regressão simples em regiões de pontos ao redor de cada observação x_0 do preditor X. Essas regiões são formadas pelos k pontos mais próximos de x_0, sendo que o parâmetro $s = k/n$, determina o quão suave ou rugosa será a curva ajustada.

O ajuste é feito por mínimos quadrados ponderados, e os pesos são inversamente proporcionais à distância do ponto em relação a x_0. Assim, os pontos na vizinhança de x_0 mais afastados recebem peso menor.

No R, modelos lineares generalizados podem ser ajustados utilizando-se a função `gam()` do pacote `mgcv`. Essa função permite a utilização de *splines* como função suavizadora. Para a utilização de regressão local, é necessário usar a função `gam()` do pacote `gam`. Também é possível utilizar o pacote `caret`, a partir da função `train()` e `method = "gam"`.

Para mais informações sobre *splines*, regressão local e modelos lineares aditivos em geral, consultar Hastie *et al.* (2008) e James *et al.* (2013).

3.4 Modelos de previsão

Às vezes, queremos explicar a série Y apenas por seus valores defasados no tempo (autocorrelação) ou pelos valores defasados dos preditores $X_1, \ldots, X_p$ (correlação cruzada). Como os modelos de regressão apresentados até aqui permitem que Y seja influenciada apenas por valores contemporâneos das variáveis explicativas, eles podem ser insuficientes para explicar toda as relações temporais presente em uma série.

Nesta seção, vamos introduzir a classe de modelos ARIMA (Box e Jenkins, 1970), que contemplam a correlação gerada por relações lineares entre observações defasadas no tempo da própria variável. A associação entre a variável resposta e valores defasados no tempo de covariáveis não será tratada aqui, mas são contemplados por modelos de regressão defasada (*lagged regression*), discutidos nas seções 4.10 e 5.6 de Shumway e Stoffer (2006).

3.4.1 Modelos autorregressivos (AR)

Modelos autorregressivos se baseiam na ideia de que Y_t pode ser explicada como uma função de p valores passados $Y_{t-1}, \ldots, Y_{t-p}$, sendo p o número de passos no passado necessários para prever o valor no instante t. Se Y_t é uma série estacionária, o modelo autorregressivo de ordem P, abreviado como AR(p), é definido como

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + w_t, \tag{3.14}$$

sendo $\phi_1, \ldots, \phi_p$ constantes com $\phi_p \neq 0$ e $w_t \sim N(0, \sigma_w^2)$, $t \geq 0$. Sem perda de generalidade, assume-se que a média de Y_t é zero[15].

Os modelos AR(p) são muito utilizados em Economia, onde é natural pensar o valor de alguma variável no instante t como função de seus valores defasados, e em algumas áreas da Física e Geofísica, onde os estimadores auto-regressivos são utilizados para estimar o espectro de certos processos.

3.4.2 Modelos autorregressivos e de médias móveis (ARMA)

Uma alternativa para o modelo AR(p) é o modelo de médias móveis de ordem q. Esse modelo assume que Y_t é gerado a partir de uma combinação linear dos erros $w_t, w_{t-1}, \ldots, w_{t-q}$. Formalmente,

[15] Se a média de Y_t é $\mu \neq 0$, então o modelo é definido para $Y_t - \mu$, o que equivale a acrescentar um intercepto $\alpha = \mu(1 - \phi_1 - \ldots \phi_p)$ ao modelo (3.14).

o modelo de médias móveis de ordem q, MA(q), é definido como

$$Y_t = w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \tag{3.15}$$

sendo $\theta_1, \ldots, \theta_q$ constantes com $\theta_q \neq 0$ e $w_t \sim N(0, \sigma_w^2)$, $t \geq 0$.

Ao contrário dos modelos auto-regressivos, representar um processo por um modelo de médias móveis puro parece não ser intuitivo.

A utilização de modelos com termos auto-regressivos e de médias móveis pode ser uma boa alternativa para muitas séries encontradas na prática, pois eles normalmente requerem um menor número de parâmetros para explicar a autocorrelação da série (Morettin e Toloi, 2004). Nesse sentido, dizemos que uma série temporal Y_t é ARMA(p, q) se ela é estacionária e se

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q Y_{t-q}, \tag{3.16}$$

com $\phi_p \neq 0$, $\theta_q \neq 0$ e $\sigma_w^2 > 0$.

Repare que os modelos AR(p) e MA(q) são casos particulares do ARMA(p, q), com $q = 0$ e $p = 0$ respectivamente.

A estimação dos parâmetros $(\phi_1, \ldots, \phi_p)$ e $(\theta_1, \ldots, \theta_q)$ pode ser feita por máxima verossimilhança ou pelo método de mínimos quadrados. Para mais informações, consulte a Seção 3.6 de Shumway e Stoffer (2006).

As três classes de modelos apresentadas até aqui consideram que a série Y_t é estacionária, o que normalmente não acontece na prática. Para flexibilizar essa restrição, apresentaremos a seguir os modelos ARIMA(p, d, q), uma extensão da classe ARMA que considera a diferenciação de grau d da série para eliminar a não-estacionariedade.

3.4.3 Modelos autorregressivos integrados e de médias móveis (ARIMA)

Vimos na Seção 3.1.2 que séries não-estacionárias podem ser diferenciadas para se alcançar a estacionariedade. De maneira geral, essa estratégia é válida para séries que não apresentam *comportamento explosivo* ou, em outros termos, que apresentam alguma homogeneidade em seu comportamento não-estacionário. Morettin e Toloi (2004) enquadram séries dessa natureza, chamadas de *séries não-estacionárias homogêneas*, em dois grupos:

- séries que oscilam ao redor de um nível médio durante algum tempo e depois saltam para outro nível temporário; e
- séries que oscilam em uma direção por algum tempo e depois mudam para outra direção temporária.

O primeiro tipo requer apenas uma diferença para torná-las estacionária, enquanto o segundo requer duas. Dessa forma, a série não-estacionária homogênea Y_t é dita ser ARIMA(p, d, q) se $\Delta^d Y_t$, como definido em (2.1), é ARMA(p, q).

Como discutido na Seção 3.8 de Shumway e Stoffer (2006) e no Capítulo 6 de Morettin e Toloi (2004), precisamos seguir alguns passos essenciais no ajuste de modelos ARIMA:

1. Construir o gráfico da série.

2. Transformar a série, se preciso.
3. Identificar a ordem de dependência do modelo.
4. Estimar os parâmetros.
5. Diagnóstico.
6. Selecionar o melhor modelo.

No primeiro passo, podemos encontrar anomalias, como heteroscedasticidade, a partir da gráfico da série contra o tempo. No passo 2, corrigimos essas anomalias utilizando alguma transformação.

No passo 3, precisamos identificar as ordens p, d e q do modelo. O próprio gráfico da série irá sugerir se alguma diferenciação será necessária. Se alguma diferenciação for realizada, calculamos ΔY_t, $t = 2, \ldots, n$, e checamos no gráfico da série ΔY_t contra o tempo t se outra diferenciação é necessária. Continuamos esse processo, sempre checando os gráficos da série diferenciada contra o tempo[16].

Com o valor de d selecionado, observamos o gráfico da função de autocorrelação amostral e da função de autocorrelação parcial amostral de $\Delta^d Y_t$. Sugestões para os valores de p e q podem ser encontrados segundo os critérios apresentados na Tabela 3.1.

Tabela 3.1: *Critérios para a escolha da ordem de modelos ARIMA.*

	AR(p)	MA(q)	ARMA(p, q)
ACF	Calda longa	Desaparece após o *lag* q	Calda longa
PACF	Desaparece após o *lag* p	Calda longa	Calda longa

A ideia nesse passo é, a partir dos gráficos da função de autocorrelação e autocorrelação parcial, escolher alguns valores para p, d e q e, no passo 4, ajustar os respectivos modelos. Assim, a partir da análise de diagnóstico realizada no passo 5, selecionar o modelo que melhor se ajustou aos dados no passo 6.

A classe ARIMA pode ser generalizada para incluir o ajuste da sazonalidade. Essa nova classe, conhecida como SARIMA, inclui termos autoregressivos e de médias móveis para termos separados por *lags* de tamanho s. Para mais informações, recomendamos a leitura do Capítulo 10 de Morettin e Toloi (2004) e da Seção 3.9 de Shumway e Stoffer (2006).

3.4.4 Outros modelos de previsão

Modelos GARCH

Os modelos para séries temporais apresentados até aqui são utilizados para modelar a média condicional de um processo quando a variância condicional (volatilidade) é constante. Em muitos problemas, contudo, a suposição de homoscedasticidade pode não ser verdadeira.

Os modelos autoregressivos com heteroscedasticidade condicional (ARCH), propostos por Engle (1982), foram desenvolvidos para contemplar mudanças da volatilidade da série. Se $\epsilon_t \sim N(0, 1)$, para $t = 1, \ldots, n$, o modelo ARCH(q) é definido por

[16] Cuidado para não introduzir dependência onde não existe. Por exemplo, $Y_t = w_t$ é serialmente não-correlacionada, mas $\Delta Y_t = w_t - w_{t-1}$ é MA(1).

$$Y_t = f(\mathbf{X}, \mathbf{Y}) + \sigma_t \epsilon_t$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \ldots \epsilon_q \epsilon_{t-q}^2, \tag{3.17}$$

com $\alpha_0 > 0$ e $\alpha_i \geq 0$, $i > 0$, sendo $f(\mathbf{X}, \mathbf{Y})$ uma função dos preditores $\{(X_{1i}, \ldots, X_{pi}), i \leq t\}$ e das variáveis defasadas $(Y_1, \ldots, Y_{t-1})$. Repare a primeira expressão de (3.17) permite o ajuste de diversas classes de modelo para a média condicional de Y_t, como modelos de regressão linear, modelos ARIMA e modelos de função de transferência, enquanto a segunda impõe um modelo autorregressivo de ordem p para a volatilidade do processo.

Bollerslev (1986) estendeu a classe ARCH, propondo os GARCH (*generalized* ARCH). Essa nova classe permite o ajuste de um modelo ARMA para a variância do erro (σ^2), modelando a volatilidade da série com menos parâmetros que um modelo ARCH (Morettin e Toloi, 2004). Esse modelo pode ser expresso por

$$Y_t = f(\mathbf{X}, \mathbf{Y}) + \sigma_t \epsilon_t$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \ldots \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_p \sigma_{t-p}^2, \tag{3.18}$$

sendo $f(\mathbf{X}, \mathbf{Y})$ definida como antes.

Por ser um modelo com muitos parâmetros, a especificação do modelo GARCH(p, q), geralmente é dividida em três passos:

1. Estimar o melhor modelo AR(q):

$$Y_t = a_0 + a_1 Y_{t-1} + \cdots a_q Y_{t-q} + \epsilon_t$$

2. Calcular e construir o gráfico das autocorrelações de ϵ^2, dadas por

$$\rho_i = \frac{\sum_{t=i+1}^{T} (\hat{\epsilon}_t^2 - \hat{\sigma}_t^2)(\hat{\epsilon}_{t-1}^2 - \hat{\sigma}_{t-1}^2)}{\sum_{t=1}^{T} (\hat{\epsilon}_t^2 - \hat{\sigma}_t^2)^2},$$

   sendo T o tamanho amostral.

3. Avaliar valores de ρ_i maiores que $1/\sqrt{T}$.

A estimação desses modelos pode ser conduzida da mesma forma que para os modelos ARMA, discutida na Seção 3.6 de Shumway e Stoffer (2006).

Modelos dinâmicos

Estudos de poluição atmosférica envolvem dados cuja coleta é naturalmente suscetível à omissão. A medição de poluentes e de variáveis meteorológicas, por exemplo, envolve equipamentos que estão sujeitos a imprecisões, falhas e precisam ser constantemente regulados. Esses dados geralmente são sustentados pela administração pública, cuja redução de verbas pode descontinuar ou reduzir os planos de coleta.

Às vezes, o próprio delineamento do estudo gera dados faltantes. Na análise feita por Salvo e Geiger (2014), os autores descartaram da amostra os meses frios (julho à setembro), devido à menor formação de ozônio nesse período. Por causa da influência do tráfego no estudo, os feriados e fins de semanas também não foram considerados. Essas exclusões geraram uma série com "buracos", inviabilizando a aplicação de modelos que fazem a suposição de observações equidistantes, como os modelos ARIMA apresentados anteriormente.

Os modelos modelos lineares dinâmicos (ou espaço-estado ou filtros de Kalman), introduzidos por Kalman (1960) e Kalman e Bucy (1961), são uma alternativa nesses casos. Eles são caracterizados por duas suposições principais. A primeira afirma que a verdadeira variável sob estudo, U_t, é um fenômeno não-observável. Neste caso, o que realmente observamos é uma transformação linear desse fenômeno, $A_t U_t$, acrescida de um ruído, v_t. A segunda suposição diz respeito sobre o processo de geração de U_t. Mais precisamente, na sua forma mais básica, temos que U_t é gerado por um processo autoregressivo de primeira ordem.

Dadas essas duas suposições, podemos escrever o modelo de espaço-estado da seguinte maneira

$$\begin{aligned} Y_t &= a_t U_t + v_t \\ U_t &= \phi U_{t-1} + w_t, \end{aligned} \tag{3.19}$$

sendo a_t e ϕ parâmetros do modelo e $w_t \sim N(0, \sigma_w)$. A primeira equação em (3.19) é chamada de *equação de estado*, enquanto a segunda é chamada de *equação de observação*.

A aplicação desses modelos nos permite ajustar a série a partir de suas observações passadas, como nos modelos ARIMA, mas, a cada passo (instante), incorporamos informação de um processo externo, que pode ser tanto a informação de variáveis explicativas quanto de outros processos autorregressivos. Assim, valores omissos no instante t são estimados a partir da informação contida em $1, \ldots, t-1$, sendo uma maneira natural e integrada para lidar com os buracos da série.

Dordonnat *et al.* (2008) apresentam um famoso uso da aplicação de modelos dinâmicos para previsão de consumo de energia elétrica na França. Mais informações sobre o modelo, no contexto de poluição do ar, podem ser encontradas no capítulo 12.3 de Zannetti (1990).

3.4.5 Outros tópicos de modelagem

Nesta seção, apresentaremos alguns tópicos adicionais sobre modelagem de dados de poluição.

Funções de impacto na saúde

Chang *et al.* (2017) utilizaram funções de dose-resposta para avaliar a associação da mortalidade com as concentrações de material particulado próximas a estradas na região central da Carolina do Norte (EUA). Eles concluíram que 72% das mortes prematuras associadas à exposição de PM2.5 aconteciam em um raio de 1km de grandes estradas, onde cerca de 50% da população vivia.

Referências de funções de impacto na saúde: Chang *et al.* (2017).

Capítulo 4

Estratégias de modelagem preditiva

> There are no routine statistical questions,
> only questionable statistical routines.
> — Sir David Cox

Nos últimos anos, uma nova abordagem de análise de dados se tornou muito popular, principalmente pela sua eficiência na resolução de problemas de predição, como detecção de imagens, transcrição de áudio e sistemas de recomendação. A chamada *modelagem preditiva*[1] envolve um conjunto de técnicas que visam gerar a estimativa mais precisa possível para uma quantidade ou fenômeno.

No capítulo anterior, apresentamos diversas classes de modelos úteis para fazer inferência em estudos de poluição do ar. A utilização desses modelos depende de suposições sobre a forma como as variáveis explicativas e as variável resposta estão relacionadas. De uma forma geral, essas suposições são feitas a partir de um modelo probabilístico para a variável resposta Y, cuja parametrização dependerá de alguma função dos preditores[2] $\mathbf{X}$. O modelo de regressão linear (3.3), por exemplo, assume as seguintes hipóteses:

- a média de Y depende das variáveis $\mathbf{X}$ a partir da relação $\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$ (linearidade e aditividade);
- a variância de Y, σ^2, é constante para todas as observações na população.

Essas suposições, embora potencialmente restritivas, permitem que o modelo seja interpretável, isto é, ao estimarmos os coeficientes $\beta_0, \beta_1, \ldots, \beta_p$, podemos avaliar como a variável Y é influenciada por cada preditor $X_1, \ldots, X_p$.

As técnicas e modelos utilizados para modelagem preditiva deixam a interpretabilidade parcialmente de lado para focar na precisão dos modelos ajustados. Além disso, as estratégias dentro dessa abordagem têm uma preocupação especial com o sobreajuste, que representa o quanto o modelo pode ser generalizado para além da amostra. Embora essa abordagem esteja focada no ajuste

[1] Também conhecida como aprendizado estatístico, aprendizagem automática ou aprendizado de máquina *machine learning*.

[2] Muita da literatura sobre modelagem preditiva vem da área da Ciência da Computação. Os computólogos, de uma maneira geral, denominam as variáveis respostas como variáveis de saída ou *outputs* e os preditores como variáveis de entrada ou *inputs*.

de modelos preditivos, muitas das estratégias adotadas por ele também podem ser aplicadas em problemas de inferência.

Neste capítulo, discutiremos com mais detalhes o conceito de sobreajuste (*over-fitting*) e apresentaremos métodos de reamostragem, seleção de variáveis e regularização. Em seguida, introduziremos alguns modelos bastante utilizados dentro desse contexto.

4.1 Sobreajuste e o balanço entre viés e variância

Ao utilizarmos um modelo estatístico para predição, estamos sujeitos a dois tipos de erros: um erro redutível e outro irredutível. No contexto apresentado na introdução do Capítulo 3, dificilmente vamos conseguir uma estimativa perfeita para a função f, e essa imprecisão introduz erro nas predições do modelo. Esse erro é chamado de *redutível*, pois sempre podemos encontrar uma candidata $\hat{f}$ mais próxima da verdadeira f. No entanto, como Y depende também do termo ϵ (3.2), mesmo se pudêssemos estimar f com 100% de precisão, ainda teríamos um erro associado. Por construção, o termo ϵ representa a variação em Y que não pode ser explicada pelos preditores $\mathbf{X}$, e como essa imprecisão não pode ser reduzida, independentemente de qual $\hat{f}$ nós escolhermos, esse erro é chamado de *irredutível*.

O grande desafio na hora de ajustar um modelo aos dados é encontrar uma $\hat{f}$ que minimize o erro redutível, isto é, queremos encontrar um modelo que, utilizando os dados disponíveis, gere as estimativas o mais precisas possível sobre o fenômeno sob estudo. Dentro da modelagem preditiva, essa tarefa é equivalente a minimizar duas quantidades do modelo: o *viés* e a *variância*.

Para entender melhor o que essas quantidades representam, imagine que precisamos ajustar um modelo para os dez pontos apresentados na Figura 4.1 (a). Podemos começar ajustando um modelo de regressão linear simples,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, 10,$$

e calcular a raiz do erro quadrático médio (definido na Seção 3.1.7) para avaliar o quanto a reta ajustada se afasta dos pontos. Uma forma de tentar melhorar o ajuste seria acrescentar um termo quadrático e verificar se o RMSE diminui. Podemos repetir esse procedimento acrescentando termos de graus cada vez maior[3], até encontrarmos o menor RMSE.

Na Tabela 4.1, apresentamos o RMSE obtido para os modelos de regressão polinomial até o nono grau[4]. Observe que, conforme aumentamos a complexidade do modelo (grau do polinômio), o RMSE diminui, até chegar em 0 para o polinômio de grau 9. Se utilizarmos puramente o RMSE como medida da performance do modelo, escolheríamos justamente esse polinômio como modelo final. No entanto, pela Figura 4.1 (b), observamos que esse modelo está claramente mal ajustado aos dados.

Considere agora, nesse mesmo exemplo, que conseguimos uma nova amostra com mais 100 observações geradas pelo mesmo fenômeno que gerou as 10 primeiras. A Figura 4.1 (c) ratifica o quanto o modelo polinomial de grau 9 se ajustou mal aos dados, enquanto os modelos de grau 1 e 2 parecem escolhas mais razoáveis. Podemos observar ainda na Tabela 4.1 que o RMSE do modelo

[3] Esses são os modelos polinomiais apresentados na Seção 3.1.5.

[4] O modelo de regressão linear simples é um modelo polinomial de grau 1.

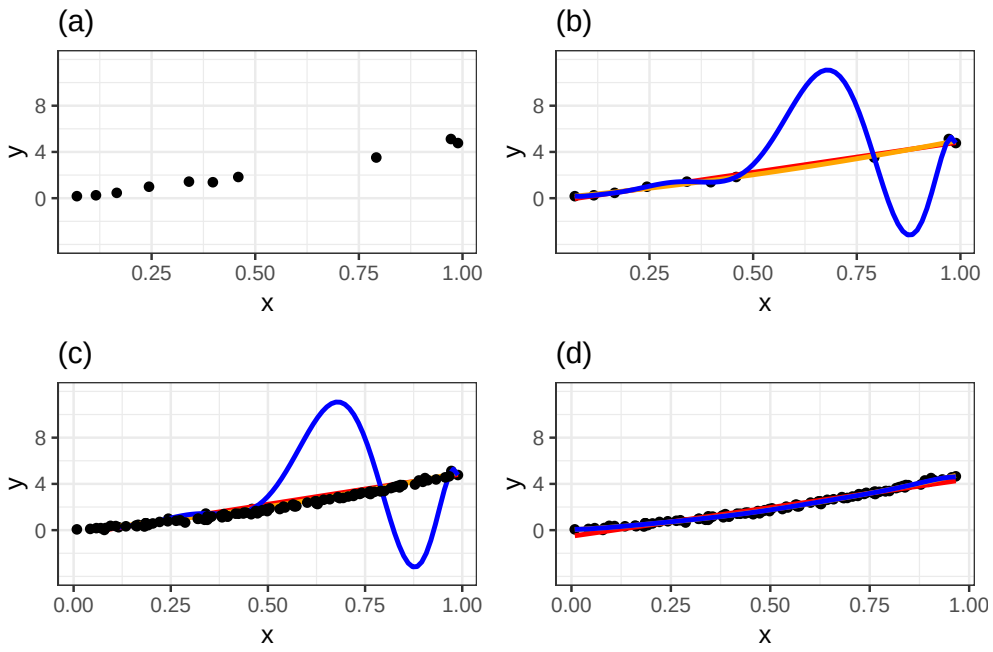

Figura 4.1: *Exemplo do trade-off entre viés e variância. (a) Conjunto de 10 pontos que gostaríamos de ajustar. (b) Modelo de regressão linear simples (vermelho), modelo de regressão polinomial de grau 2 (amarelo) e modelo de regressão polinomial de grau 9 (azul), ajustados aos 10 pontos. (c) Amostra de 100 novas observações plotadas juntas dos modelos polinomiais ajustados nas 10 observações iniciais. (d) Modelos de regressão polinomial de graus 1 (vermelho), 2 (amarelo) e 9 (azul) ajustados aos 100 novos pontos.*

Tabela 4.1: *Raiz do erro quadrático médio (RMSE) para os modelos polinomiais de grau 1 a 9 ajustados com 10 e 110 observações no exemplo da Figura 4.1.*

Grau do polinômio	RMSE (10 obs.)	RMSE (100 obs.)
1	0.204	0.360
2	0.149	0.226
3	0.140	0.199
4	0.140	0.198
5	0.102	0.289
6	0.086	0.360
7	0.063	0.320
8	0.031	1.152
9	0.000	3.904

polinomial de grau 9 calculado nas 100 novas observações[5] é o maior entre todos os candidatos. Por fim, observe na Figura 4.1 (d) como a curva desse modelo muda quando o ajustamos agora usando as 100 novas observações.

[5] Aqui, os modelos não foram reajustados. Foram considerados os modelos ajustados apenas com as 10 primeiras observações

Durante a modelagem, estamos sempre em busca de modelos que se ajustem bem à amostra, mas que também possam ser generalizados para a população. Nesse sentido, chamamos de *viés* o quanto o modelo ajustado está distante dos dados da amostra e de *variância* o quanto o modelo erra ao o utilizarmos para prever novas observações. O viés representa o erro induzido por aproximar um fenômeno real, que pode pode ser extremamente complicado, por um modelo muito mais simples e a *variância* o quanto as estimativas dos parâmetros do modelo mudariam se nós tivéssemos usado uma amostra diferente. Assim, dizemos que modelos mal ajustados apresentam alto viés e modelos que erram muito na predição de novas observações apresentam alta variância.

É muito comum utilizarmos estratégias que se preocupam apenas com a minimização do viés. Essas estratégias geram modelos mais complexos, visando um ajuste cada vez melhor aos dados, sem levar em conta o quanto isso será representativo em um contexto mais geral. No exemplo anterior, isso fica claro com o ajuste de polinômios de grau cada vez maior aos dados. Além disso, o polinômio de grau 9 ilustra, de forma bem simplificada, o conceito de sobreajuste, que ocorre quando o modelo absorve de forma inadequada comportamentos da amostra que não são generalizáveis para a população. Modelos sobreajustados apresentam baixo viés, mas alta variância, não sendo apropriados para representar o fenômeno de interesse. Controlar o balanço entre o viés e a variância é um dos maiores desafios da modelagem preditiva.

Na presença de muitos preditores, não é possível visualizar graficamente o sobreajuste, como mostrado no exemplo. Por isso, na prática, nem sempre é trivial identificar um modelo sobreajustado. Para contornar esse problema, apresentaremos na próxima seção medidas utilizadas para quantificar o viés e a variância de um modelo.

4.2 Estimando a performance do modelo

Na Seção 3.1.7, vimos que o R^2 e a raiz do erro quadrático médio (RMSE) podem ser utilizados para avaliar a qualidade do ajuste de um modelo de regressão linear. Embora o R^2 seja utilizado apenas nessa classe de modelos, o RMSE pode ser calculado em qualquer contexto. Em alguns casos, podemos querer utilizar o erro absoluto médio (MAE), que dá menos peso para erros em valores muito altos da variável resposta. Em outros, pode ser razoável penalizar mais justamente os valores mais altos, e então construímos uma medida que distribui os pesos dessa maneira.

A escolha da métrica de performance vai depender sempre do objetivo do estudo. Independentemente da medida escolhida, ao calculá-la para as próprias observações utilizadas no ajuste, temos uma estimativa do viés do modelo, isto é, o quanto o modelo escolhido se ajusta bem à amostra. Essa quantidade é chamada de *erro de treino*. Para obtermos uma estimativa da variância, precisamos calcular a medida de performance para observações não utilizadas no ajuste, que representem uma nova amostra do fenômeno sob estudo. Essa quantidade é chamada de *erro de teste*.

Na prática, nem sempre teremos à disposição uma nova base de dados para a estimação da variância. Uma alternativa nesses casos é utilizar técnicas de *validação cruzada*, que permite separar a base em observações utilizadas para *treinar* o modelo e observações para estimar sua performance. Essas técnicas serão o tema da próxima seção.

4.3 Métodos de reamostragem

A reamostragem consiste na técnica de gerar novas amostras a partir de uma base principal. As técnicas de reamostragem mais utilizadas são a *validação cruzada* e o *bootstraping*.

4.3.1 Validação cruzada

Como na maioria dos estudos não é possível obter facilmente novas observações, podemos calcular o erro de teste, a estimativa da variância do modelo, dividindo a amostra original em duas partes: uma utilizada para o ajuste do modelo (amostra de treino) e a outra para o cálculo do erro (amostra de teste), essa última agindo como se fosse um conjunto de novas observações. Essa técnica é conhecida como validação cruzada (James *et al.*, 2013). Há diversos tipos de validação cruzada, que variam a depender da forma utilizada para dividir a amostra. Nesta seção, apresentaremos os principais tipos de validação cruzada e discutiremos as vantagens e desvantagens de cada um.

Amostra de validação

A amostra de validação é a forma mais simples de validação cruzada. A estratégia consiste em dividir aleatoriamente as observações em um conjunto de treino, usado para ajustar o modelo, e outro de teste, utilizado exclusivamente para estimar o erro de teste.

A proporção de observações em cada uma depende do tamanho amostral. Costuma-se utilizar 30% da amostra original no conjunto de teste, mas esse número pode ser menor para amostras muito grandes (mais de 100 mil observações, por exemplo).

As maiores vantagens dessa técnica é a sua simplicidade e a necessidade de se ajustar o modelo uma única vez. No entanto, conforme discutido em James *et al.* (2013), a amostra de validação apresenta duas potenciais desvantagens:

- a estimativa do erro de teste pode apresentar alta variabilidade, dependendo de quais observações ficaram na amostra de treino e quais ficaram na amostra de validação;
- como a acurácia de modelos estatísticos é menor quando ajustados com menos observações, e apenas parte das observações são utilizadas para treinar o modelo, o erro de teste pode estar sendo superestimado.

A seguir, apresentaremos o LOOCV, um método de validação cruzada que não possui essas limitações.

Validação cruzada *leave-one-out* (LOOCV)

Considere uma amostra com n observações. A validação cruzada *leave-one-out* (LOOCV) consiste em rodar o modelo escolhido n vezes, sendo que, em cada ajuste, deixamos de fora a i-ésima observação, $i = 1, \ldots, n$, e a utilizamos para calcular o erro de teste. A estimativa final do erro de teste será então a média das n medidas parciais. Uma esquematização dessa técnica está representada na Figura 4.2.

Repare que, neste caso, todas as observações são utilizadas no ajuste do modelo e na estimativa do erro de teste, o que elimina as limitações da amostra de validação. No entanto, uma desvantagem

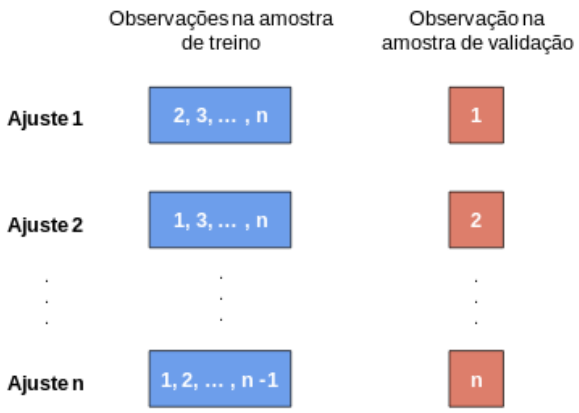

Figura 4.2: *Esquematização da validação cruzada leave-one-out.*

aqui é a necessidade de ajustar o modelo n vezes. Quando n é muito grande, a LOOCV pode exigir muito esforço computacional, inviabilizando a sua utilização em muitos casos.

Vale ressaltar que esse procedimento é utilizado para estimar as métricas de performance do modelo, sendo que ajuste do modelo final da análise contempla todas as observações da amostra. Dessa forma, ao fim desse procedimento, $n + 1$ modelos são ajustados: as n interações da LOOCV e o ajuste do modelo com todas as observações.

A seguir, apresentamos validação cruzada *k-fold*, uma generalização da LOOCV que não possui a contrapartida computacional.

K-fold

Podemos generalizar a LOOCV dividindo a amostra original aleatoriamente em k grupos com aproximadamente a mesma quantidade de observações. Então ajustamos o modelo k vezes, sendo que em cada ajuste selecionamos um grupo diferente como amostra de teste. Essa abordagem é chamada de *k-fold*. Note que a LOOCV é o caso especial em que $k = n$. Na prática, escolhemos valores de k entre 3 e 10, sendo que $k = 5$ é o mais utilizado (Figura 4.3).

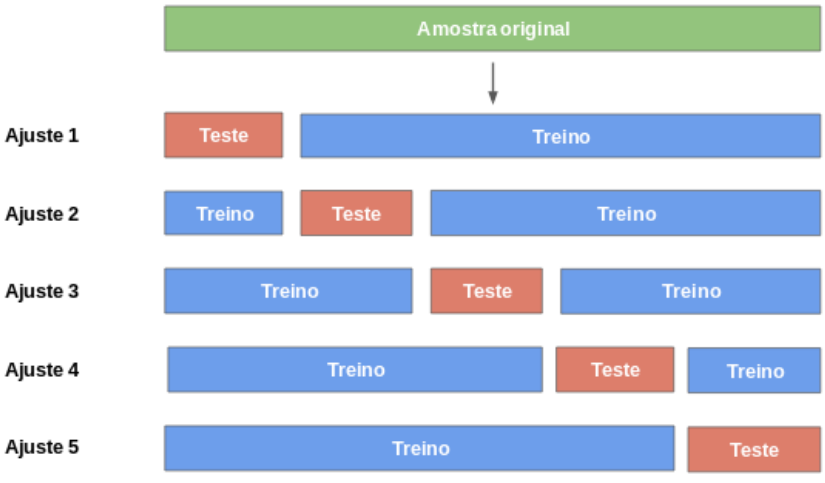

Figura 4.3: *Esquematização da validação cruzada k-fold, com* $k = 5$.

A maior vantagem da validação cruzada *k-fold* sobre a LOOCV é computacional. Em vez de ajustarmos o modelo n vezes, ajustamos apenas k, sendo que $k << n$. E como estamos utilizando todas as observações para treinar o modelo, não temos as limitações de se utilizar uma única amostra de validação.

Assim como na LOOCV, o objetivo desse procedimento é estimar o erro de predição. Ao fim, ajustamos o modelo utilizando todas as observações na amostra, que será considerado o modelo final. Assim, o modelo é ajustado $k + 1$ vezes: as k iterações da validação k-fold e o ajuste do modelo com todas as observações.

Como discutimos até agora, a validação cruzada é geralmente utilizada para avaliar a performance do modelo. A seguir, apresentaremos uma técnica de reamostragem muito utilizada também para a estimação de quantidades acerca dos parâmetros do modelo.

4.3.2 Bootstrapping

O *bootstrapping* é uma poderosa ferramenta estatística utilizada para quantificar incertezas associadas a estimadores e modelos estatísticos. Ela consiste em gerar m novas amostras a partir de sorteios com reposição da amostra original. Para cada uma das amostras geradas, ajustamos o modelo escolhido e guardamos as estimativas dos parâmetros. Ao repetirmos esse processo para as m amostras, teremos m estimativas diferentes para cada parâmetro do modelo. Assim, para cada parâmetro, podemos, por exemplo, calcular o desvio-padrão dessas m estimativas e utilizar essa medida como o erro-padrão associado ao coeficiente. Repare que os parâmetros do modelo devem ser estimados utilizando a amostra original. Nesse exemplo, o *bootstrapping* seria usado apenas para estimar a variabilidade dos coeficientes.

Essa técnica é utilizada principalmente quando não conhecemos a distribuição dos estimadores do modelo ou quando precisamos controlar outras fontes de variabilidade. Salvo e Geiger (2014) e Salvo *et al.* (2017), por exemplo, utilizaram o *bootstrapping* para estimar o erro-padrão dos coeficientes do modelo de regressão linear ajustado para associar a concentração de ozônio na cidade de São Paulo com a proporção estimada de veículos bicombustíveis rodando a gasolina. Segundo os autores, essa estratégia foi utilizada para contemplar a variação causada pelo erro de medida presente na estimação da proporção de carros rodando a gasolina e na medição das condições climáticas.

O *bootstrapping* também pode ser utilizado para a estimação da performance do modelo. Neste caso, cada uma das m amostras é utilizada como conjunto de treino e as observações que foram sorteadas em cada amostra é utilizada com conjunto de teste. Assim como na LOOCV, se m for muito grande, essa estratégia pode gerar um esforço computacional muito alto.

Mais informações sobre o *bootstrapping* podem ser encontradas em James *et al.* (2013).

4.4 Seleção de variáveis

Muitas vezes, na construção do modelo, incluímos variáveis que não são associadas com o fenômeno sob estudo. Isso acontece principalmente quando temos pouco conhecimento sobre o mecanismo gerador do fenômeno ou quando estamos justamente investigando quais fatores estão associados a ele.

Como variáveis irrelevantes geram uma complexidade desnecessária no modelo, é apropriado pesarmos em estratégias para retirá-las da análise, aumentando assim a interpretabilidade dos

resultados.

Nesta seção, apresentaremos algumas técnicas de seleção de variáveis que podem ser utilizadas em qualquer classe de modelos estatísticos.

4.4.1 Selecionando o melhor subconjunto de preditores

A maneira mais simples que podemos pensar para selecionar variáveis em um modelo é ajustar todas as possíveis combinações dos p preditores e avaliar qual produz o melhor ajuste segundo alguma métrica. Essa estratégia é chamada de *melhor subconjunto de preditores* (*best subset selection*, em inglês) e seu procedimento de seleção pode ser resumido pelos passo abaixo:

1. Ajustar o modelo nulo, sem nenhum preditor.
2. Para $k = 1, \ldots, p$, ajustar todos os modelos com k preditores e escolher o melhor entre eles, isto é, aquele com menor RSME ou maior R^2 por exemplo.
3. Para cada um dos $p+1$ modelos escolhidos, selecionar o melhor usando o R^2 ajustado, RMSE calculado por validação cruzada (erro de teste), AIC ou BIC[6].

Observe que a métrica utilizada para selecionar o modelo final deve ser penalizada pelo número de parâmetros, pois, caso contrário, escolheríamos sempre o modelo com mais preditores.

Para um número relativamente pequeno de variáveis, selecionar o melhor subconjunto de preditores é uma estratégia conceitualmente simples e de fácil execução. No entanto, conforme p cresce, essa técnica pode se tornar computacionalmente inviável. Na Tabela 4.2 apresentamos os 7 modelos que precisaríamos ajustar se tivéssemos 3 preditores, X_1, X_2 eX_3, e um modelo de regressão linear (Seção 3.1). Para $p = 20$, por exemplo, precisaríamos rodar mais de um milhão de modelos, o que poderia inviabilizar a execução dessa estratégia.

Tabela 4.2: *Modelos de regressão linear que devem ser ajustados para selecionar o melhor subconjunto de variáveis no caso com 3 preditores.*

Uma variável	Duas variáveis	Três variáveis
$Y = \beta_0 + \beta_1 X_1 + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$	
$Y = \beta_0 + \beta_1 X_2 + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
$Y = \beta_0 + \beta_1 X_3 + \epsilon$	$Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \epsilon$	

A seguir, apresentamos algumas estratégias computacionalmente eficientes para aplicarmos em problemas com muitos preditores.

4.4.2 Stepwise

Os métodos *stepwise* são algoritmos de seleção de variáveis que visam encontrar o melhor subconjunto de preditores dentro de um conjunto restrito de combinações em vez de ajustar todos os 2^p modelos possíveis.

A diferença entre cada método *stepwise* está em como as variáveis são adicionas ou retiradas do modelo em cada passo. Os mais utilizados são o *foward stepwise* e o *backward stepwise*.

O *foward stepwise* consiste na execução dos seguintes passos:

[6] O AIC e o BIC são medidas da qualidade do ajuste penalizadas pelo número de parâmetros do modelo. Mais informações, consultar James *et al.* (2013)

1. Ajuste o modelo nulo (M_0), sem preditores.
2. Ajuste todos os p modelos com 1 preditor e escolha o melhor[7] (M_1).
3. Ajuste todos os $p-1$ modelos com 2 preditores que contenham o preditor selecionado no passo anterior e escolha o melhor (M_2).
4. De forma análoga, ajuste os modelos com 3, 4, ..., p preditores, mantendo sempre como base o modelo obtido anteriormente, e em cada passo escolha o melhor (M_3, M_4, ..., M_p).
5. Escolha o melhor modelo entre M_0, M_1, ..., M_p utilizando erro preditivo, AIC, BIC ou R^2 ajustado.

Repare que o *foward stepwise* diminui o número de modelos ajustados de 2^p para $1+p(p+1)/2$. Para $p=20$, o número de modelos diminui de 1.048.576 para 211.

A ideia do método *backward stepwise* é parecida com a do *foward*. A diferença é que começamos no passo 1 com o modelo completo (M_p), com todos os preditores, e nos passos seguintes retiramos cada um dos preditores e ajustamos os modelos correspondentes, selecionando sempre aquele com maior R^2 ($M_{p-1}, M_{p-2}, \ldots, M_0$). Ao fim, escolhemos o melhor entre os modelos M_0, M_1, ..., M_p utilizando erro preditivo, AIC, BIC ou R^2 ajustado. O número de modelos ajustados nesse caso é igual ao do *foward stepwise*.

Ainda existem métodos *stepwise* híbridos, nos quais os preditores são adicionados sequencialmente, assim como no *foward stepwise*, mas em cada etapa é avaliado se um dos preditores já incluídos deve ou não sair do modelo. Essa estratégia tenta considerar mais modelos, chegando mais próximo da seleção do melhor sub-conjunto discutida na seção anterior. Para mais informações, consultar Nelder e Wedderburn (1972).

4.5 Regularização

Os métodos de seleção de sub-conjuntos de preditores apresentados nas seções anteriores envolvem o ajuste de diversos modelos e a escolha do melhor segundo alguma métrica. Uma outra forma de selecionar variáveis é a partir das *técnicas de regularização*. Essas técnicas, a princípio, envolvem o ajuste de um único modelo e introduzem no processo de estimação penalizações que limitam as estimativas dos coeficientes, encolhendo seus valores na direção do zero.

A utilização da regularização pode levar a uma redução substancial da variância do modelo ao custo de um pequeno aumento no viés. Apresentaremos nesta seção as formas mais utilizadas de regularização: a *regressão ridge* e o LASSO.

Regressão Ridge

De uma forma geral, o processo de estimação dos parâmetros de um modelo consiste na minimização de uma função de perda $L(y, f(x))$ que depende dos dados observados (x, y) e do modelo escolhido ($f(\cdot)$). As técnicas de regularização consistem em adicionar uma penalidade nessa função de perda, de tal forma que os coeficientes dos preditores pouco associados à variável resposta sejam encolhidos na direção do zero.

[7] Maior R^2, por exemplo.

No caso da regressão ridge (James *et al.*, 2013), essa penalização é dada por

$$L(y, f(x)) + \lambda \sum_{j=1}^{p} \beta_j^2,$$

sendo $\beta_1, \ldots, \beta_p$ os parâmetros do modelo $f(\cdot)$ e λ um hiperparâmetro[8] que controla o impacto da penalização nas estimativas dos coeficientes. Quando $\lambda = 0$, o termo é anulado e as estimativas são calculadas sem penalização. Conforme $\lambda \longrightarrow \infty$, os coeficientes β_j passam a ser penalizados, encolhendo seus valores na direção do zero. A vantagem desse comportamento está na potencial redução da variância do modelo, em troca de um pequeno aumento do viés, já que os coeficientes menos importantes recebem cada vez menos peso. Assim, a regularização é uma alternativa para lidarmos com o balanço entre viés e variância discutido na Seção 4.1.

No caso da regressão ridge, é possível mostrar que, para qualquer $i = 1, \ldots, p$, $\beta_i = 0$ apenas se $\lambda = \infty$. Isso significa que não estamos fazendo seleção de variáveis, isto é, o modelo ajustado sempre terá todos os preditores. Apesar de estarmos melhorando a performance do modelo diminuindo o peso dos preditores menos importantes, isso pode não ser o ideal quando quisermos de fato eliminar variáveis do modelo. Nesses casos, uma boa alternativa é utilizar o LASSO.

Least absolute shrinkage and selection operator (LASSO)

O LASSO (*least absolute shrinkage and selection operator*) é uma técnica análoga à regressão ridge, mas com penalização dada por

$$L(y, f(x)) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Para λ grande o suficiente, essa penalização força que alguns dos coeficientes sejam estimados exatamente como 0 e os correspondentes preditores associados serão eliminados do ajuste. Assim, ao utilizarmos o LASSO, estamos ao mesmo tempo reduzindo a variância do modelo e executando seleção de variáveis.

Um ponto importante sobre a aplicação das técnicas de regularização é a escala dos preditores. A maioria dos processos de estimação usuais são invariantes à escala em que os preditores foram medidos, isto é, ajustar o modelo usando o preditor X_1 ou cX_1, c uma constante qualquer, não mudará a interpretação dos resultados. No caso da regressão ridge e do LASSO, a escala dos preditores influenciam não só a estimativa dos próprios coeficientes, mas também a estimativa dos outros parâmetros do modelo. Dessa forma, um passo importante anterior à aplicação dessas técnicas é a padronização dos preditores, de tal forma que todos fiquem com a mesma média e variância. Essa padronização pode ser feita a partir da fórmula

$$\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}, \tag{4.1}$$

sendo o denominador dessa expressão a estimativa do desvio-padrão do j-ésimo preditor. Consequentemente, todos os preditores terão média 0 desvio-padrão igual a 1.

[8] Hiperparâmetros são parâmetros que não são estimados diretamente pelos dados.

Embora haja muita discussão sobre a validade de testes de hipóteses do tipo $\beta = 0$ para o LASSO, já que o algoritmo zera automaticamente os coeficientes menos importantes, alguns trabalhos vêm surgindo nos últimos anos sobre o calculo do erro-padrão e o desenvolvimento de testes paras as estimativas (Javanmard e Montanari, 2014; Lockhart *et al.*, 2014). Uma boa alternativa para avaliar a variabilidade das estimativas dos coeficientes é utilizar o *bootstrapping*.

Para uma discussão mais aprofundada sobre a interpretação da regressão ridge e do LASSO, consulte o Capítulo 6 de James *et al.* (2013). Para o desenvolvimento matemático dessas técnicas, o Capítulo 5 de Hastie *et al.* (2008) é uma ótima referência.

4.6 Quantificando a importância dos preditores

Nas últimas seções, discutimos técnicas para removermos do modelo as variáveis que não ajudam a explicar a variabilidade da variável resposta. Em alguns casos, também gostaríamos de saber, entre os preditores que permaneceram no modelo, quais são os mais importantes.

Os valores p são amplamente utilizados para definir as variáveis estatisticamente significantes para explicar a variável resposta. Dado um coeficiente β, o valor p associado representa uma medida de evidência a favor da hipótese $\beta = 0$ e pode ser utilizado tanto para seleção de variáveis quanto para quantificar a magnitude de uma associação. Podemos interpretar o valor p como quanto a estimativa encontrada seria inverossímil caso o verdadeiro valor de β fosse 0. Se o valor p for muito baixo (próximo de zero), significa que a estimativa obtida teria baixa probabilidade caso β fosse 0 e então rejeitamos a hipótese de que esse coeficiente é nulo. Caso contrário, se o valor p for alto, significa que a estimativa obtida não é inverossímil em um cenário em que β é zero, e então não rejeitamos a hipótese de que $\beta = 0$.

Ao cálculo do valor p está associado uma estatística de teste, que também pode ser usada para quantificar a importância dos preditores. No modelo de regressão linear, por exemplo, o valor da estatística do teste t pode ser utilizada, de tal forma que, quanto maior o valor absoluto da estatística, maior será a importância do preditor para explicar a variável resposta.

Em alguns casos, a variação no erro preditivo quando um preditor é eliminado do modelo é utilizada como medida de importância. Essa métrica mais geral é bastante utilizada em modelos de regressão que envolvem funções suavizadoras, como os modelos aditivos generalizados.

Já para para a regressão ridge ou o LASSO, em que padronizamos as variáveis explicativas, uma medida de importância pode ser o próprio valor do coeficiente.

As métricas de importância vão depender sempre do modelo utilizado. De uma forma geral, os programas estatísticos já possuem métricas de importância implementadas. No R, a função `varImp()` do pacote `caret` calcula uma medida de importância para a maioria dos modelos disponíveis.

4.7 Modelos de árvores

Modelos baseados em árvores (Hastie e Tibshirani, 1990; James *et al.*, 2013) são algoritmos bastante utilizados tanto para regressão quanto para classificação. Esses métodos envolvem a segmentação do espaço gerado pelas variáveis explicativas em algumas regiões mais simples, onde a média ou a moda da variável resposta são utilizadas como predição.

As chamadas árvores de decisão são modelos conceitualmente e computacionalmente simples, bastante populares pela sua interpretabilidade, apesar da precisão inferior quando comparados com modelos mais flexíveis. Generalizações desse modelos, como as *random forests*, costumam apresentar alta precisão, mesmo quando comparadas a modelos lineares, porém são pouco interpretáveis.

Nesta seção, introduziremos os principais conceitos por trás das árvores de decisão e das *random forests*.

4.7.1 Árvores de decisão

As árvores de decisão se baseiam no particionamento das variáveis explicativas, de tal forma que as regiões formadas gerem predições para a variável resposta com baixo erro segundo alguma métrica (geralmente RMSE para regressão, o foco deste trabalho). O particionamento é feito a partir de *regras* que dividem o espeço gerado pelas variáveis explicativas. Cada regra é representada por um *nó* e a cada partição criada podem ser acrescentadas mais regras.

Na Figura 4.4, apresentamos um exemplo de árvore de decisão para a concentração de ozônio explicada pela temperatura. Para interpretá-la, basta começarmos pelo primeiro nó (a caixa mais alta da figura) e seguir as regras de decisão até encontrarmos um nó final (uma das caixas no nível mais baixo). Cada nó final apresenta a estimativa para as observações que caíram naquela partição e o número de observações dentro da partição (absoluto e proporcional ao tamanho da amostra). A Figura 4.4 indica que dias com temperatura menor de 26° C serão preditos com concentração de ozônio igual a 38 μ/m^3; dias com temperatura entre 27 e 29° C serão preditos com concentração de ozônio igual a 69 μ/m^3; e dias com temperatura maior de 92° C serão preditos com temperatura igual a 92° C. Além de conseguirmos predizer facilmente o nível de ozônio a partir da temperatura, também podemos observar que a concentração de ozônio aumenta com a temperatura. Na prática, podemos utilizar as árvores de decisão com quantos preditores forem necessários, sendo que cada nó poderá conter uma regra com um preditor diferente.

Observe que no exemplo temos 3 nós finais, mas, teoricamente, poderíamos continuar particionando a temperatura até cada possível valor ter o seu próprio nó. Esse seria um caso de árvore de decisão sobreajustada, apresentando pouco poder de generalização. A escolha do número de nós finais é feita a partir de uma técnica conhecida como *poda*, que consiste em usar validação cruzada para definir a melhor altura para a árvore.

As árvores de decisão copiam bem o processo de tomada de decisão do cérebro humano, e por isso são mais simples de interpretar até mesmo que o modelo de regressão linear. No entanto, elas apresentam baixo poder preditivo e raramente são utilizadas para descrever processos muito complexos. A seguir, apresentaremos as *random forests*, que abrem mão da interpretabilidade em troca de um alto grau de precisão.

4.7.2 Random Forests

As chamadas *random forests* utilizam diversas árvores de decisão para predizer o valor de uma observação. Em cada iteração do algoritmo, selecionamos uma amostra dos p preditores e construímos uma árvore de decisão para esses m preditores. Esse procedimento é repetido M vezes, gerando M árvores diferentes. A amostragem dos preditores garante que as árvores geradas não sejam muito correlacionadas e permite que o modelo aprenda também com preditores de menor peso. A predição final será então a média das predições de cada árvore. Geralmente, m é escolhido como $\sqrt{p}$, mas

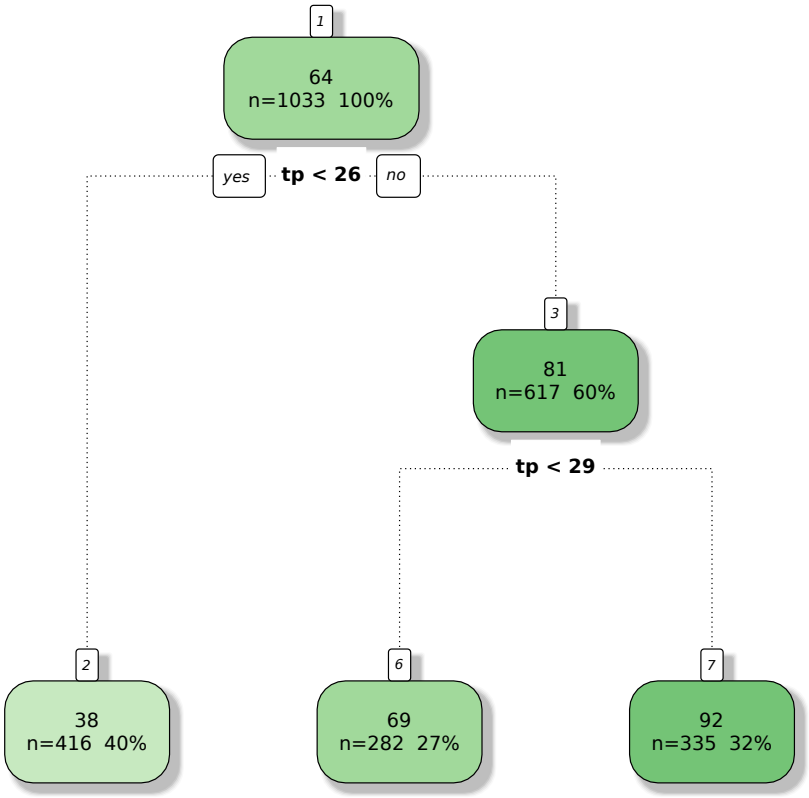

Figura 4.4: *Exemplo de uma árvore de decisão para a concentração de ozônio explicada pela temperatura.*

esse hiperparâmetro pode ser definido utilizando validação cruzada. M por volta de 200 costuma ser suficiente para gerar bons resultados[9].

Ao contrário das árvores de decisão, as *random forests* não são diretamente interpretáveis, já que se baseiam nas estimativas de centenas de árvores. Medidas de importância das variáveis podem ser utilizadas para avaliar quais preditores são os mais relevantes para explicar a variável resposta. Em geral, usa-se a redução média do RMSE gerada por todas partições do preditor nas M árvores.

Uma outra técnica utilizada para interpretar modelos como as *random forests* é o LIME (*Local snterpretable Model-Agnostic Explanations*) (Ribeiro *et al.*, 2016). Dado qualquer modelo estatístico, essa técnica consiste em explicar quais preditores foram responsáveis pela predição de uma observação específica. Nos últimos anos, o LIME vem sendo muito utilizado para explicar, corrigir e aperfeiçoar modelos não interpretáveis.

Essa técnica assume que todo modelo complexo prevê valores parecidos para duas observações muito próximas, sendo possível ajustar um modelo simples ao redor de uma única observação para copiar o comportamento global do modelo complexo. Assim, a partir de um modelo interpretável podemos obter para cada observação uma explicação sobre a predição feita pelo modelo não-interpretável.

O algoritmo consiste em:

1. Para cada predição a ser explicada, permutar a observação n vezes.

[9]Para valores de M muito altos (maior que 200), a amostragem dos preditores passa a criar árvores mais parecidas com as já existentes, o que não gera maiores ganhos de precisão.

2. Predizer cada uma das observações permutadas usando o modelo complexo.

3. Calcular uma medida de distância e similaridade entre as permutações e a observação original. Geralmente a distância de Gower é utilizada (Gower, 1971).

4. Selecionar as m variáveis mais importantes utilizadas pelo modelo complexo para explicar os dados permutados.

5. Ajustar um modelo interpretável aos dados permutados, utilizando as predições do modelo complexo como variável resposta e as m variáveis selecionadas no passo anterior como preditores, ponderando pela medida de similaridade com a observação original.

6. Usar as estimativas desse modelo simples como as explicações para o comportamento local do modelo complexo.

A implementação do LIME exige definir como as observações serão permutadas, qual medida de similaridade será usada, qual o valor de m e qual modelo interpretável será utilizado. Uma boa discussão sobre esses pontos pode ser encontrada no *vignette* do pacote `lime`, no qual a técnica foi implementada na linguagem R. O texto pode ser acessado por este link: https://goo.gl/Nu9ZsA.

Capítulo 5

Poluição e uso de combustíveis

Neste Capítulo, vamos utilizar as técnicas apresentadas até aqui para analisar um problema comum em estudos de poluição do ar: a relação entre uso de combustíveis e a concentração de poluentes.

5.1 Etanol e ozônio

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 diminuir o tamanho da frota de veículos (ou ao menos restringir o seu uso) esbarram em fatores políticos e econômicos, os estudos nessa área têm como objetivo encontrar combustíveis menos poluentes, alternativas ao diesel e à gasolina.

O bio-etanol é uma fonte quase-renovável de energia que pode ser produzida a partir de matéria prima agrícola. É amplamente utilizado no Brasil e nos Estados Unidos, seja puro (conhecido como E100) ou como aditivo da gasolina (gasohol; conhecido como E20 ou E25, de acordo com a porcentagem adicionada à gasolina, 20% ou 25%). Comparado com a gasolina convencional, o etanol é considerado um combustível cuja queima gera menores concentrações de material particulado (PM), óxidos de nitrogênio (NO_X), monóxido de carbono (CO) e dióxido de carbono (CO_2), sendo uma boa opção para reduzir a poluição do ar e o aquecimento global.

Em um experimento controlado na cidade de Fairbanks, no Alasca, por exemplo, 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 em dias mais frios. Yoon *et al.* (2009) conduziram uma investigação similar e concluíram que a combustão de etanol e da mistura E85 (85% etanol e 15% gasolina) emitia concentrações inferiores de hidrocarbonetos, monóxido de carbono e óxidos de nitrogênio quando comparados com a gasolina sem aditivos sob diversas condições experimentais.

Apesar de diversos trabalhos, além do senso comum, considerarem o etanol como uma alternativa menos poluente à gasolina, quando as emissões de veículos abastecidos com etanol são associadas à concentração ambiente de ozônio (O_3), os estudos têm apontado para uma direção diferente. Pereira *et al.* (2004), por exemplo, expuseram câmaras contendo etanol puro e gasool (mistura de 22-24% etanol em gasolina) ao sol 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. Jacobson (2007) juntou modelos de previsão para a poluição do ar e o clima com inventários de emissões futuras e dados populacionais e epidemiológicos para examinar o efeito da troca da gasolina por E85 na incidência de câncer e casos de mortalidade e hospitalização em Los Angeles, em particular, e nos Estados Unidos, como um todo. O autor concluiu que o risco de câncer era o mesmo com a utilização dos dois combustíveis, mas uma frota futura de veículos rodando com E85 aumentaria a hospitalização por complicações relacionadas à poluição por 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 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 em São Paulo está associado a maiores concentrações do poluente. Esse estudo foi ampliado por Salvo *et al.* (2017), utilizando dessa vez dados de 2008 a 2013 e analisando também o efeito na concentração de partículas ultrafinas. Os resultados apontaram novamente associação entre o maior uso de etanol e aumento na concentração de ozônio, mas queda no número de partículas ultrafinas. Por fim, Salvo e Wang (2017) sumarizam essa discussão analisando a variação nos níveis de ozônio em quatro períodos de maior penetração do etanol. Eles concluíram que a química atmosférica da cidade de São Paulo é limitada em compostos orgânicos voláteis[2] e que esses períodos estavam associados a maiores concentrações do poluente.

Neste capítulo, utilizaremos os dados disponibilizados por Salvo *et al.* (2017) para explorar análises cujo objetivo é entender o mecanismo de formação dos poluentes a partir da relação entre as variáveis associadas.

5.2 Dimensionando a análise

> Far better an approximate answer
> to the right question, which is often vague,
> than an exact answer to the wrong question,
> which can always be made precise.
> — John Tukey

Dimensionar a análise significa entender a complexidade do problema a ser resolvido. Embora pareça uma tarefa simples, é muito comum nos perdermos durante a modelagem quando o objetivo do estudo não está muito bem definido. Essa deve ser sempre a primeira etapa da modelagem e cumpri-la envolve conhecimento sobre o fenômeno de interesse e, principalmente, pensamento crítico.

Conforme discutido no Capítulo 3, o mecanismo por trás do fenômeno sob estudo pode ser muito complexo. Muitas vezes não conseguimos nem mesmo identificar todas as variáveis envolvidas no processo. A formação do ozônio troposférico certamente é um exemplo disso. É um processo espaço-temporal que depende de reações químicas complicadas e um grande número de variáveis, a maior parte delas difícil de ser medida com precisão.

Sendo assim, diante esses cenários complicados, devemos ter cuidado para:

[1] Proporção de carros a gasolina entre os carros bicombustíveis.
[2] Um dos principais percursores do ozônio. Sua fonte primária é a queima parcial ou evaporação de etanol.

1. não escolher um modelo muito complexo para responder uma pergunta que poderia ser respondida por uma análise mais simples;

2. não escolher um modelo muito simples para descrever associações complexas entre as variáveis.

Essa medida de complexidade reflete o balanço entre viés e variância discutido na Seção 4.1. No primeiro caso, teríamos problema com a interpretabilidade do modelo e, no segundo, um problema de baixa precisão. Dimensionar a análise está não só ligada com a interpretação correta do estudo, mas também interfere diretamente na escolha do modelo a ser ajustado.

O objetivo do estudo feito por Salvo *et al.* (2017), tomado como exemplo neste capítulo, é investigar como a proporção de carros rodando a gasolina em relação àqueles rodando a etanol pode ajudar a explicar a variabilidade da concentração de ozônio. De maneira geral, queremos descobrir qual é relação entre a variável resposta e um dos preditores, isto é, entender melhor o mecanismo escondido dentro da caixa preta apresentada na Figura 3.1. Os dados disponibilizados pelos autores, além de variáveis de calendário e da proporção estimada de carros rodando a gasolina na cidade de São Paulo, contêm informação horária da concentração de ozônio, clima e trânsito. Uma ideia inicial seria propor um modelo para relacionar as observações horárias do ozônio com os preditores. Idealmente, esse modelo deveria contemplar a correlação das observações medidas no mesmo dia, o que pode ser feito, por exemplo, usando um modelo misto. No entanto, se examinarmos a proporção estimada de carros rodando a gasolina, verificamos que ela varia, no máximo, de um dia para o outro. Isso significa que, para responder a pergunta do estudo, não precisamos modelar a relação entre as variáveis dentro de cada dia, já que o principal preditor não apresenta essa granularidade.

Uma boa estratégia nesse caso é agregar as observações feitas no mesmo dia, utilizando, por exemplo, a média ou máxima diária. Dessa forma, eliminamos necessidade de modelar a correlação entre as medidas feitas no mesmo dia. Salvo *et al.* (2017), além de considerarem a média diária, excluíram da análise os meses de junho a setembro, período no qual a concentração de ozônio é menor devido às baixas temperaturas. Repare que retirar os meses frios também diminui a complexidade do problema, já que estamos eliminando um forte componente sazonal, que precisaria ser modelado. No entanto, neste caso, essa escolha restringe a inferência do modelo, que não poderá ser generalizada para a temporada de inverno. Como o maior interesse em estudos de poluição está na redução dos índices mais altos, é razoável limitar a conclusão da análise apenas ao período de maiores concentrações.

Outra consequência da exclusão desses meses está na escolha do modelo. Removendo esses dias da amostra, não teremos mais um problema de série temporal usual, já que as medidas não são mais equidistantes no tempo. Nesse sentido, modelos de regressão se tornam mais atraentes para modelar os dados.

As estratégias apresentadas até aqui consideram apenas o objetivo do estudo e conhecimento prévio sobre as variáveis disponíveis. Podemos continuar reduzindo a complexidade do problema, além de melhorarmos nossa intuição sobre o modelo mais adequado, extraindo informação dos dados a partir de uma análise exploratória.

5.3 Análise exploratória

O objetivo desta análise exploratória é investigar o comportamento dos preditores e entender como eles se relacionam com a concentração de ozônio.

Por pragmatismo e como a maior parte dos resultados são compartilhados pelas outras estações, como apresentado na Seção 2.1, apresentaremos aqui apenas a análise da estação Dom Pedro II. A análise exploratória completa pode ser visualizada em https://rpollution.com/categories/ozônio.

Na Seção 2.1.1, vimos como a concentração média diária do ozônio se comporta ao longo do dia (Figura 2.2). Como no período da manhã o ozônio ainda está sendo gerado e, no final do dia, ele já foi quase inteiramente consumido, é razoável analisarmos apenas a média diária no intervalo de pico, entre o meio-dia e as 17 horas. Isso significa que vamos relacionar o uso de etanol com o pico diário do ozônio, o que é suficiente para responder a pergunta de interesse.

Mesmo considerando apenas a concentração de ozônio medida entre meio-dia e 17 horas, as condições climáticas e de tráfego no período da manhã podem ser fatores importantes na formação do poluente. Assim, além das médias dessas variáveis no período da tarde, poderíamos considerar também valores médios pela manhã (entre 8 e 11 horas).

Vamos investigar inicialmente a associação entre a concentração de ozônio e a proporção estimada de carros a gasolina rodando na cidade. Pela Figura 5.1, observamos que existem dois picos de utilização de gasolina durante o período analisado, um no começo de 2010 e outro no começo de 2011. Após o segundo pico, a proporção estimada de carros a gasolina varia pouco, próximo ao 50%. Analisando o gráfico de dispersão na Figura 5.2, não encontramos indícios claros de associação entre o ozônio e a proporção estimada de carros a gasolina. Isso significa que, se as variáveis estão associadas, essa relação está sendo mascarada pelo efeito dos outros preditores.

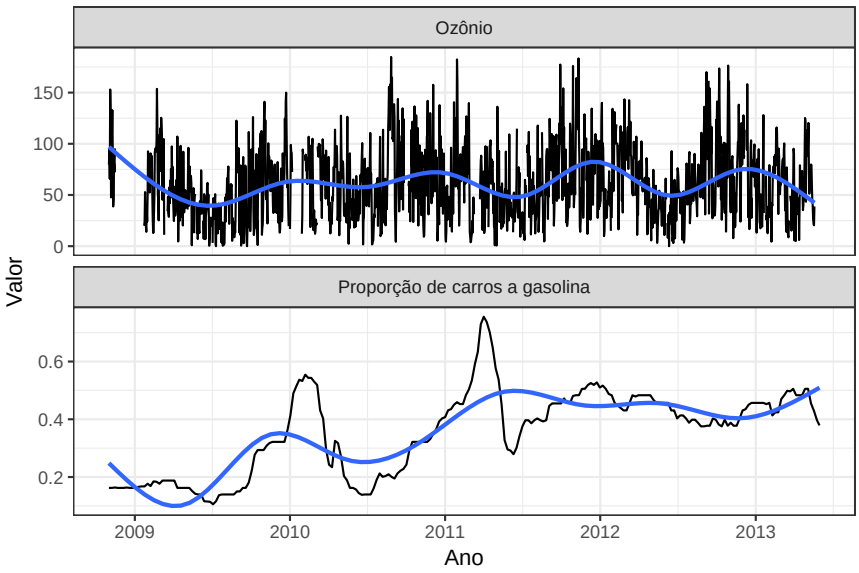

Figura 5.1: *Séries da concentração de ozônio diária média e da proporção estimada de carros a gasolina rodando na cidade. Dados da estação Dom Pedro II, de 2008 a 2013.*

Nas Figuras 5.3 e 5.4, podemos observar a distribuição da temperatura média, pela manhã e pela tarde, para cada mês do ano, e também comparar as séries de cada variável. Repare que a temperatura de manhã é muito mais sensível às estações do ano, enquanto a temperatura à tarde varia mais. Pela Figura 5.5, observamos que a concentração de ozônio parece mais associada com a temperatura pela tarde, o que é razoável devido ao papel da luz solar no mecanismo gerador do poluente.

Repetindo a mesma análise para as outras variáveis climáticas[3], podemos concluir:

[3]Os gráficos para os outros preditores podem ser encontrados em https://www.rpollution.com/flexdashboards/ozonio-

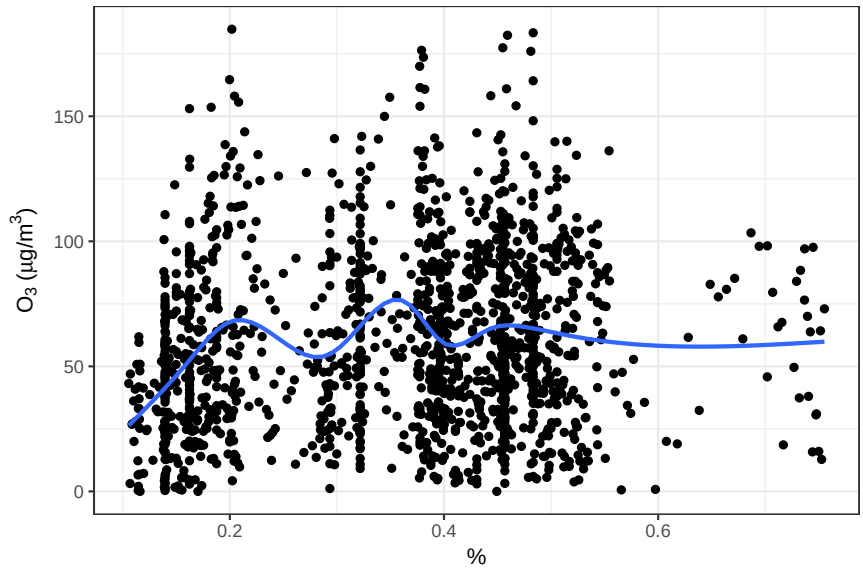

Figura 5.2: *Gráfico de dispersão da concentração de ozônio contra a proporção estimada de carros rodando a gasolina na cidade. Dados da estação Dom Pedro II, de 2008 a 2013.*

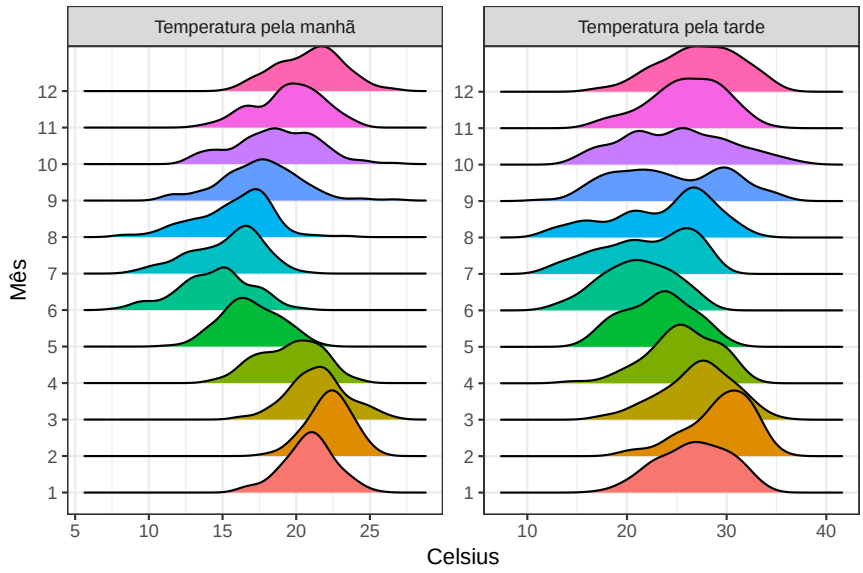

Figura 5.3: *Gráficos ridge da temperatura diária média nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.*

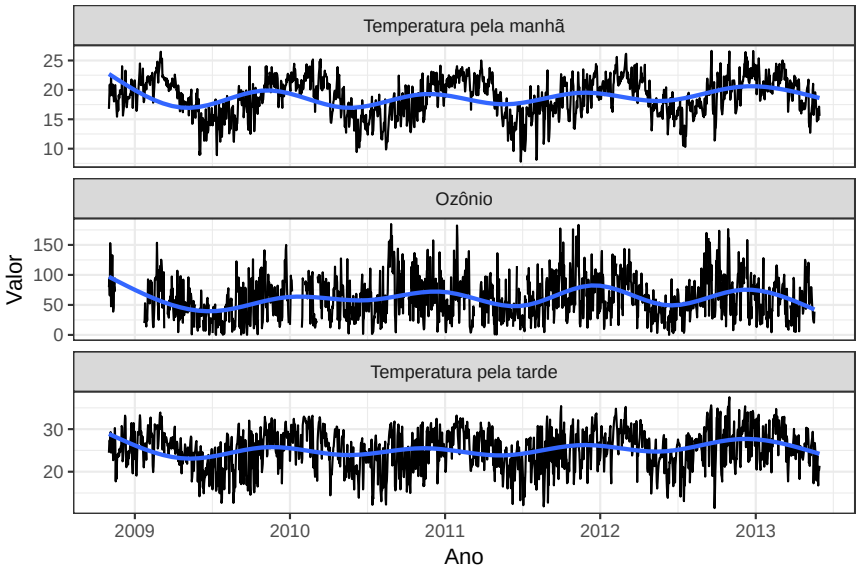

Figura 5.4: *Gráficos das séries da concentração de ozônio e da temperatura diária média nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.*

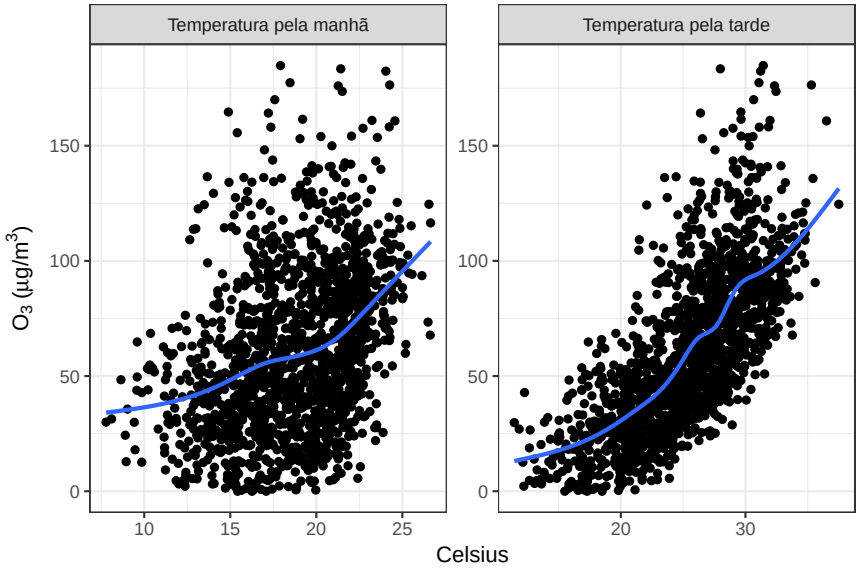

Figura 5.5: *Gráficos de dispersão da concentração de ozônio pela temperatura diária média nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.*

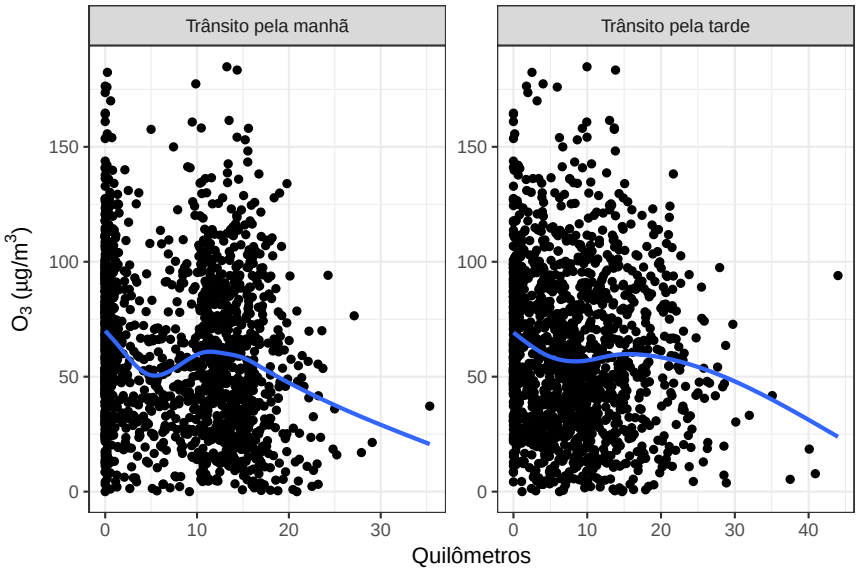

Figura 5.6: *Gráficos de dispersão da concentração de ozônio pelo congestionamento diário médio, na região da estação de monitoramento, nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.*

- dias com maior radiação estão associados a maiores níveis de ozônio;
- categorizando a variável precipitação em "Choveu no período" e "Não choveu no período"[4], os períodos sem chuva estão associados a maiores concentrações de ozônio;
- umidade alta, principalmente à tarde, está associada com menores concentrações de ozônio;
- a relação entre a velocidade do vento, tanto de manhã quanto à tarde, e a concentração de ozônio não é muito clara; e
- parece haver uma leve associação entre a ocorrência de inversões térmicas e maiores concentrações de ozônio.

Analisando agora o trânsito diário médio na região da estação de monitoramento (Figura 5.6), não parece ficar clara qual a relação com a concentração de ozônio. No entanto, se observamos agora a Figura 5.7, observamos que a concentração diária média de ozônio é, em geral, maior nos fins de semana, enquanto o congestionamento tende a ser menor nesses dias. Como não há motivos para acreditar que as condições climáticas sejam diferentes nos fins de semana, é razoável supor que a concentração de ozônio é maior em dias de pouco tráfego.

Se essa relação for verdadeira, isso implica que, independentemente de qual for a relação entre o uso de etanol a concentração de ozônio, as emissões veiculares tendem a diminuir os níveis do poluente. Para avaliar essa hipótese melhor, vamos estudar a concentração de ozônio em dias com maior proporção estimada de carros rodando a álcool e em dias com maior proporção estimada de carros rodando a gasolina. Como podemos observar pela Figura 5.8, a concentração de ozônio é maior em dias de menor tráfego independentemente de qual combustível está sendo mais utilizado na cidade.

clima-sp/dash-ozonio-clima-sp.html.

[4] A escolha pela categorização se deve ao alto número de zeros (dias sem chuva) que essa variável possui.

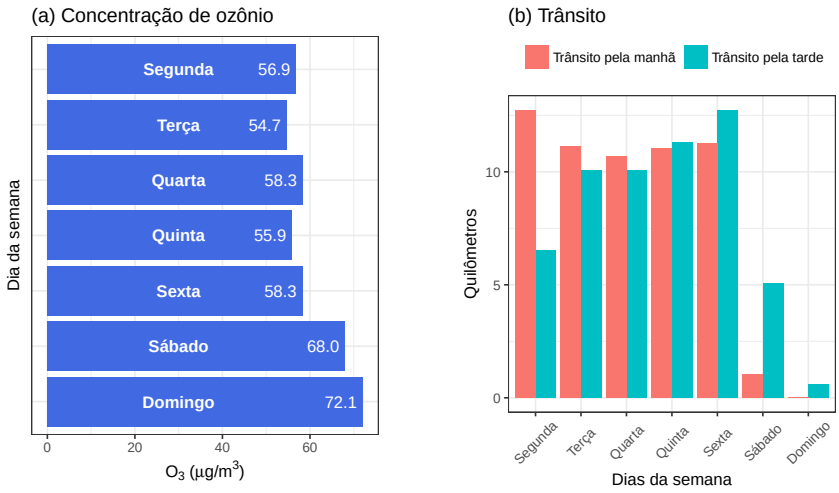

Figura 5.7: *Relação entre a concentração de ozônio e o congestionamento na região da estação de monitoramento ao longo da semana. (a) Concentração de ozônio diária média ao longo da semana. (b) Congestionamento diário médio, no período da manhã e da tarde, na região da estação de monitoramento ao longo da semana. Dados da estação Dom Pedro II, de 2008 a 2013.*

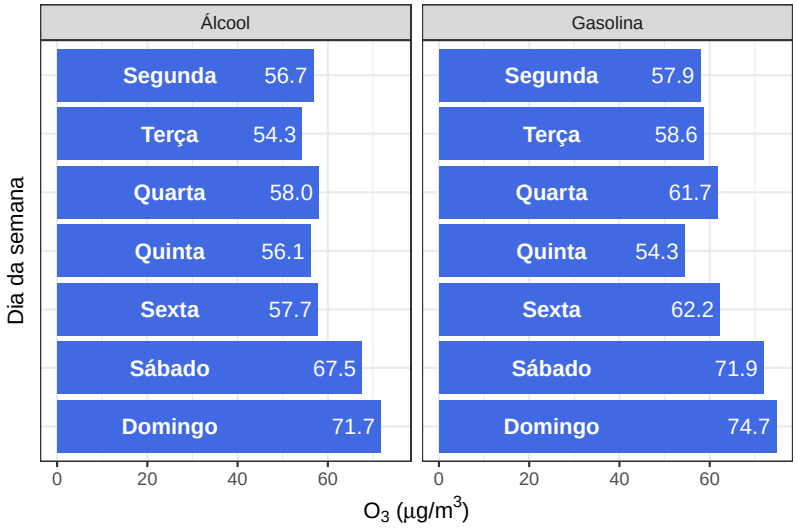

Figura 5.8: *Concentração de ozônio diária média ao longo da semana em dias com maior proporção de estimadas de carros rodando a álcool e em dias com maior proporção estimada de carros rodando a gasolina. Dados da estação Dom Pedro II, de 2008 a 2013.*

Devido à complexidade do problema e ao grande número de variáveis, a análise exploratória pode seguir em diversas direções. Poderíamos, por exemplo, analisar a associação entre os preditores para buscar indícios de interação, investigar a concentração de ozônio fora e durante as férias escolares, avaliar também o congestionamento médio em toda cidade em vez de apenas o congestionamento na região da estação e, é claro, generalizar a análise para as outras estações de monitoramento. No entanto, como o objetivo dessa seção era passar uma visão geral da análise exploratória, vamos prosseguir com o ajuste dos dados.

5.4 A análise conduzida por Salvo *et al.* (2017)

Antes de discutir as diferentes estratégias que adotamos para investigar a relação entre a concentração de ozônio e o uso de gasolina/etanol, vamos apresentar os resultados da análise feita por Salvo *et al.* (2017).

Como discutido anteriormente, os autores optaram por remover da análise os dias meses de junho a setembro, optando por uma amostra mais homogênea em relação às condições climáticas. Além disso, eles juntaram as observações de todas as 12 estações de monitoramento, formando uma única amostra em que cada linha da base representa as medidas de um dia para uma das estações.

O modelo final apresentado por Salvo *et al.* (2017) foi o modelo de regressão linear 3.3, considerando as variáveis apresentadas na Tabela 5.1.

Tabela 5.1: *Preditores considerados pelo modelo para a concentração de ozônio ajustado em Salvo et al. (2017).*

Tipo	Variáveis	Número de parâmetros
Etanol	Proporção estimada de carros a gasolina.	1
Estação	Indicador de estação.	11
Calendário	Indicadores de dia da semana, semana do ano, férias e feriados públicos.	44
Tendência	Termo de tendência geral e específica para cada estação.	12
Clima	Temperatura, radiação, umidade, velocidade do vento e indicadores de precipitação e de inversão térmica.	9
Trânsito	Indicadores de congestionamento na região da estação de monitoramento, na cidade como um todo e inauguração de vias importantes.	18
Total	**16 preditores + intercepto**	**96 parâmetros***

*95 parâmetros dos preditores + 1 parâmetro do intercepto.

A estimativa reportada para o parâmetro referente à proporção de carros rodando a gasolina foi -16.66 ± 10.01 (mais ou menos dois desvios-padrão)[5], o que indica que o aumento da proporção estimada de carros rodando a gasolina na cidade está associada com a diminuição da concentração

[5]Esse foi o resultado de um dos modelos, no qual foi utilizado mínimos quadrados ordinários para estimação dos parâmetros e *bootstrapping* para o cálculo do erro padrão das estimativas.

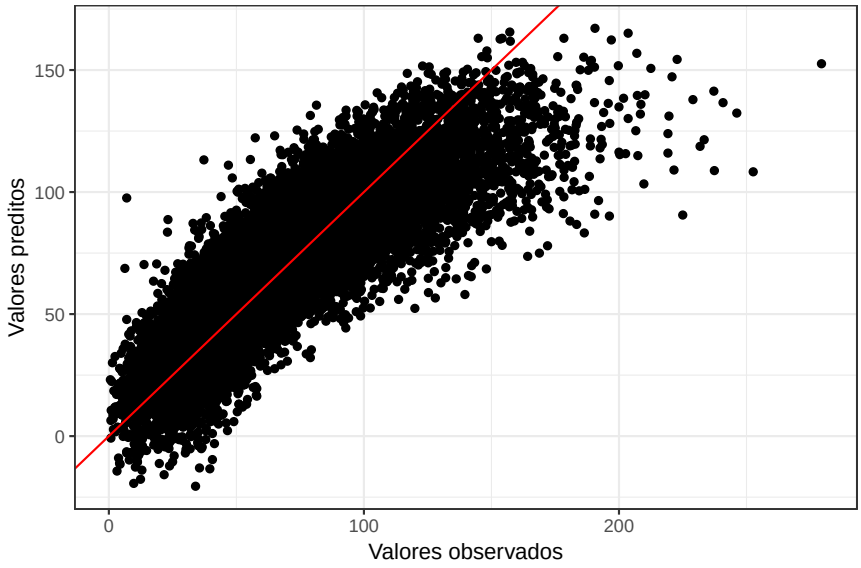

Figura 5.9: *Valores da concentração de ozônio preditos pelo modelo de regressão linear ajustado por Salvo et al. (2017) contra os valores observados.*

de ozônio. Como medida de qualidade de ajuste, os autores reportaram a proporção da variância da concentração de ozônio explicada pelo modelo (R^2): 70.65%.

Adicionalmente, calculamos o erro de teste do modelo, usando validação cruzada (*5-fold*) e o RMSE como métrica. O valor encontrado foi 19.74. Utilizamos também o valor da estatística do teste t de cada coeficiente como medida de importância dos preditores. Os cinco preditores mais importantes foram: temperatura, velocidade do vento, radiação, umidade e variável indicadora para a estação São Caetano do Sul. A variável correspondente à proporção estimada de carros a gasolina foi considerada apenas a décima quinta mais importante.

Na Figura 5.9, apresentamos o gráfico dos valores preditos pelo modelo contra os valores observados. Os pontos acima da reta vermelha representam as observações superestimadas pelo modelo, isto é, aquelas cujo valor predito foi maior do que o valor observado. Os pontos abaixo da curva representam os valores subestimados. Podemos observar que o modelo subestima valores muito altos da concentração de ozônio, isto é, o modelo não consegue explicar os picos do poluente. Esse comportamento é essencialmente ruim dado o objetivo do estudo, já que gostaríamos de avaliar principalmente os dias nos quais os níveis de ozônio estão altos.

Nas próxima seções, vamos comparar o resultado obtido pelos autores com outras estratégias de análise, visando corrigir o viés do modelo para concentrações altas de ozônio e obter um ajuste mais preciso de forma geral. Inicialmente, vamos utilizar as mesmas variáveis para ajustar explorar outros modelos, mais flexíveis que o modelo de regressão linear. Em seguida, vamos testar novos preditores e formas de dimensionar os dados.

5.5 Ajustando outros modelos

O modelo de regressão linear, embora muito utilizado pela sua simplicidade e interpretabilidade, pode ser muito restritivo, já que faz muitas suposições sobre a geração do fenômeno sob estudo. Para avaliar se modelos mais flexíveis são mais adequados para descrever a relação entre a concentração de ozônio e a proporção estimada de carros rodando a gasolina, vamos ajustar modelos aditivos

generalizados, o modelo de regressão linear com regularização e uma *randon forest*. A variável resposta e os preditores serão os mesmos utilizados por Salvo *et al.* (2017). As performances serão comparadas a partir do erro de teste, calculado usando validação cruzada 5-*fold* e o RMSE. Também discutiremos como interpretar cada modelo, e se as interpretações obtidas por ele são suficientes para responder a pergunta de interesse.

5.5.1 Modelos aditivos generalizados

Conforme a especificação em (3.10), ajustamos três modelos aditivos generalizados, um considerando a distribuição Normal, o segundo com distribuição Gama e o terceiro com distribuição Normal Inversa. As funções não-lineares foram atribuídas a todos os preditores quantitativos: proporção estimada de carros rodando a gasolina, temperatura, radiação, umidade, velocidade do vento e tendência. Os demais preditores entraram no modelo de forma linear. *Splines* suavizados foram utilizados na estimação das funções e o grau de suavização foi escolhida automaticamente por meio de validação cruzada. Nos três modelos, a proporção estimada de carros a gasolina foi considerada estatisticamente significante para explicar a concentração de ozônio. A performance de cada modelo está descrita na Tabela 5.2.

Tabela 5.2: *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[6]. Na Figura 5.10, apresentamos esse gráfico para o preditor referente à proporção estimada de carros a gasolina utilizando o modelo com distribuição Normal. O gráfico indica que conforme a proporção de carros a gasolina aumenta, a concentração de ozônio tende a diminuir, mesma conclusão encontrada no modelo de regressão linear.

A Figura 5.11 apresenta o gráfico dos valores preditos contra os valores observados para o modelo

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

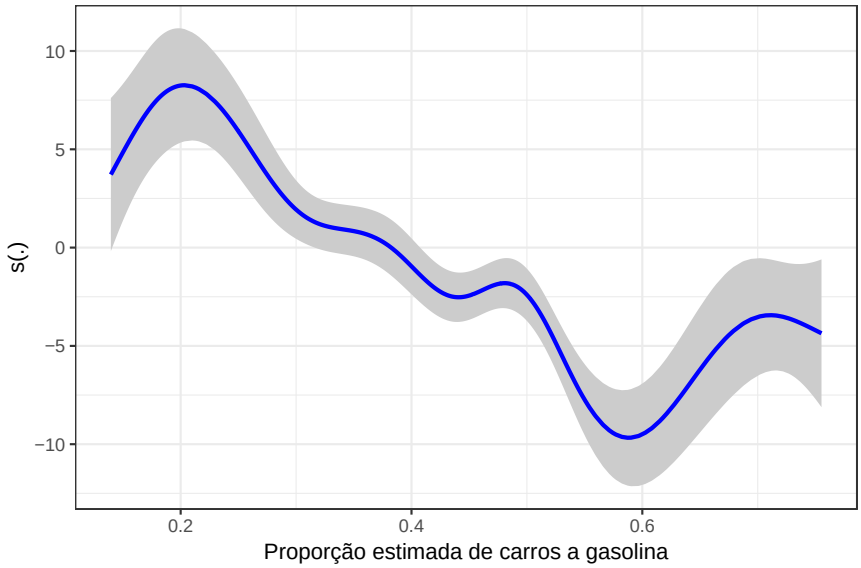

Figura 5.10: *Função não-linear estimada pelo modelo aditivo generalizado com distribuição Normal para a proporção estimada de carros rodando a gasolina. A área cinza em volta da curva representa o intervalo de confiança com 2 erros-padrão para cima e para baixo.*

com distribuição Normal (a) e para o modelo com distribuição Gama (b). O modelo com distribuição Normal possui o mesmo do modelo de regressão linear: ele subestima valores altos da concentração de ozônio. Já o modelo com distribuição Gama corrige em parte esse comportamento. De maneira geral, ambos os modelos tendem a errar mais conforme o valor da concentração de ozônio aumenta.

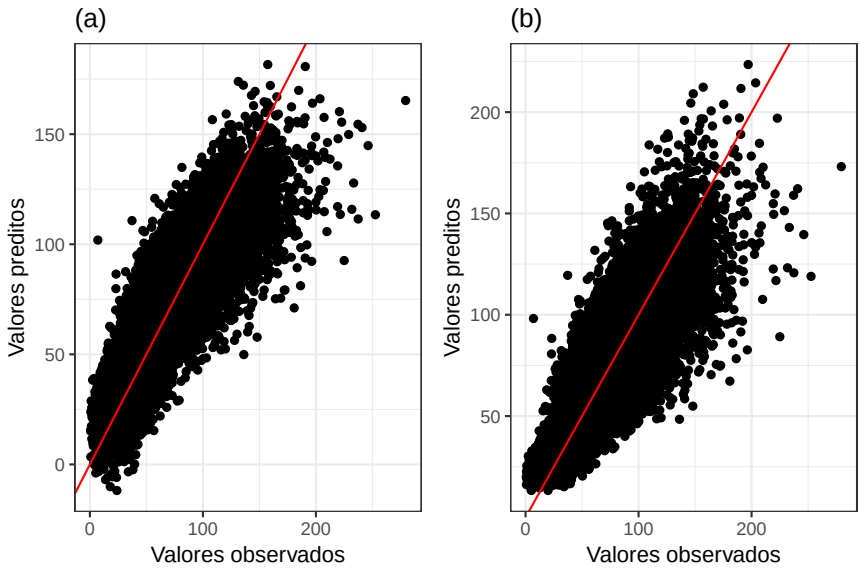

Figura 5.11: *Valores da concentração de ozônio preditos pelo modelo com distribuição Normal (a) e pelo modelo com distribuição Gama (b) contra os valores observados.*

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 5.12, apresentamos o resultado para a função estimada da variável referente à proporção estimada de carros a gasolina. As curvas cinzas são as 200 funções estimadas, uma para cada amostra de *bootstrapping*, e representam a variabilidade da função apresentada na

Figura 5.10. 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 5.1).

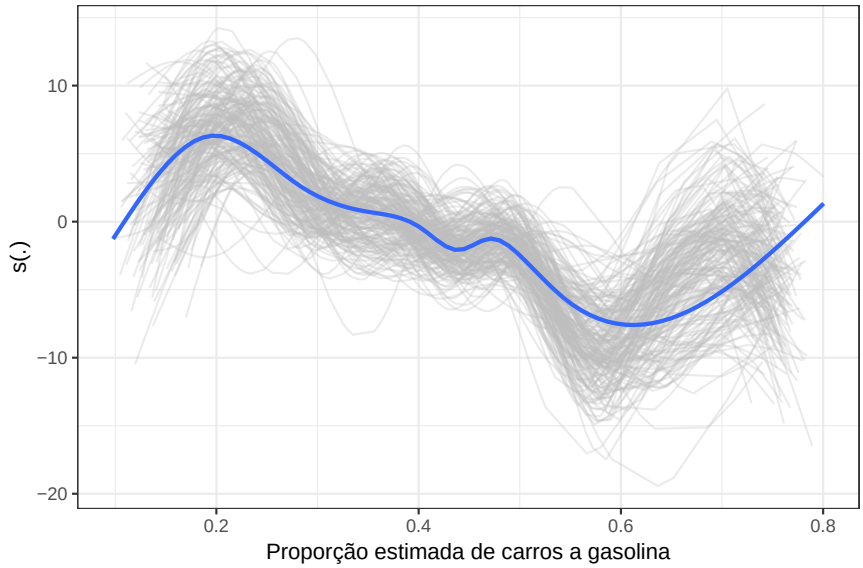

Figura 5.12: *Em cinza, as funções estimadas da variável referente à proporção estimada de carros a gasolina para cada uma das 200 amostras de bootstrapping. Em azul, a curva suavizada por splines cúbicos.*

Quando comparamos os resultados do modelo de regressão linear com os dos modelos aditivos, observamos que não houve ganho de precisão com o aumento da complexidade do modelo. Portanto o modelo linear, por simplicidade, é preferível.

5.5.2 LASSO e regressão ridge

Como apresentado na Tabela 5.1, o modelo de regressão linear ajustado por Salvo *et al.* (2017) tem 96 parâmetros. Podemos tentar diminuir esse número removendo as variáveis menos importantes do modelo. Como discutido na Seção 4.5, podemos utilizar regularização para encolher os coeficientes dos preditores menos importantes e fazer seleção de variáveis.

Com esse objetivo, ajustamos um modelo de regressão linear utilizando o LASSO como técnica de regularização. No entanto, o valor do hiperparâmetro de penalização escolhido por validação cruzada foi igual a 0, indicando que a estimação sem penalização produz o modelo com melhor relação entre viés e variância.

Também testamos a regressão *ridge*, para avaliar se apenas encolher os coeficientes na direção do zero diminuiria a variância do modelo, mas novamente o modelo completo foi selecionado como o melhor.

5.5.3 Random Forest

Em busca de resultados mais precisos, ajustamos também uma *random forest* aos dados. Os resultados estão resumidos na Tabela 5.3.

Tabela 5.3: *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

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 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.

A Figura 5.13 apresenta o gráfico dos valores preditos contra os valores observados para a *random forest*. Fica nítido, ao compararmos com os outros modelos, a redução da diferença entre os valores preditos e observados. No entanto, ainda observamos que os valores altos da concentração de ozônio tendem a ser subestimadas pelo modelo.

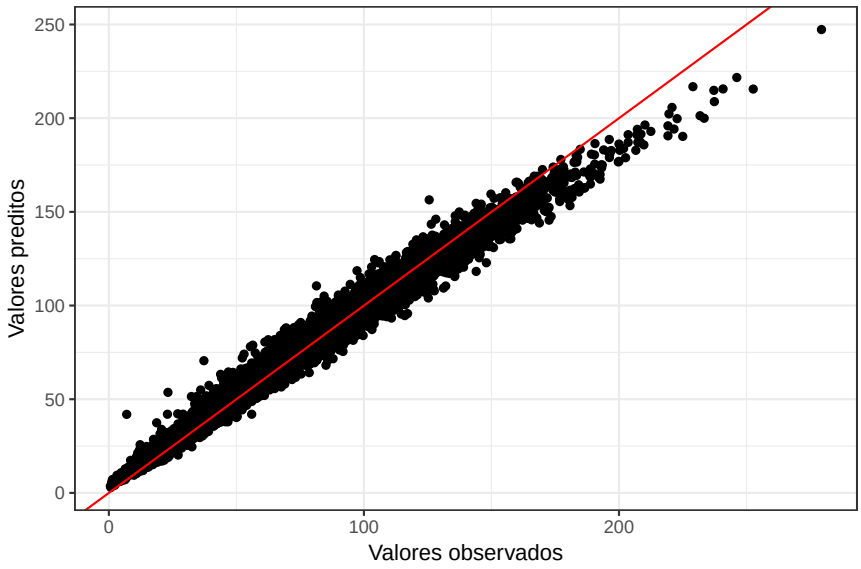

Figura 5.13: *Valores da concentração de ozônio preditos pelo modelo random forest contra os valores observados.*

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 4.7.2. 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 5.14 sugere que, para os dias com maiores médias de ozônio, o efeito protetor da proporção estimada de carros a 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.

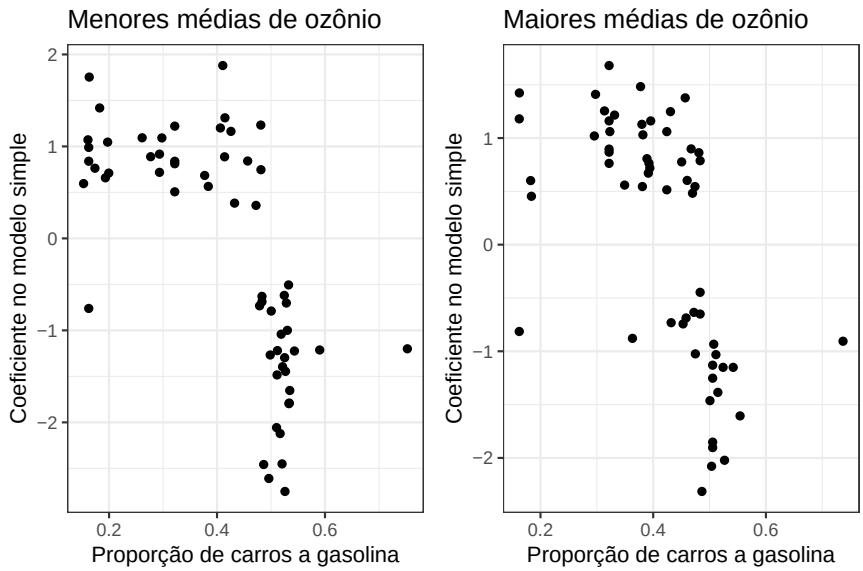

Figura 5.14: *Proporção estimada de carros gasolinas contra o coeficiente estimado para esse preditor no modelo simples (regressão ridge). À esquerda, utilizamos os 100 dias com as menores médias de ozônio e, à direita, os 100 dias com as maiores médias.*

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 5.15, 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.

5.6 Transformando a variável resposta

Vimos na última seção que os modelos ajustados subestimam a concentração de ozônio quando tentamos predizer valores muito altos do poluente. A nossa suspeita é que existe alguma variável importante para explicar os valores esses níveis elevados de ozônio que não foi inserida na análise. No entanto, existem situações em que esse problema de ajuste se deve pelas restrições impostas pelo modelo.

No caso do modelo de regressão linear, supomos que a associação entre a concentração de ozônio e cada um dos preditores era linear, o que pode não ser razoável, principalmente quando a variável resposta tem distribuição assimétrica (Figura 5.16).

Como tentativa de melhorar o ajuste do modelo ajustado por Salvo *et al.* (2017), podemos aplicar alguma transformação à concentração de ozônio na tentativa de reduzir a assimetria da variável. Na Tabela 5.4, apresentamos os resultados do modelo de regressão linear ajustado com as

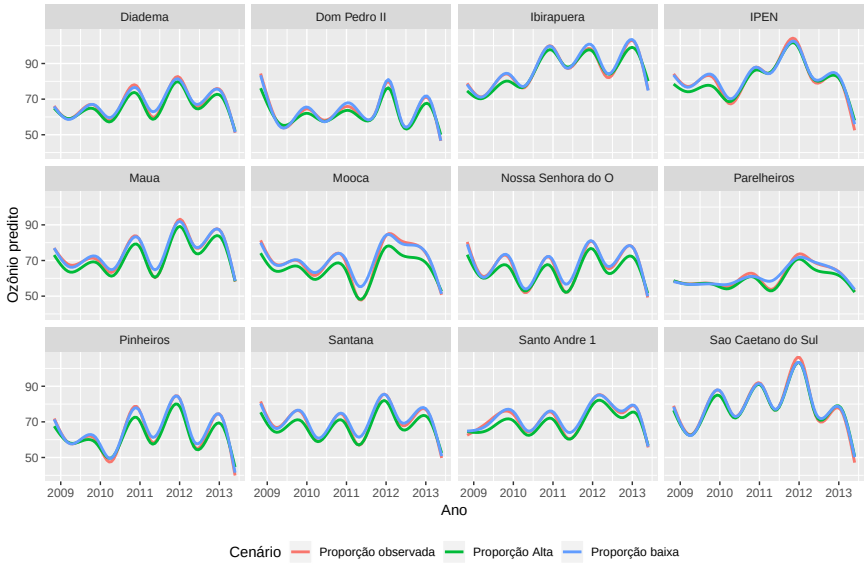

Figura 5.15: *Curvas suavizadas do ozônio predito para os cenários observado, com alta proporção de carros a gasolina (70% durante todo o período) e baixa proporção de carros a gasolina (20% durante todo o período).*

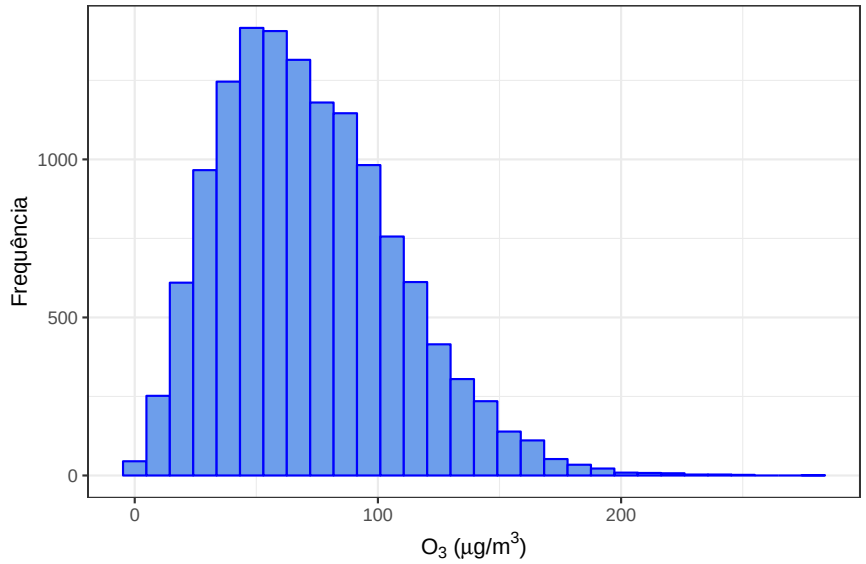

Figura 5.16: *Distribuição da concentração de ozônio na amostra considerada por Salvo et al. (2017).*

transformações *log* e Box-Cox (com lambda = 0.51). Com a ajuda da Figura 5.17, observamos que, para o modelo de regressão linear, a transformação logarítmica melhora o ajuste dos valores mais elevados, mas causa um maior viés nos valores pequenos. Já a transformação de Box-Cox melhora o ajuste tanto dos valores baixos quanto dos altos, diminuindo o RMSE do modelo e aumento a porcentagem da variância explicada.

Como a *random forest* não faz restrições sobre a distribuição da variável resposta, o ganho encontrado ao transformamos a variável foi muito pequeno.

Tabela 5.4: *Resultado dos modelos ajustados com a variável resposta transformada.*

Modelo	Transformação	RMSE	% var. explicada	Variáveis mais importantes
Regressão linear	log	21.18	71.31	Temperatura, radiação, vento, umidade e var. indicadora da estação São Caetano
Regressão linear	Box-Cox	19.48	74.02	Temperatura, radiação, vento, umidade e var. indicadora da estação São Caetano
Random Forest	Box-Cox	14.04	86.87	Temperatura, umidade, radiação, vento e tendência

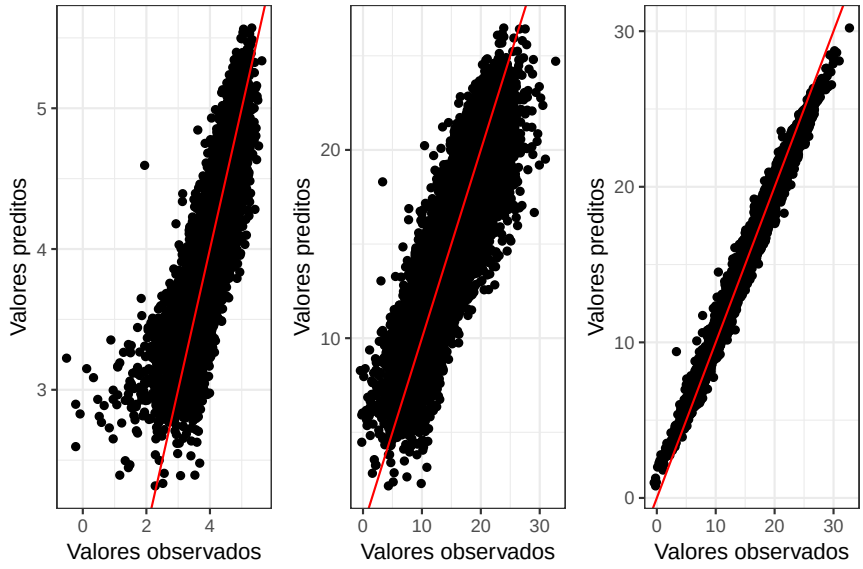

Figura 5.17: *Gráficos dos valores da concentração de ozônio preditos pelos modelos contra os valores observados. No painel da esquerda, modelo de regressão linear com transformação log. No painel do meio, modelo de regressão linear com transformação Box-Cox. No painel da direita, random forest com transformação Box-Cox.*

5.7 Ajustando a máxima diária

Para verificar se os resultados são flexíveis quanto à métrica escolhida para agregar os dados, vamos ajustar o modelo de regressão linear e a *random forest* agora para a máxima diária da concentração de ozônio.

Na Tabela 5.5, apresentamos os resultados dos ajustes. Observamos uma queda considerável na performance dos modelos em relação aos resultados para a média diária, mostrando que, ou os preditores considerados não explicam muito bem a variabilidade da máxima diária, ou precisamos encontrar um modelo mais adequado para essa nova variável.

Também observamos pela Figura 5.18 que os modelos para essa nova variável também subestima valores altos da máxima diária.

Apesar da redução na precisão dos modelos, os resultados obtidos apontam na mesma direção: a proporção de ozônio é estatisticamente significante para explicar variações na concentração de ozônio e mais carros a gasolina tende a diminuir os níveis do poluente.

Tabela 5.5: *Resultado dos modelos ajustados com a variável resposta transformada.*

Modelo	RMSE	% var. explicada	Variáveis mais importantes
Regressão linear	28.00	61.01	Vento, radiação, temperatura e variáveis indicadoras das estações Ibirapuera e São Caetano
Random Forest	20.45	79.86	Temperatura, radiação, umidade, vento e tendência

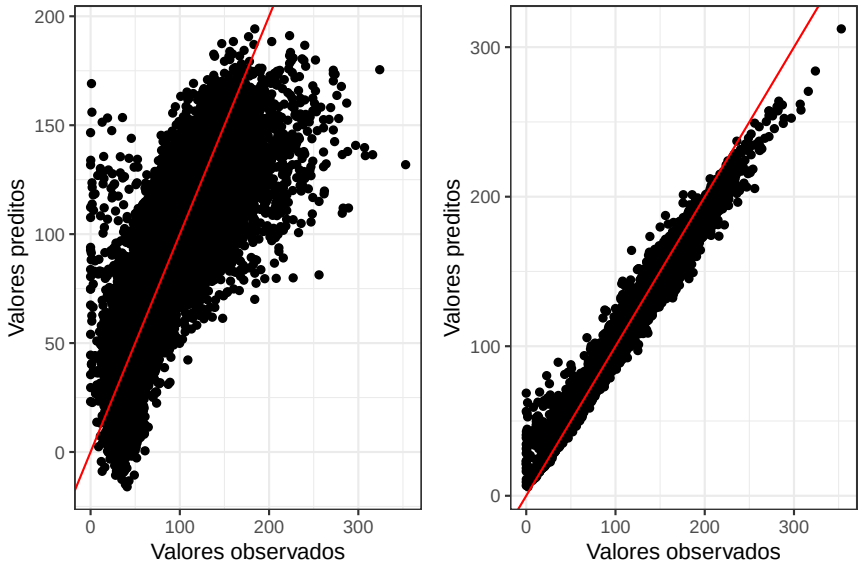

Figura 5.18: *Gráficos dos valores da máxima diária da concentração de ozônio preditos pelos modelos contra os valores observados. No painel da esquerda, modelo de regressão linear e, no painel do meio, a random forest.*

5.8 Ajustando cada estação separadamente

A estratégia adotada por Salvo *et al.* (2017) e por nós até aqui foi juntar as informações de todas as estações de monitoramento em uma única amostra, analisando os dados conjuntamente e incluindo variáveis indicadoras para controlar o efeito de cada estação.

O mecanismo gerador do ozônio, assim como qualquer poluente, é altamente dependente da química atmosférica local. Para avaliar se os resultados encontrados são robustos em relação à posição geográfica de cada estação, podemos refazer a análise ajustando um modelo para cada uma das estações.

Na Tabela 5.6, apresentamos os resultados do modelo de regressão linear para cada uma das 12 estações medidoras de ozônio. O modelo ajustado foi o mesmo considerado pelos autores, com exceção das variáveis relacionadas à estação. Analisando as estimativas dos coeficientes da proporção de carros a gasolina, observamos que essa variável é positivamente associada com a concentração de ozônio nas estações IPEN e São Caetano do Sul[7], contrariando a conclusão para as outras estações e para os resultados com a amostra agregada.

Em relação à performance dos modelos, o RMSE para cada estação não difere muito do encontrado para o modelo global. A proporção da variabilidade explicada só está muito abaixo para a

[7] Para a estação IPEN, o coeficiente é marginalmente significativo significativo.

estação Parelheiros. De uma forma geral, as variáveis meteorológicas são as mais importantes para explicar a concentração de ozônio, sendo que a temperatura ficou entre as cinco primeiras em todos os modelos.

Essa diferença levanta algumas questões que devem ser estudadas mais profundamente:

- O modelo de regressão linear é o melhor para explicar a associação entre as variáveis?
- Existem variáveis importantes para explicar a concentração de ozônio que não foram incluídas no estudo?
- O efeito da proporção de carros rodando a gasolina na concentração de ozônio vai depender da região estudada, mesmo quando analisamos diferentes áreas de uma mesma cidade?

Pela análise realizada aqui, vimos que as conclusões do estudo parecem ser robustas em relação ao modelo adotado. Discutiremos melhor os demais pontos nos comentários da próxima seção.

5.9 Comentários

As análises realizadas neste capítulo indicaram uma associação negativa entre a proporção de carros rodando a gasolina e a concentração de ozônio na região metropolitana de São Paulo. Mostramos que os resultados são robustos à especificação do modelo, mas podem variar a depender das características da química atmosférica local.

Embora essa associação pareça contra-intuitiva, dado o senso comum de que o etanol é uma opção menos poluente aos combustíveis fósseis, ela pode ser explicada se estudarmos com mais cuidado a formação do ozônio troposférico. Dentre outros fatores, o processo gerador desse poluente depende da presença de NO_2, gerado principalmente pela queima de gasolina, e VOCs (compostos orgânicos voláteis), gerado em maior quantidade por veículos a etanol. Como explicado com mais detalhes em Madronich (2014), ambientes limitados em NO_2 tendem a gerar mais ozônio quando mais NO_2 é lançado no atmosfera e ambientes limitados em VOCs tendem a gerar mais ozônio quando mais VOCs é lançado. Os resultados encontrados sugerem que a atmosfera na cidade de São Paulo é, em grande parte, limitada em VOCs e, portanto, a formação de ozônio depende de mais veículos rodando a etanol. Além disso, a maior queima de gasolina (e diesel) pela manhã gera mais NO, que reage com o ozônio durante a tarde, diminuindo seus níveis. Por isso verificamos uma relação negativa entre congestionamento e concentração do poluente.

Essa interpretação também pode explicar os resultados diferentes encontrados para as estações IPEN e São Caetano do Sul, sendo preciso estudar melhor a química atmosférica da região onde as estações estão para confirmar a existência de ambientes limitados em NO_2.

Um outro ponto importante a ser considerado é a relevância prática dos resultados. Usando a estimativa encontrada por Salvo *et al.* (2017), quando a proporção de carros a gasolina sobe de 30% para 80%, mantendo todas as outras variáveis constantes, a concentração diária média de ozônio tende a diminuir 8.3 $\mu g/m^3$. Considerando a concentração média de ozônio em toda a amostra, uma redução média de 72.2 $\mu g/m^3$ para 63.9 $\mu g/m^3$ pode não ter relevante prática para a criação de políticas públicas para a redução de emissões de etanol.

Para avaliar o real impacto da redução do ozônio, podemos estudar o efeito desse poluente no número de casos de doenças e mortes relacionadas com a poluição. Nesse sentido, apresentaremos no próximo capítulo algumas análises associando dados de poluição com dados epidemiológicos.

Tabela 5.6: *Resultados dos modelos para cada estação. A estimativa apresentada na segunda coluna se refere ao coeficiente da proporção de carros a gasolina rodando na cidade.*

Estação	Estimativa (erro-padrão)	RMSE	% var. explicada	Variáveis mais importantes
Diadema	-15.46 (5.01)	17.05	76.87	Vento, temperatura, radiação, variável ind. para a abertura do anel viário e umidade
Dom Pedro II	-18.78 (6.35)	18.90	69.81	Radiação, vento, temperatura, umidade e variável ind. para a semana 40
Ibirapuera	-5.84 (6.5)	20.24	70.88	Temperatura, vento, radiação, variável ind. para a abertura do anel viário e umidade
IPEN	12.7 (7.2)	22.11	66.94	Radiação, temperatura, vento, variável ind. para a abertura do anel viário e tendência
Mauá	-28.09 (6.17)	20.76	67.35	Temperatura, vento, variável ind. para a abertura do anel viário, umidade, radiação
Moóca	-60.02 (7.77)	21.35	63.05	Radiação, proporção de carros a gasolina, vento, temperatura e umidade
Nossa Senhora do Ó	-32.6 (4.53)	17.19	74.59	Temperatura, radiação, vento, proporção de carros a gasolina, variável ind. para a semana 45
Parelheiros	-24.61 (5.57)	18.28	59.64	Vento, variável ind. para inversão térmica, umidade, radiação, temperatura
Pinheiros	-24.41 (5.78)	19.24	65.96	radiação, variável ind. para a abertura do anel viário, vento, umidade, temperatura
Santana	-21.91 (5.91)	18.31	72.70	Temperatura, vento, radiação, umidade, variável ind. para a semana 44
Santo André	-23.1 (6.45)	20.20	67.52	Temperatura, vento, radiação, umidade, proporção de carros a gasolina
São Caetano do Sul	38.37 (7.53)	21.17	70.26	Vento, tendência, radiação, variável ind. para a abertura do anel viário, temperatura

Capítulo 6

Poluição e saúde pública

Embora o estudo da poluição do ar só ter passado a ser tratada como ciência no século 20, relatos milenares de problemas de saúde e ambientais envolvendo a queima de madeira e o derretimento de cobre foram encontrados em cidades como a Grécia antiga e a Roma antiga (Jacobson, 2002). Nos países que formam o Reino Unido, há registros descrevendo as consequências da queima de madeira, carvão e o derretimento de metais ao longo de toda a Idade Média. A criação das máquinas a vapor no século XVII e XVIII e a utilização de combustíveis fósseis iniciada no século XIX pioraram os eventos de poluição do ar nos países industrializados, exigindo a criação de regulamentações e órgãos de controle e monitoramento (Jacobson, 2002).

No último capítulo, discutimos estudos que associavam os índices de poluição ao uso de combustíveis. Agora, o foco será o impacto da poluição do ar na saúde pública. A literatura contemporânea sobre o tema é vasta. Schwartz e Dockery (1992), por exemplo, concluíram que a concentração de partículas suspensas no ar estava positivamente associada com a mortalidade no dia seguinte em Steubenville, Ohio. Saldiva *et al.* (1995) encontraram associação positiva entre a mortalidade diária em idosos (com idade maior que 65 anos) e a concentração de PM10. Os autores não concluíram não existir um nível seguro para a concentração do poluente no cenário estudado. Peters *et al.* (2000) estudaram a chance de intervenções de desfibriladores cardiovasculares implantados em pacientes com histórico alto de arritmia. A partir dos resultados de um modelo logístico, eles concluíram que havia associação positiva entre o aumento de óxidos de nitrogênio e o número de arritmias que geravam intervenções. Hoek *et al.* (2002) acompanharam uma coorte de 5000 holandeses entre 55 e 59 para investigar a associação entre exposição a material particulado e morte por doenças cardiopulmonares. Os autores concluíram que o risco de morte estava associado com os níveis atmosféricos do poluente e, mais consistentemente, com viver perto de vias de tráfego intenso. Utilizando uma função de impacto na saúde, Fann *et al.* (2012) estimaram que 80 mil mortes prematuras seriam evitadas se os níveis de PM2.5 nos Estados Unidos fossem reduzidos em 5 $\mu g/m^3$ e que, em 2005, os níveis de PM2.5 causaram cerca de 130 mil mortes prematuras em pessoas com mais de 29 anos de idade. Cox (2012), no entanto, em nota ao editor, afirmou que Fann *et al.* (2012) interpretaram coeficientes de dose-resposta como se eles representassem relações causais entre a concentração de poluentes e o número de mortes. Segundo o autor, a análise realizada por Fann *et al.* (2012) não garante a causalidade proposta em suas conclusões, que deveriam ser interpretadas apenas como associação estatística.

Embora a poluição esteja associada com morbidade e mortalidade em geral, alguns grupos merecem atenção especial. Segundo relatório da OMS (WHO, 2004), crianças (até 5 anos) são

mais suscetíveis aos efeitos adversos da poluição pois, entre outros fatores, elas respiram uma maior quantidade de ar proporcionalmente ao seu peso do que adultos, costumam praticar mais atividades físicas ao ar livre e possuem sistema imunológico pouco desenvolvido. Idosos compreendem outro grupo de risco, sendo altamente vulneráveis a casos extremos de poluição devido a maior incidência de doenças pré-existentes e sistema imunológico vulnerável.

Neste capítulo, utilizaremos dois problemas como exemplo. Primeiro, vamos expandir a análise do capítulo anterior e avaliar se a proporção estimada de carros rodando a gasolina está associada com a mortalidade na cidade de São Paulo, com foco em crianças e idosos. Em seguida, avaliaremos modelos para estudar a associação também da mortalidade na cidade de São Paulo nesses grupos com as médias diárias de material particulado.

6.1 Etanol, ozônio e mortalidade

6.2 Material particulado e mortalidade

Capítulo 7

Obtendo dados de poluição

> The fact that data science exists as a field
> is a colossal failure of statistics.
> To me, what I do is what statistics is all about.
> It is gaining insight from data using modelling and visualization.
> Data munging and manipulation is hard
> and statistics has just said that's not our domain.
> — Hadley Wickham

Uma etapa crucial da análise de dados, muitas vezes menosprezada pelos estatísticos como não sendo parte da Estatística, é a coleta e a estruturação de dados. De uma maneira geral, a obtenção de dados para análise requer a realização de experimentos, medições ou aplicação de questionários, e a sua estruturação se refere à transferência dos registros obtidos na coleta para uma base de dados retangular[1]. Às vezes, no entanto, os dados que queremos analisar já foram coletados por terceiros e estão disponíveis (mas nem sempre acessíveis) na internet.

Dados de poluição, pela sua relevância para a saúde pública, costumam ser medidos e disponibilizados gratuitamente por órgãos públicos. Embora dificilmente haja interesse político para dificultar o acesso desses dados, como muitas vezes acontece com dados públicos governamentais e de tribunais, o acesso a eles nem sempre é construído maneira ótima e raramente a base se encontra estruturada para análise. A principal causa disso é falta de comunicação entre o responsável pela disponibilização dos dados e a pessoa que vai de fato analisá-los.

Com o aumento da disponibilização de dados na internet ao lado da dificuldade de acesso e estruturação, um *framework* de coleta de dados conhecido *web scraping* (ou raspagem de dados web) vem se tornando cada vez mais popular. Seu objetivo é criar rotinas computacionais para baixar dados de páginas e sistemas na internet de forma automática e estruturada. Embora essas rotinas exijam conhecimento de programação web, elas podem ser realizadas no mesmo ambiente da análise de dados quando utilizamos linguagens como o R ou o Python.

Nesta seção, discutiremos os conceitos básicos de web scraping e apresentaremos algumas formas de se obter dados de poluição do ar, tanto no Brasil quanto em outros lugares do mundo.

[1] Em que cada linha representa uma observação (unidade amostral) e cada coluna representa uma variável.

7.1 Web scraping

7.2 Dados no Brasil

7.3 Dados nos EUA

7.4 Dados na Europa

Capítulo 8

Discussão

Referências Bibliográficas

Achen(2005) Christopher H. Achen. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, 2: 327–339. Citado na pág. 30

Barros *et al.***(2009)** Michelli Barros, Gilberto A. Paula e Victor Leiva. An r implementation for generalized birnbaum-saunders distributions. *Computational Statistics and Data Analysis*, 53(4): 1511–1528. Citado na pág. 36

Beer *et al.***(2011)** Tom Beer, John Carras, David Worth, Nick Coplin, Peter K. Campbell, Bin Jalaludin, Dennys Angove, Merched Azzi, Steve Brown, Ian Campbell, Martin Cope, Owen Farrell, Ian Galbally, Stephen Haiser, Brendan Halliburton, Robert Hynes, David Jacyna, Melita Keywood, Steven Lavrencic, Sarah Lawson, Sunhee Lee, Imants Liepa, James McGregor, Peter Nancarrow, Michael Patterson, Jennifer Powell, Anne Tibbett, Jason Ward, Stephen White, David Williams e Rosemary Wood. The health impacts of ethanol blend petrol. *Energies*, (4): 352–367. Citado na pág. 2

Belusic *et al.***(2015)** Andreina Belusic, Ivana Herceg-Bulic e Zvjezdana Bencetic 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, 27, 38

Bollerslev(1986) T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *J. Econ.*, 31: 307–327. Citado na pág. 44

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. 41

Breiman(2001) Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–231. Citado na pág. 23

Breiman e Friedman(1985) Leo Breiman e J. H. Friedman. Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association*, 80(391): 580–619. Citado na pág. 38

Cameron e Miller(2015) A. Colin Cameron e Douglas L. Miller. A practitioner's guide to cluster-robust inference. *J. Human Resources*, 50(2): 317–372. Citado na pág. 29

Carslaw *et al.***(2007)** David C. Carslaw, Sean D. Beevers e James E. Tate. Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach. *Atmospheric Environment*, 41: 5289–5299. Citado na pág. 1, 2

Casella e Berger(2001) George Casella e Roger L. Berger. *Statistical Inference*. Duxbury Press; 2nd edition. Citado na pág. 25, 33

Chang *et al.***(2017)** Shih Ying Chang, William Vizuete, Marc Serre, Lakshmi Pradeepa Vennam, Mohammad Omary, Vlad Isakov, Michael Breen e Saravanan 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, 45

Conceição *et al.*(2001a) Gleice M.S. Conceição, Simone G.E.K. Miraglia, Humberto S. Kishi, Paulo Hilário Nascimento Saldiva e Julio 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

Conceição *et al.*(2001b) Gleice M.S. Conceição, Paulo Hilário Nascimento Saldiva e Julio 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, 35, 37

Cox(2012) L. A. Cox. Miscommunicating risk, uncertainty, and causation: Fine particulate air pollution and mortality risk as an example. *Risk Analysis*, 32(5). Citado na pág. 81

Demidenko(2013) E. Demidenko. *Mixed models: theory and applications with R.* Wiley, New York. Citado na pág. 29

Dobson(1990) A. J. Dobson. *An introduction to generalized linear models.* Chapman and Hall, New York. Citado na pág. 34

Dordonnat *et al.*(2008) V. Dordonnat, S. J. Koopman, M. Ooms, A. Dessertaine e J. Collet. An hourly periodic state space model for modelling french national electricity load. *International Institute of Forecasters*, 24: 566–587. Citado na pág. 45

Eckner(2018) Andreas Eckner. A framework for the analysis of unevenly spaced time series data. *Journal of the American Statistical Association*. URL http://eckner.com/papers/unevenly_spaced_time_series_analysis.pdf. Citado na pág. 5

Engle(1982) R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50: 987–1007. Citado na pág. 43

European Commission(1999) European Commission. EU focus on clean air. *Office for Official Publications of the European Communities*. Citado na pág. 1

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. 61

Fann *et al.*(2012) Neal Fann, A. D. Lamson, S. C. Anenberg, K Wesson, D. Risley e B. J. Hubbel. Estimating the national public health burden associated with exposure to ambient PM2.5 and ozone. *Risk Analysis*, 32: 81–95. Citado na pág. 81

Gower(1971) C. John Gower. A general coefficient of similarity and some of its properties. *Biometrics*. Citado na pág. 60

Hastie e Tibshirani(1990) Trevor Hastie e Robert Tibshirani. *Generalized additive models.* London:Chapman & Hall. Citado na pág. 3, 4, 38, 57

Hastie *et al.*(2008) Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer. Citado na pág. 3, 23, 25, 32, 38, 40, 41, 57

Hoek *et al.*(2002) Gerard Hoek, Bert Brunekreef, Sandra Goldbohm, Paul Fischer e Piet A van den Brandt. Association between mortality and indicators of traffic-related air pollution in the netherlands: a cohort study. *The Lancet*, 360: 1203:1209. Citado na pág. 81

Jacobson(2002) Mark Z. Jacobson. *Atmospheric Pollution History, Science, and Regulation.* Cambridge. Citado na pág. 81

Jacobson(2007) Mark 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. 2, 62

James *et al.*(**2013**) Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani. *An Introduction to Statistical Learning*. Springer Series in Statistics. Springer, New York. Citado na pág. 4, 25, 27, 31, 32, 40, 41, 51, 53, 54, 56, 57

Jasarevic *et al.*(**2014**) Tarik Jasarevic, Glenn Thomas e Nada 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. 1

Javanmard e Montanari(2014) Adel Javanmard e Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *arXiv:1301.4240v3 [stat.ME]*. Citado na pág. 57

Kalman(1960) R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME J. of Basic Eng.*, (82): 35–45. Citado na pág. 45

Kalman e Bucy(1961) R. E. Kalman e R. S. Bucy. New results in filtering and prediction problems. *Trans. ASME J. of Basic Eng.*, (83): 95–108. Citado na pág. 45

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

Kloog *et al.*(**2012**) Itai Kloog, Francesco Nordio, Brent A. Coull e Joel 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

Leiva(2015) Victor Leiva. *The Birnbaum-Saunders distribution*. Academic Press; 1 edition. Citado na pág. 36

Leiva *et al.*(**2008**) Victor Leiva, Michelli Barros, Gilberto A. Paula. e Antonio Sanhueza. Generalized birnbaum-saunders distribution applied to air pollutant concentration. *Environmetrics*, 19: 235–249. Citado na pág. 36

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

Lockhart *et al.*(**2014**) Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani e Robert Tibshirani. A significance test for the lasso. *Annals of Statistics*, (2). Citado na pág. 57

Madronich(2014) Sacha Madronich. Ethanol and ozone. *Nature Geoscience: news & views*, 7: 395–397. Citado na pág. 79

McCulloch e Searle(2001) C. E. McCulloch e S. R. Searle. *Generalized, linear, and mixed models*. Wiley, New York. Citado na pág. 29

Morettin e Toloi(2004) Pedro A. Morettin e Clelia M.C. Toloi. *Análise de Series Temporais*. ABE - Projeto Fisher e Editora Edgard Blucher, São Paulo. Citado na pág. 16, 42, 43, 44

Mulawa *et al.*(**1997**) Patricia A. Mulawa, Steven H. Cadle, Kenneth Knapp, Roy Zweidinger, Richard Snow, Randy Lucas e Joseph 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. 61

Nelder e Wedderburn(1972) J. A. Nelder e R. W. M. Wedderburn. Generalized linear models. *Stat Soc A*, 135: 370–384. Citado na pág. 3, 34, 55

Nicholson(2001) W. Keith Nicholson. *Elementary Linear Algebra*. McGraw-Hill Ryerson, 2ª edição. Citado na pág. 17

Paula(2013) Gilberto A. Paula. *Modelos de Regressão com apoio computacional*. São Paulo. URL https://www.ime.usp.br/~giapaula/texto_2013.pdf. Citado na pág. 29, 34, 35, 36

Pereira *et al.*(**2004**) Pedro Afonso Pereira, Leilane Maria B. Santos, Eliane Teixeira Sousa e Jailson 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. 61

Peters *et al.*(**2000**) Annette Peters, Emerson Liu, Richard L. Verrier, Joel Schwartz, Diane R. Gold, Murray Mittleman, Jeff Baliff, J. Annie Oh, George Allen, Kevin Monahan e Douglas W. Dockery. Air pollution and incidence of cardiac arrhythmia. *Lippincott Williams & Wilkins*, 11 (1): 11–17. Citado na pág. 81

R Core Team(2016) R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. Citado na pág. 4

Ribeiro *et al.*(**2016**) Marco Tulio Ribeiro, Sameer Singh e Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. *arXiv:1602.04938v3 [cs.LG]*. Citado na pág. 59

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

Saldiva *et al.*(**1995**) Paulo H. N. Saldiva, C. Arden Pope, Joel Schwartz, Douglas W. Dockery, Ana Julia Lichtenfels, Joao Marcos Salge, Ivana Barone e Gyorgy Miklos 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, 24, 37, 81

Salvo e Geiger(2014) Alberto Salvo e Franz 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. xii, 2, 5, 17, 18, 33, 45, 53, 62

Salvo e Wang(2017) Alberto Salvo e Yi Wang. Ethanol-blended gasoline policy and ozone pollution in sao paulo. *JAERE*, 4(3). Citado na pág. 62

Salvo *et al.*(**2017**) Alberto Salvo, Joel Brito, Paulo Artaxo e Franz M. Geiger. Reduced ultrafine particle levels in São Paulo's atmosphere during shifts from gasoline to ethanol use. *Nature Communications*, 8: 1–14. Citado na pág. viii, xiii, xiv, xv, 2, 6, 7, 24, 28, 29, 53, 62, 63, 69, 70, 71, 73, 74, 75, 76, 78, 79

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, 36, 81

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

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

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

Schwartz e Marcus(1990) Joel Schwartz e Allan 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

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

Shumway e Stoffer(2006) Robert H. Shumway e David S. Stoffer. *Time Series Analysis and Its Applications (with R examples)*. Springer Texts in Statistics. Springer, New York, 2ª edição. Citado na pág. 13, 15, 16, 27, 41, 42, 43, 44

Singer *et al.*(2012) J.M. Singer, J.S. Nobre e F.M.M. Rocha. *Análise de Dados Longitudinais (versão parcial preliminar)*. http://www.ime.usp.br/ jmsinger/Textos, São Paulo. Citado na pág. 32

WHO(2004) World Health Organization WHO. Health aspects of air pollution. Relatório Técnico E83080. Citado na pág. 81

Wickham(2010) Hadley Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1): 3–28. Citado na pág. 7

Wickham e Grolemund(2017) Hadley Wickham e Garrett Grolemund. *R for Data Science*. O'Reilly, 1ª edição. Citado na pág. 3, 4, 5

Wilkinson(2005) Leland Wilkinson. *The Grammar of Graphics*. Statistics and Computing. Springer. 2nd edition. Citado na pág. 7

Williams(1987) A. D. Williams. Generalized linear model diagnostic using the deviance and single case deletion. *Applied Statistics*, 36: 181–191. Citado na pág. 36

Wood(2006) Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, 1ª edição. Citado na pág. 36

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. 61

Zannetti(1990) P. Zannetti. *Air pollution modelling: theories, computational methods and available software*. Springer Science+Business Media, LLC, New York. Citado na pág. 1, 45