

Data science para estudos 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

Data science para estudos 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

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, 2019.

A Estatística é uma ferramenta imprescindível para a aplicação do método científico, estando presente em todos os campos de pesquisa. As metodologias estatísticas usuais estão bem estabelecidas entre os pesquisadores das mais diversas áreas, sendo que a análise de dados em muitos trabalhos costumam ser feita pelos próprios autores. Nos últimos anos, a área conhecida como *data science* vem exigindo de estatísticos e não-estatísticos habilidades que vão muito além de modelagem, começando na obtenção e estruturação das bases de dados e terminando na divulgação dos resultados. Dentro dela, uma abordagem conhecida como *machine learning* reuniu diversas técnicas e estratégias para modelagem preditiva, que, com alguns cuidados, podem ser aplicadas também para inferência. Essas novas visões da Estatística foram pouco absorvidas pela comunidade científica até então, principalmente pela ausência de estatísticos de formação em grande parte dos estudos. Embora pesquisa de base em Probabilidade e Estatística seja importante para o desenvolvimento de novas metodologias, a criação de pontes entre essas disciplinas e suas áreas de aplicação é essencial para o avanço da ciência. O objetivo desta tese é aproximar a ciência de dados, discutindo metodologias novas e usuais, da área de pesquisa em poluição do ar, que, segundo a Organização Mundial da Saúde, é o maior risco ambiental à saúde humana. Para isso, apresentaremos diversas estratégias de análise e as aplicaremos em dados reais de poluição do ar. Os problemas utilizados como exemplo foram o estudo realizado por [Salvo et al. \(2017\)](#), cujo objetivo foi associar a proporção de carros rodando a gasolina com a concentração de ozônio na cidade de São Paulo, e uma extensão desse trabalho, na qual analisamos o efeito do uso de gasolina/etanol na mortalidade de idosos e crianças. Concluímos que suposições como linearidade e aditividade, feitas por alguns modelos usuais, podem ser muito restritivas para problemas essencialmente completos, sendo que o ajuste de diferentes modelos leva a diferentes conclusões, nem sempre sendo fácil identificar qual delas é inadequada.

Palavras-chave: *data science, machine learning, poluição do ar, regressão*

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.

Keywords:

Sumário

Lista de Figuras	xii
Lista de Tabelas	xviii
1 Introdução	1
2 Análise exploratória	7
2.1 Gráficos	8
2.1.1 O gráfico da série	8
2.1.2 Gráficos de dispersão	11
2.1.3 Gráficos de distribuição	13
2.2 Componentes temporais	14
2.2.1 Tendência	15
2.2.2 Sazonalidade	17
2.2.3 Autocorrelação	18
2.2.4 Função de correlação cruzada	20
2.3 Visualizando dados de poluição durante a greve de caminhoneiros	21
3 Estratégias usuais de modelagem	29
3.1 Regressão linear	31
3.1.1 Especificação do modelo	31
3.1.2 Incorporando tendência e sazonalidade	32
3.1.3 Tratando erros correlacionados	34
3.1.4 Contornando a suposição de homoscedasticidade	36
3.1.5 Contornando a suposição de linearidade	37
3.1.6 Contornando a suposição de aditividade	39
3.1.7 Avaliando a qualidade do ajuste	40
3.2 Modelos lineares generalizados	41
3.2.1 Especificação do modelo	41
3.2.2 Modelos para dados positivos assimétricos	42
3.2.3 Modelos para dados de contagem	44
3.3 Modelos aditivos generalizados	45
3.3.1 Especificação do modelo	45
3.3.2 Splines e regressão local	46
3.4 Modelos de séries temporais	48

3.4.1	Modelos autorregressivos (AR)	49
3.4.2	Modelos autorregressivos e de médias móveis (ARMA)	49
3.4.3	Modelos autorregressivos integrados e de médias móveis (ARIMA)	50
3.5	Modelos não-supervisionados	51
3.5.1	Análise de componentes principais	51
3.5.2	Análise Fatorial	52
3.6	Outros modelos	53
3.6.1	Modelos mistos	53
3.6.2	Modelos GARCH	54
3.6.3	Modelos dinâmicos	55
4	Estratégias de machine learning	57
4.1	Sobreajuste e o balanço entre viés e variância	58
4.2	Estimando a performance do modelo	60
4.2.1	Validação cruzada	61
4.2.2	Bootstrapping	63
4.3	Seleção de variáveis	63
4.3.1	Selecionando o melhor subconjunto de preditores	64
4.3.2	Stepwise	64
4.4	Regularização	65
4.5	Quantificando a importância dos preditores	67
4.6	Modelos de árvores	67
4.6.1	Árvores de decisão	68
4.6.2	Random Forests	69
4.6.3	XGBoost	70
4.7	Interpretando modelos caixa-preta	70
4.7.1	Gráfico de dependência parcial	71
4.7.2	Gráfico da esperança condicional individual	71
4.7.3	Gráfico de efeitos locais acumulados	71
4.7.4	LIME	72
4.7.5	Exemplo	73
5	Poluição e uso de combustíveis	77
5.1	Etanol e ozônio	77
5.2	Entendendo o problema	78
5.3	Análise exploratória	80
5.4	A análise conduzida por Salvo <i>et al.</i> (2017)	84
5.5	Ajustando outros modelos	87
5.5.1	Modelos aditivos generalizados	88
5.5.2	Modelo de regressão segmentada	90
5.5.3	Random Forest	92
5.6	XGBoost	93
5.7	Outras estratégias de análise	95
5.7.1	Seleção de variáveis	96

5.7.2	Transformando a variável resposta	96
5.7.3	Ajustando a máxima diária	98
5.7.4	Ajustando cada estação separadamente	99
5.8	Comentários	99
6	Poluição e saúde pública	103
6.1	Uso de etanol e mortalidade	104
6.1.1	Modelo linear Poisson	105
6.1.2	Modelo aditivo Poisson	106
6.1.3	Random forest	107
6.1.4	Análises complementares	108
6.2	Concentração de ozônio e mortalidade	110
7	Obtendo dados de poluição	113
7.1	Web scraping	114
7.2	Dados no Brasil	116
7.3	Dados em outros países	117
8	Discussão	119
	Referências Bibliográficas	121

Listas de Figuras

1.1	Esquematização do ciclo da ciência de dados.	4
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 <i>splines</i> cúbicos.	9
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.	10
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 <i>splines</i> cúbicos.	10
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 <i>splines</i> cúbicos.	11
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 <i>splines</i> cúbicos.	12
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.	12
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.	13
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.	14
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.	15
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.	16
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.	17

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.	18
2.13 Função de autocorrelação 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.	19
2.14 Função de autocorrelação 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.	20
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.	21
2.16 Médias horárias por dia da semana durante o período observado dos poluentes considerados na análise.	23
2.17 Série observada para o monóxido de carbono. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.	24
2.18 Série observada para o ozônio. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.	25
2.19 Série observada para o monóxido de nitrogênio. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.	26
2.20 Série observada para o dióxido de nitrogênio. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.	27
2.21 Série observada para o material particulado (PM10). O intervalo entre as retas pontilhadas corresponde ao período de paralisações.	28
3.1 Esquematização do mecanismo gerador dos dados.	29
3.2 Exemplos de séries com tendência linear e quadrática, ambas positivas.	33
3.3 Exemplos de uma série com tendência não-constante.	34
3.4 Comparação entre os gráficos dos resíduos de um modelo linear contra o tempo para dados auto-correlacionados e dados não correlacionados.	35
3.5 Gráfico dos resíduos contra os valores preditos. Exemplo de nuvem de pontos em forma de funil, indicando heteroscedasticidade.	36
3.6 A estimativa $\hat{\beta}$ representa a variação em Y quando acrescemos X em uma unidade, não importando o valor de X	37

3.7	Gráfico dos resíduos contra os valores preditos, um exemplo de nuvem de pontos em forma de “U”, indicando não-linearidade.	38
3.8	Exemplo de regressão segmentada. À esquerda, o gráfico de concentrações médias diárias de ozônio pela temperatura média diária. À direita, um modelo de regressão segmentada ajustada aos pontos com um ponto de corte.	39
3.9	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.	43
3.10	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.	43
3.11	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 $x = 1$	47
4.1	Exemplo do <i>trade-off</i> 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.	59
4.2	Esquematização da validação cruzada <i>leave-one-out</i>	62
4.3	Esquematização da validação cruzada <i>k-fold</i> , com $k = 5$	62
4.4	Exemplo de uma árvore de decisão para a concentração de ozônio explicada pela temperatura.	68
4.5	Gráficos de dispersão da média diária de NO_x contra as médias diárias de temperatura (Celsius) e umidade relativa do ar (%).	74
4.6	Gráficos de dispersão da média diária de NO_x contra as médias diárias de temperatura (Celsius) e umidade relativa do ar (%).	74
4.7	Valores preditos do NO_x para cada modelo em função da umidade e da temperatura.	75
4.8	Gráficos de dependência parcial (PDP), esperança condicional individual (ICE) e efeitos locais acumulados (ALE) para a <i>random forest</i> . A curva amarela no ICE representa a média de todas as retas individuais, isto é, o PDP.	76
4.9	Gráfico de dispersão entre as médias diárias da temperatura e umidade.	76
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.	81
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.	82
5.3	Boxplot da proporção estimada de carros a gasolina para cada mês.	82
5.4	Gráficos <i>ridge</i> 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.	83

5.5	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.	83
5.6	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.	84
5.7	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.	85
5.8	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.	85
5.9	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.	86
5.10	Valores da concentração de ozônio preditos pelo modelo de regressão linear ajustado por Salvo <i>et al.</i> (2017) contra os valores observados.	87
5.11	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.	89
5.12	Valores da concentração de ozônio preditos pelo modelo com distribuição Normal (a) e pelo modelo com distribuição Gama (b) contra os valores observados.	90
5.13	Em cinza, as funções estimadas da variável referente à proporção estimada de carros a gasolina para cada uma das 200 amostras de <i>bootstrapping</i> . Em azul, a curva suavizada por <i>splines</i> cúbicos.	91
5.14	Retas de regressão segmentada para a proporção estimada de carros a gasolina. O efeito representa o valor da concentração de ozônio para cada valor da proporção estimada de carros a gasolina se todos os outros preditores tivessem valor igual a 0. Essa medida não tem interpretação prática, mas pode ser utilizada para calcular a variação no ozônio quando variamos a proporção de carros a gasolina.	92
5.15	Valores da concentração de ozônio preditos pelo modelo <i>random forest</i> contra os valores observados.	93
5.16	Gráficos de dependência parcial (PDP) e de efeitos locais acumulados para o modelo <i>random forest</i> .	94
5.17	Gráficos de dependência parcial (PDP) e de efeitos locais acumulados para a proporção estimada de carros a gasolina do modelo <i>XGBoost</i> .	95
5.18	Gráficos efeitos locais acumulados para as variáveis climáticas do modelo <i>XGBoost</i> .	95
5.19	Distribuição da concentração de ozônio na amostra considerada por Salvo <i>et al.</i> (2017).	96

5.20 Gráficos dos valores da concentração de ozônio preditos pelos modelos contra os valores observados. No painel da esquerda, modelo de regressão linear com transformação <i>log</i> . No painel do meio, modelo de regressão linear com transformação Box-Cox. No painel da direita, <i>random forest</i> com transformação Box-Cox.	97
5.21 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 <i>random forest</i>	98
6.1 Séries da mortalidade diária para crianças e idosos.	105
6.2 Gráfico dos valores preditos versus os valores observados para o modelo linear generalizado com distribuição Poisson para cada um dos grupos.	106
6.3 Funções estimadas para a proporção estimada de carros a gasolina pelo modelo aditivo Poisson para cada grupo.	107
6.4 Gráficos de dependência parcial (PDP) e de efeitos acumulados (ALE) da <i>random forest</i> para a variável proporção estimada de carros a gasolina.	108
6.5 Gráfico de correção cruzada da mortalidade diária para idosos contra a proporção estimada de carros a gasolina.	109
6.6 Gráficos de dependência parcial (PDP) e de efeitos acumulados (ALE) da <i>random forest</i> para a variável proporção estimada de carros a gasolina defasada em 3 dias.	110
6.7 Gráfico da função não-linear estimada pelos modelos aditivos generalizados para a proporção estimada de carros a gasolina.	111
6.8 Gráficos de dependência parcial (PDP) e de efeitos acumulados (ALE) da <i>random forest</i> (idosos) para a concentração de ozônio.	111
7.1 O fluxo do web scraping.	114
7.2 Mapa de estações de monitoramento disponíveis na Plataforma de Qualidade do Ar do Instituto de Energia e Meio Ambiente.	116
7.3 Exemplo de visualização do portal AirVisual para a estação Parque Dom Pedro II, em São Paulo.	118

Listas de Tabelas

2.1	Variação da média dos poluentes em cada estação no período de greve em relação à média nos períodos anterior e posterior à greve	23
3.1	Critérios para a escolha da ordem de modelos ARIMA.	51
4.1	Raiz do erro quadrático médio (RMSE) para os modelos polinomiais de grau 1 a 9 ajustados com 10 e 100 observações no exemplo da Figura 4.1.	59
4.2	Modelos de regressão linear que devem ser ajustados para selecionar o melhor subconjunto de variáveis no caso com 3 preditores.	64
5.1	Preditores considerados pelo modelo para a concentração de ozônio ajustado em Salvo <i>et al.</i> (2017).	86
5.2	Resultados dos modelos aditivos generalizados em comparação com o modelo utilizado por Salvo <i>et al.</i> (2017).	88
5.3	Resultado do modelo de regressão segmentada.	91
5.4	Resultado do modelo <i>random forest</i> aplicado aos dados de Salvo <i>et al.</i> (2017). Os hiperparâmetros referentes ao tamanho mínimo de cada nó e o número de preditores sorteados em cada amostra foram definidos por validação cruzada.	92
5.5	Performance do modelo <i>XGBoost</i> aplicado aos dados de Salvo <i>et al.</i> (2017) em comparação com os outros modelos ajustados.	94
5.6	Resultado dos modelos ajustados com a variável resposta transformada.	97
5.7	Resultado dos modelos ajustados com a variável resposta transformada.	98
5.8	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. .	100
6.1	Resultados do modelo Poisson para mortalidade geral em crianças, idosos e na população como um todo.	105
6.2	Resultados do modelo aditivo Poisson para mortalidade geral em crianças, idosos e na população como um todo.	106
6.3	Resultados do modelo da Random forest com proporção estimada de carros a gasolina defasada no tempo. Os valores estão ordenados do menor ao maior RMSE.	109
6.4	Resultados dos modelos ajustados para a mortalidade diária utilizando a concentração de ozônio como preditor.	110

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 cardiovesselares, problemas digestivos e no sistema nervoso, câncer e aumento da mortalidade infantil (European Commission, 1999). No meio ambiente, a poluição do ar é responsável por dois grandes problemas: o efeito estufa, relacionado ao aumento da temperatura média do planeta, e a destruição da camada de ozônio, aumentando a incidência de radiação solar nociva na superfície terrestre.

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 interesses econômicos, mesmo quando a redução a longo prazo nos gastos com tratamentos de saúde poderia gerar números positivos nesse balanço.

Como discutido em Zannetti (1990), os estudos de poluição do ar são de extrema importância pois, a partir deles, podemos

- acompanhar as concentrações dos poluentes ao longo do tempo;
- 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;
- prever e 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;
- desenvolver soluções mais limpas (ou menos poluentes) que ainda sejam economicamente viáveis.

A maioria dos estudos de poluição do ar utiliza modelagem, que pode envolver tanto modelos físicos quanto modelos matemáticos. Os modelos físicos correspondem a experimentos laboratoriais que tentam representar os fenômenos naturais. Já os modelos matemáticos são construções teóricas (analíticas ou numéricas) que visam descrever da forma mais precisa possível o fenômeno natural, sendo possível dividi-los em duas classes: determinísticos e estatísticos. Os modelos determinísticos, geralmente formulados a partir de equações de balanço químico, descrevem matematicamente os processos atmosféricos. Por sua vez, os modelos estatísticos utilizam observações coletadas sobre os agentes (supostamente) envolvidos no processo para inferir relações entre eles e produzir previsões. Nesta tese, abordaremos apenas os modelos estatísticos.

A literatura sobre modelagem em estudos de poluição do ar é vasta. *Katsouyanni et al.* (1996), por exemplo, utilizaram um modelo Poisson autorregressivo para investigar o efeito da poluição do ar na saúde pública a partir da concentração de diversos poluentes. *Carslaw et al.* (2007) usuram modelos aditivos generalizados para modelar 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) utilizaram funções de impacto na saúde para estudar os impactos epidemiológicos e econômicos de 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 modelos mistos para prever concentrações diárias de material particulado na costa leste dos Estados Unidos a partir de medidas de profundidade óptica de aerossóis feitas por satélites.

Embora esses estudos ataquem diferentes problemas utilizando variadas técnicas estatísticos, uma característica presente em todos os estudos citados é a dificuldade inerente da modelagem de dados de poluição do ar. A formação de poluentes envolve reações químicas complexas, que envolvem diversos variáveis climáticas e, nas cidades, é altamente dependente das flutuações do trânsito de veículos causadas por feriados, férias escolares ou eventos esportivos e culturais. Todos esses fatores são difíceis de serem controlados em laboratório e, por isso, as grandes cidades acabam se transformando em laboratórios naturais para estudos de poluição do ar. Comumente túneis servem como "ambientes controlados" para a coleta de dados e estações de monitoramento serem instaladas próximas a grandes vias e zonas industriais. Com a disponibilidade de dados meteorológicos e de tráfego, é possível avaliar grande parte dos fatores que influenciam na formação dos poluentes.

Além do grande número de variáveis a serem consideradas nos estudos de poluição atmosférica, também é razoável imaginar que a relação entre elas não deva ser simples, o que limitaria o uso de modelos muito restritivos. Apesar disso, na literatura, muitos estudos de poluição do ar utilizam metodologias estatísticas usuais, que fazem suposições fortes como linearidade, aditividade e independência¹, nem sempre coerentes com a realidade. Exemplos dessas metodologias são os modelos de regressão linear, utilizados por *Salvo e Geiger* (2014); *Salvo et al.* (2017), os modelos lineares generalizados (*Conceição et al.*, 2001b; *Lin et al.*, 1999; *Saldiva et al.*, 1994, 1995; *Schwartz e Dockery*, 1992b), os modelos aditivos generalizados (*Carslaw et al.*, 2007; *Conceição et al.*, 2001a,b; *Schwartz et al.*, 1996; *Schwartz*, 1994, 1996) e várias de suas generalizações (modelos de regressão segmentada, modelos mistos, modelos autorregressivos, entre outros).

Modelos de regressão linear (*Hastie et al.*, 2008) são muito utilizados para modelagem devido à facilidade de implementação e interpretação de seus coeficientes. Por outro lado, eles supõem

¹ Esses conceitos serão tratados com detalhes no Capítulo 3.

que as observações tenham variância constante e que relação entre as variáveis seja linear, o que pode ser muito restritivo na prática. Os modelos lineares generalizados (Nelder e Wedderburn, 1972) flexibilizam a suposição de homoscedasticidade, permitindo modelar também a dispersão dos dados, mas ainda estão presos à suposição de linearidade. Já os modelos aditivos generalizados (Hastie e Tibshirani, 1990) são uma boa alternativa para modelar relações não-lineares, pois permitem a inclusão de termos não-paramétricos para ajustar funções suavizadas não-lineares da variável resposta em função das variáveis explicativas. Contudo, assim como as outras duas classes, supõem relações aditivas, o que dificulta o ajuste de interações complexas entre os preditores. De uma forma geral, autocorrelação, heteroscedasticidade, superdispersão, tendência, sazonalidade, componentes espaciais 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.

Tomando o estudo conduzido por Salvo e Geiger (2014); Salvo *et al.* (2017) como exemplo, observamos que os autores utilizam um modelo de regressão linear para investigar a associação entre o uso de gasolina/etanol com a formação troposférica de ozônio, supondo que cada variável explicativa está linearmente associada com a concentração de ozônio, o que, na prática, pode não ser verdade. Isso pode levar a conclusões erradas ou superficiais, direcionando o poder público a tomar medidas equivocadas ou insuficientes.

Nos últimos anos, a área popularmente conhecida como *data science* (ou ciência dos dados) vem expandindo as preocupações e habilidades que estatísticos e não estatísticos precisam possuir para analisar dados. O que antes era um trabalho primordialmente de visualização e modelagem, agora agrupa tarefas mais gerais de importação, estruturação e manipulação de bases de dados, assim como etapas de implementação, automatização e divulgação dos resultados. Outra característica dessa área é o menor rigor matemático/probabilístico dado aos problemas, em função de uma abordagem mais prática/computacional. Isso deixa seu conteúdo mais acessível a pesquisadores e profissionais de qualquer área, já que exige um menor conhecimento formal de Estatística. A Figura 1.1 representa uma esquematização do chamado *ciclo da ciência de dados*, cujo objetivo é ressaltar a importância dessas outras etapas da análise que antes eram colocadas em segundo plano pela Estatística. O foco desta tese será a etapa de modelagem, mas também falaremos sobre visualização e importação de dados.

Dentro da etapa de modelagem, a abordagem conhecida como *machine learning* vem ganhando bastante notoriedade por criar estratégias de análise robustas para modelagem preditiva, em especial por utilizarem modelos que não fazem suposições sobre a forma como as variáveis estão relacionadas, permitindo ajustes muito mais precisos, e incluírem conceitos como sobreajuste, validação cruzada e regularização, essenciais no contexto de previsão. Embora diversas dessas técnicas possam ser generalizadas para além da modelagem preditiva, ainda há muita confusão e resistência na sua utilização para inferência, principalmente pela falta de interpretabilidade dos modelos mais famosos.

Em estudos de poluição do ar, podemos encontrar alguns trabalhos mais recentes que aplicaram *machine learning* sobretudo em problemas de previsão (Cortina-Januchs *et al.*, 2015; Elangasinghe *et al.*, 2014; Feng *et al.*, 2015). O uso contudo ainda é insípiente e há bastante espaço para aprimorar a modelagem de estudos inferenciais utilizando essa abordagem. Stingone *et al.* (2017), por exemplo, utilizaram árvores de decisão para associar exposição à poluição do ar com habilidades cognitivas em crianças, nos Estados Unidos, que, apesar de serem altamente interpretáveis, poderiam ter sido

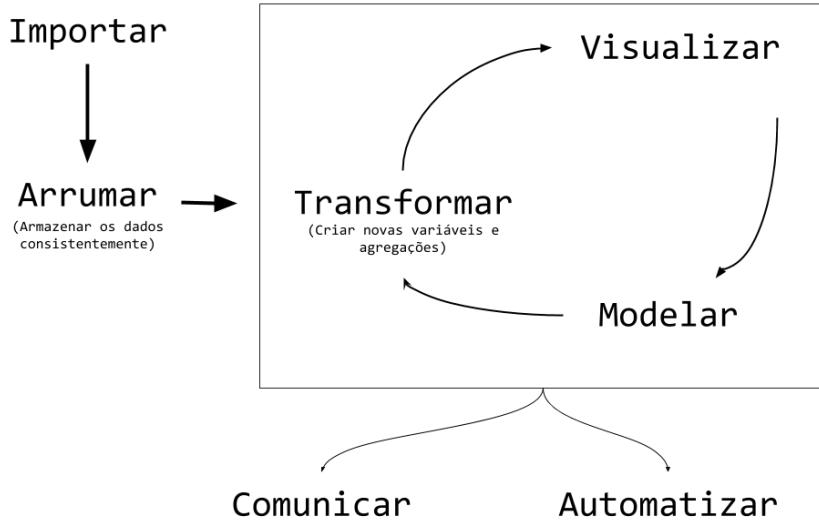


Figura 1.1: Esquematização do ciclo da ciência de dados.

substituídas por modelos mais precisos. Já [Polezer et al. \(2018\)](#) utilizam redes neurais para investigar a associação entre material particulado e doenças respiratórias, mas não utilizaram métodos de interpretação para avaliar o sentido e a força dessa associação.

O objetivo desta tese é criar e discutir estratégias robustas para a análise de dados de poluição do ar usando a roupagem da ciência de dados. Vamos avaliar em quais contextos as metodologias usuais são adequadas e podem ser aplicadas e propondo abordagens com técnicas mais recentes, como o *machine learning*, ainda pouco utilizado em estudos de poluição do ar. Os resultados deste trabalho visam criar uma ponte entre a ciência de dados e a pesquisa em poluição do ar, auxiliando pesquisadores de outras áreas que têm pouco contato com a Estatística. Dado seu caráter multidisciplinar, é essencial que existam materiais acessíveis sobre cada uma das disciplinas que compõem esse campo de pesquisa para o adequado desenvolvimento dos trabalhos, aumentando a chance de criação de políticas públicas que promovam melhorias na saúde pública 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 *machine learning* 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. No Capítulo 3, apresentaremos modelos usuais utilizados no ajuste de dados de poluição. No Capítulo 4, discutiremos técnicas de *machine learning* úteis também para modelagem inferencial. Nos Capítulos 5 e 6, aplicaremos diversas estratégias discutidas nesta tese em estudos reais de poluição do ar. No Capítulo 7, apresentaremos formas de se extrair dados públicos de poluição do ar da internet. Por fim, no Capítulo 8, concluiremos a tese discutindo e sumarizando as conclusões mais importantes.

Seguindo a abordagem prática da ciência de dados, o texto a seguir busca ser acessível a pesquisadores de diversas áreas, sem abandonar totalmente o formalismo matemático. Um certo grau de conhecimento estatístico será exigido em muitos pontos. [Bussab e Morettin \(2013\)](#) e [Magalhães e Lima](#)

(2013) podem ser utilizados como referências de Estatística básica e Hastie *et al.* (2008); James *et al.* (2013) são ótimos livros para os tópicos de modelagem e *machine learning*. 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 corresponde a etapa “Visualizar” do ciclo da ciência de dados (Figura 1.1) e caracteriza a primeira tentativa de se extrair informação dos dados. Seu objetivo é gerar conhecimento inicial acerca do fenômeno sob estudo para guiar a etapa de modelagem. 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 grandes períodos de tempo. Por simplicidade computacional, as metodologias usuais para a análise de séries temporais supõem que as observações são realizadas em intervalos equidistantes. 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.

Séries temporais são normalmente representadas pela notação $\{Y_t, t \geq 0\}$. Neste texto, utilizaremos a forma simplificada Y_t . Aqui, Y representa o fenômeno sob estudo, denominado como *variável resposta*. O índice inteiro t representa o instante em que essa variável foi avaliada ($1, 2, 3, \dots$), podendo ser medido em minutos, horas, dias, anos etc. Na maioria dos casos, estaremos interessados em associar Y_t com p outras variáveis, chamadas variáveis explicativas ou preditores. Essas variáveis serão denotadas aqui por $X_{1t}, X_{2t}, \dots, X_{pt}$. Quando não houver risco de ambiguidade, omitiremos o índice t tanto de Y_t quanto de $X_{1t}, X_{2t}, \dots, X_{pt}$.

Sob o contexto de estudos de poluição do ar, apresentaremos nas próximas seções 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>. Em seguida, apresentaremos uma aplicação real de análise exploratória estudando os níveis de poluição durante a greve de caminhoneiros de 2018.

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 geral 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", contempla os principais modelos gráficos já conhecidos e abre caminho para a criação de estratégias bem estruturadas para construção de gráficos.

Wickham (2010), por exemplo, utilizou 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 heteroscedasticidade², 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,

¹ A layered grammar of graphics, em inglês.

² Variância não-constante ao longo do tempo.

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 região metropolitana de 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 Parque 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.

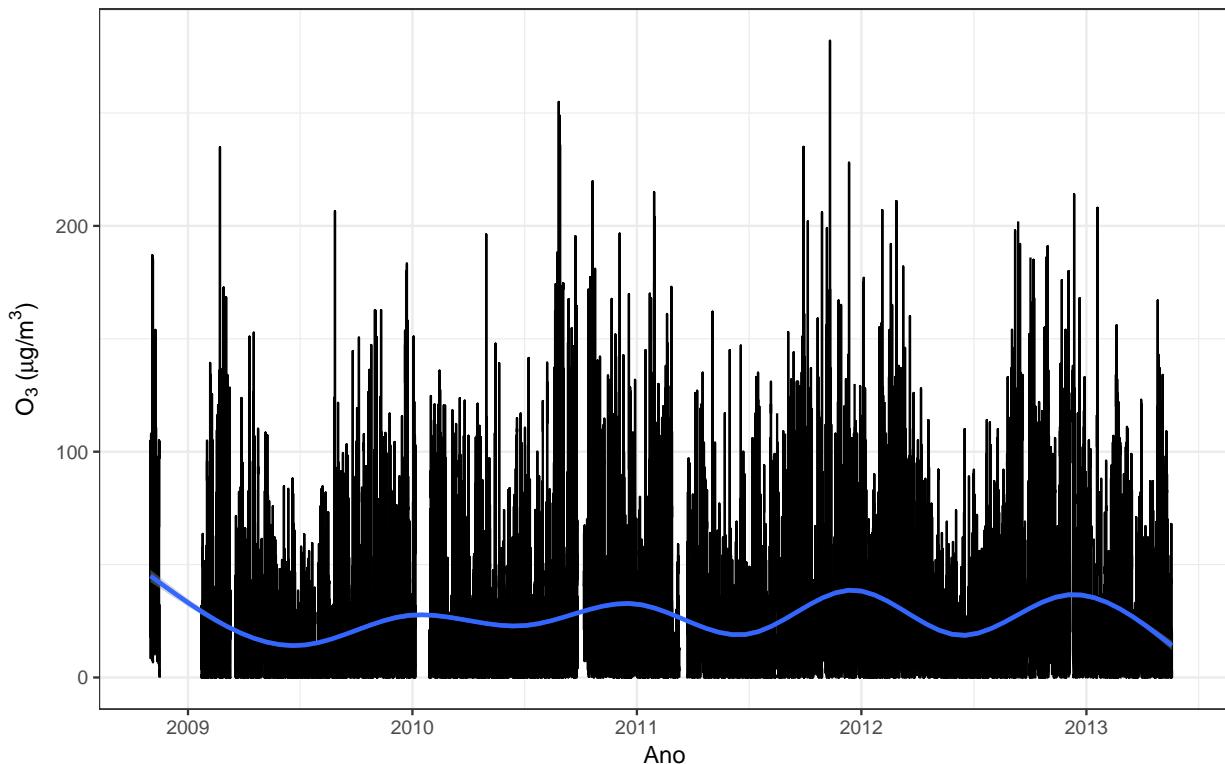


Figura 2.1: Série da concentração de ozônio para a estação Dom Pedro II, na cidade de São Paulo, no período de 2008 a 2013. Em azul, a série suavizada usando splines cúbicos.

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, das 12 às 16 horas.

Podemos então considerar a média diária dentro desse período para avaliar apenas o horário em que a concentração de ozônio normalmente 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.

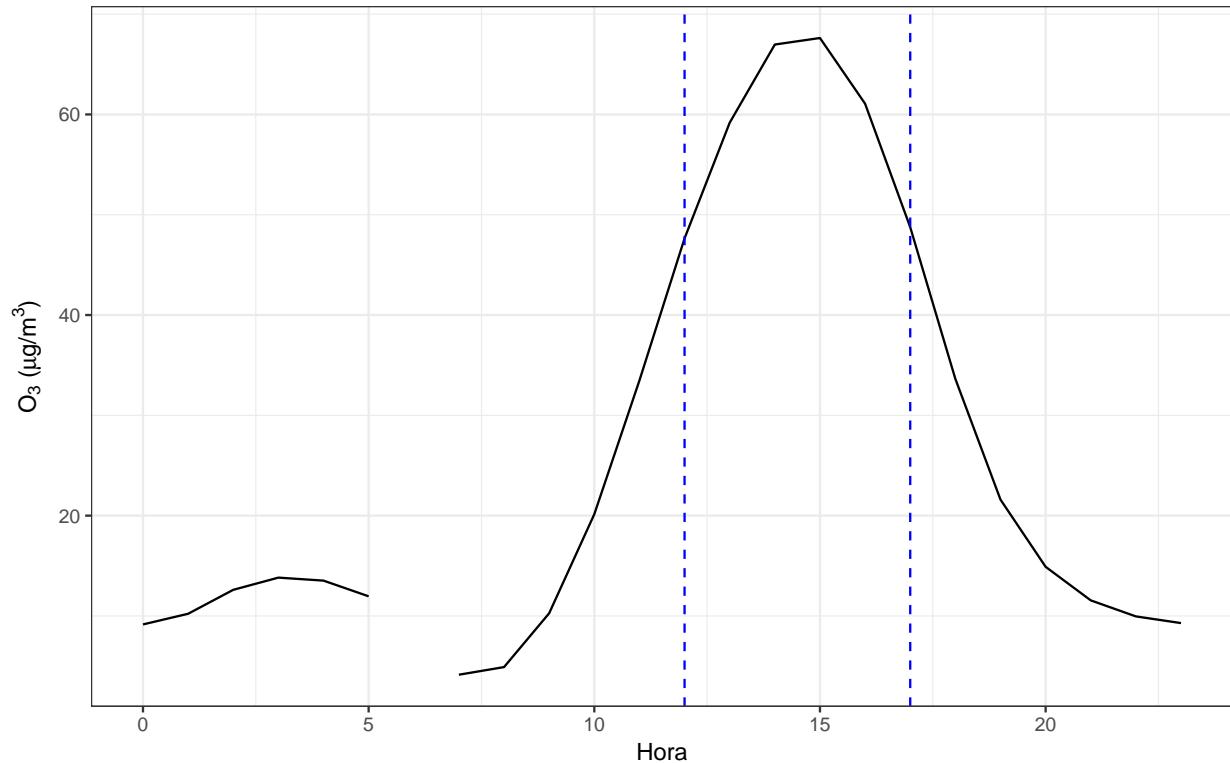


Figura 2.2: Série da concentração média de ozônio ao longo do dia para a estação Dom Pedro II, na cidade de São Paulo, no período de 2008 a 2013. Não existe informação para as 6 da manhã pois é o horário em que o equipamento sofre manutenção.

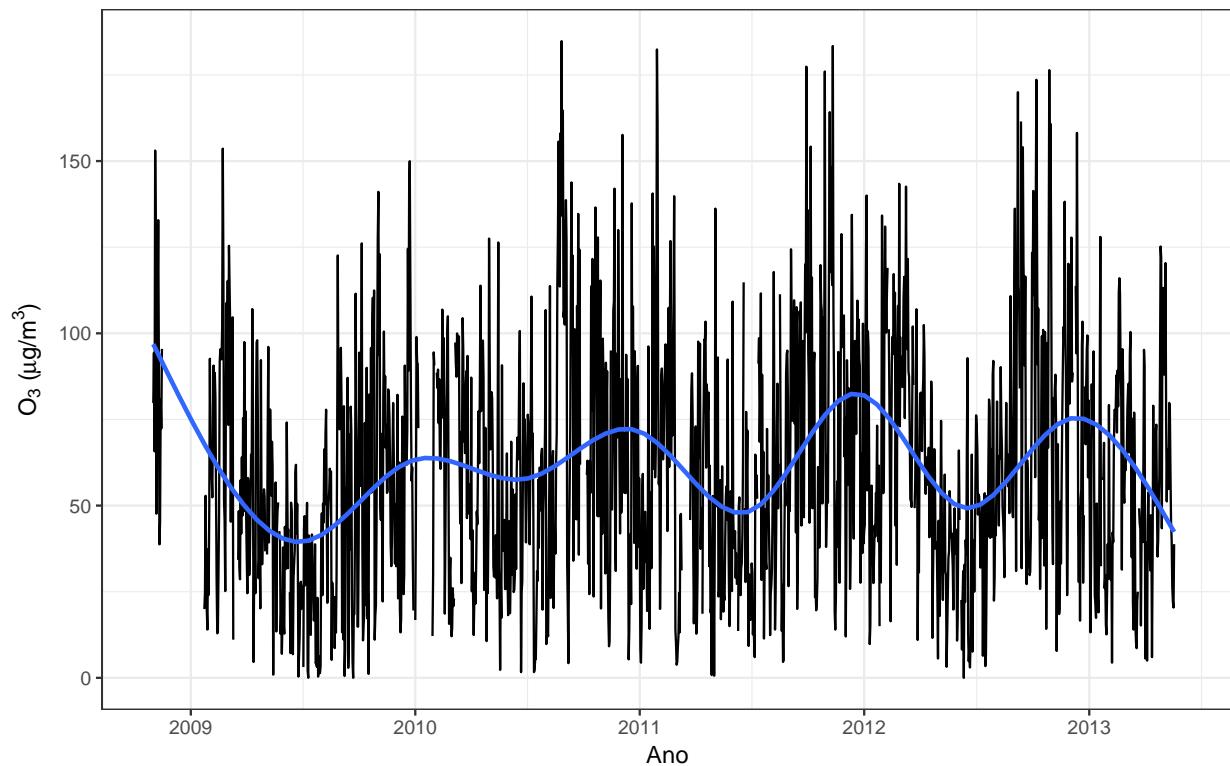


Figura 2.3: Série diária da concentração média de ozônio medida no começo da tarde para a estação Dom Pedro II, na cidade de São Paulo, no período de 2008 a 2013. Em azul, a série suavizada usando splines cúbicos.

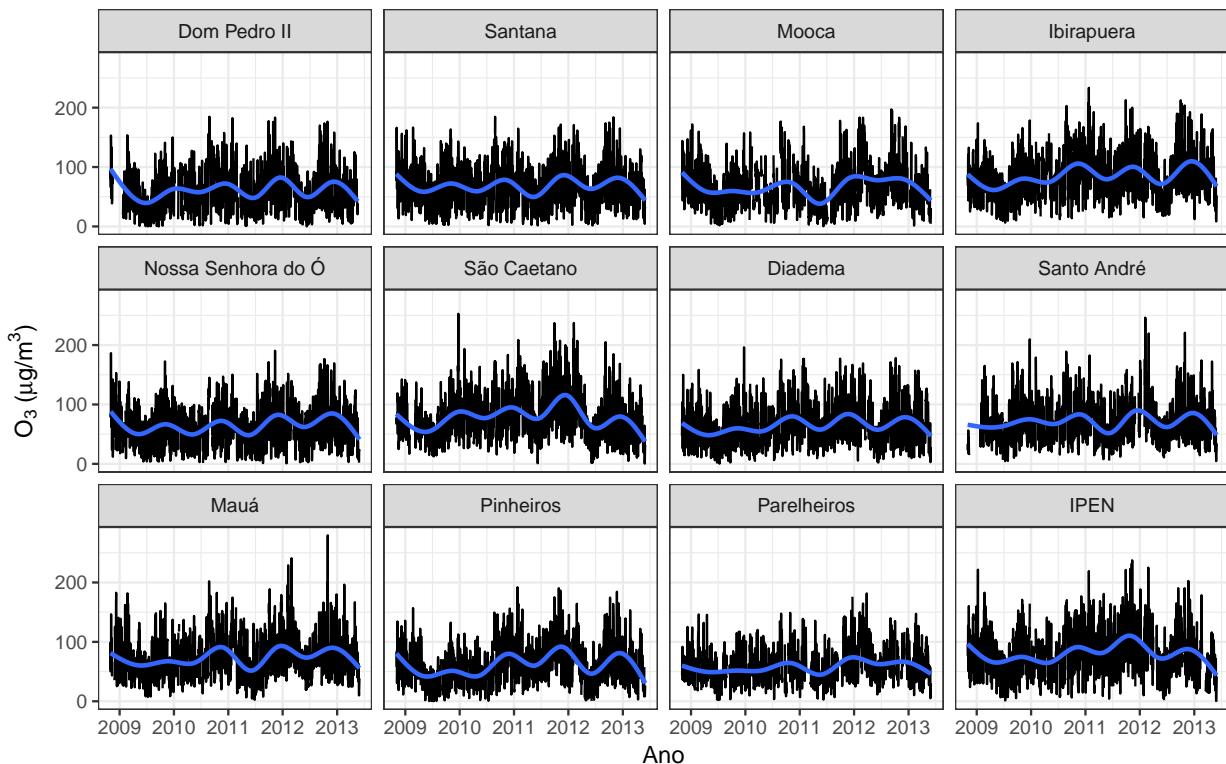


Figura 2.4: Série diária da concentração média de ozônio medido no começo da tarde para todas as estações, na cidade de São Paulo, no período de 2008 a 2013. Em azul, as séries suavizadas usando splines cúbicos.

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.

Embora plotar o gráfico de duas séries na mesma figura possa trazer informações sobre como essas variáveis estão relacionadas, gráficos de dispersão são mais eficientes nessa tarefa. Na próxima seção, traremos alguns exemplos de como construir e interpretar esses gráficos.

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 relação, 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.

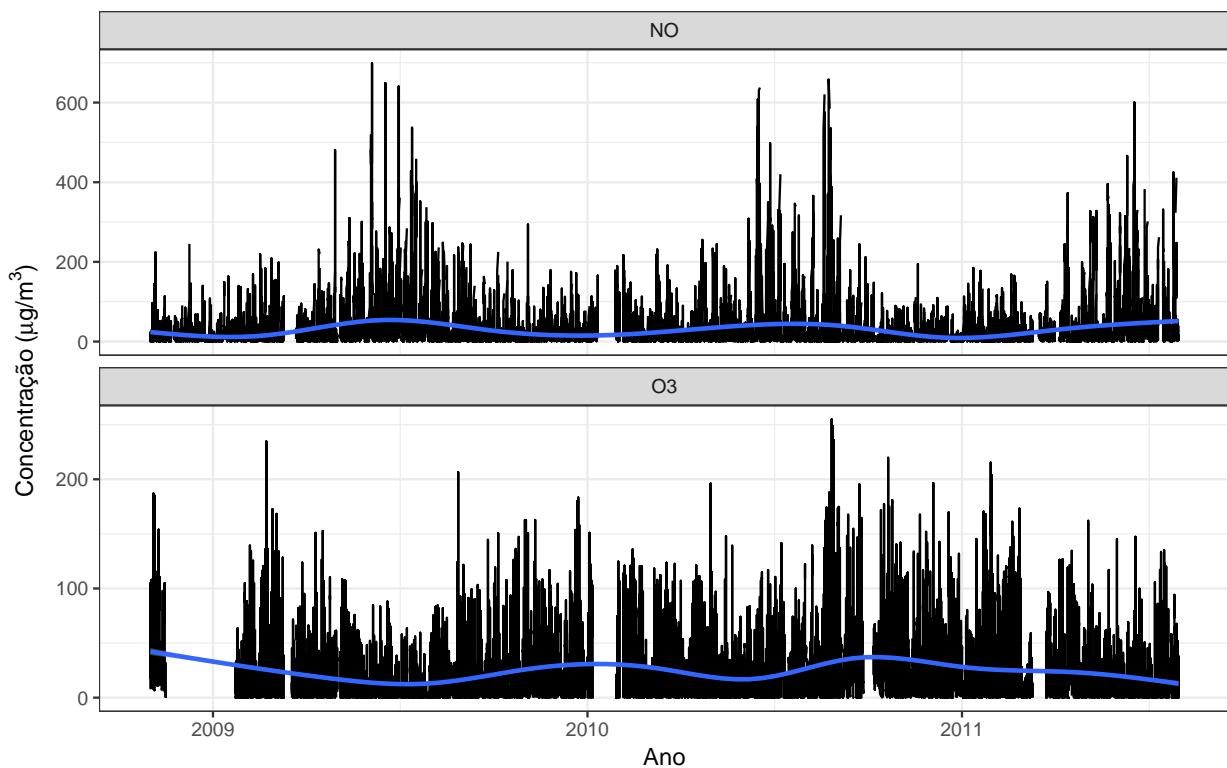


Figura 2.5: Séries horárias de ozônio e de óxido de nitrogênio (NO), ambos medidos na estação Dom Pedro II, em São Paulo, no período de 2008 a 2011. Em azul, as séries suavizadas usando splines cúbicos.

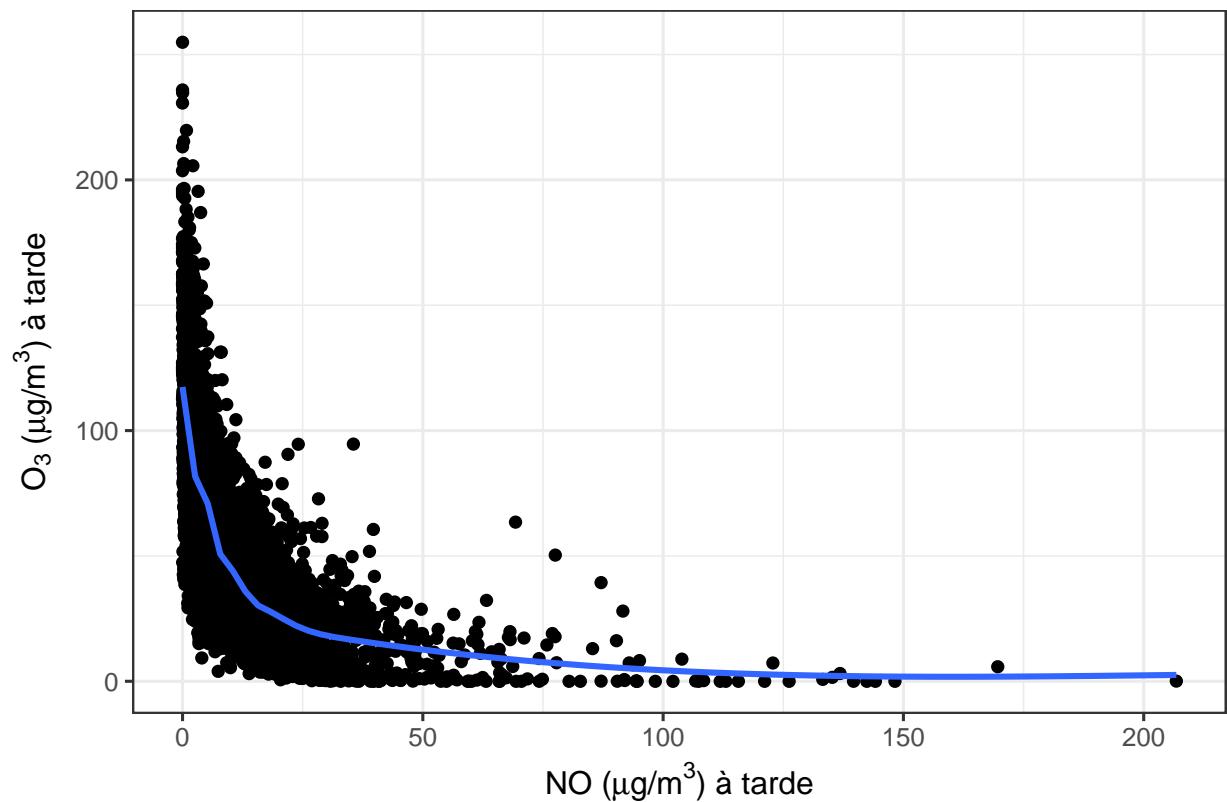


Figura 2.6: Gráfico de dispersão da concentração de ozônio contra a concentração de óxido de nitrogênio medidas das 12 às 16 horas na estação de monitoramento Dom Pedro II, em São Paulo, de 2008 a 2011.

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 NO₂ 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 de 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.

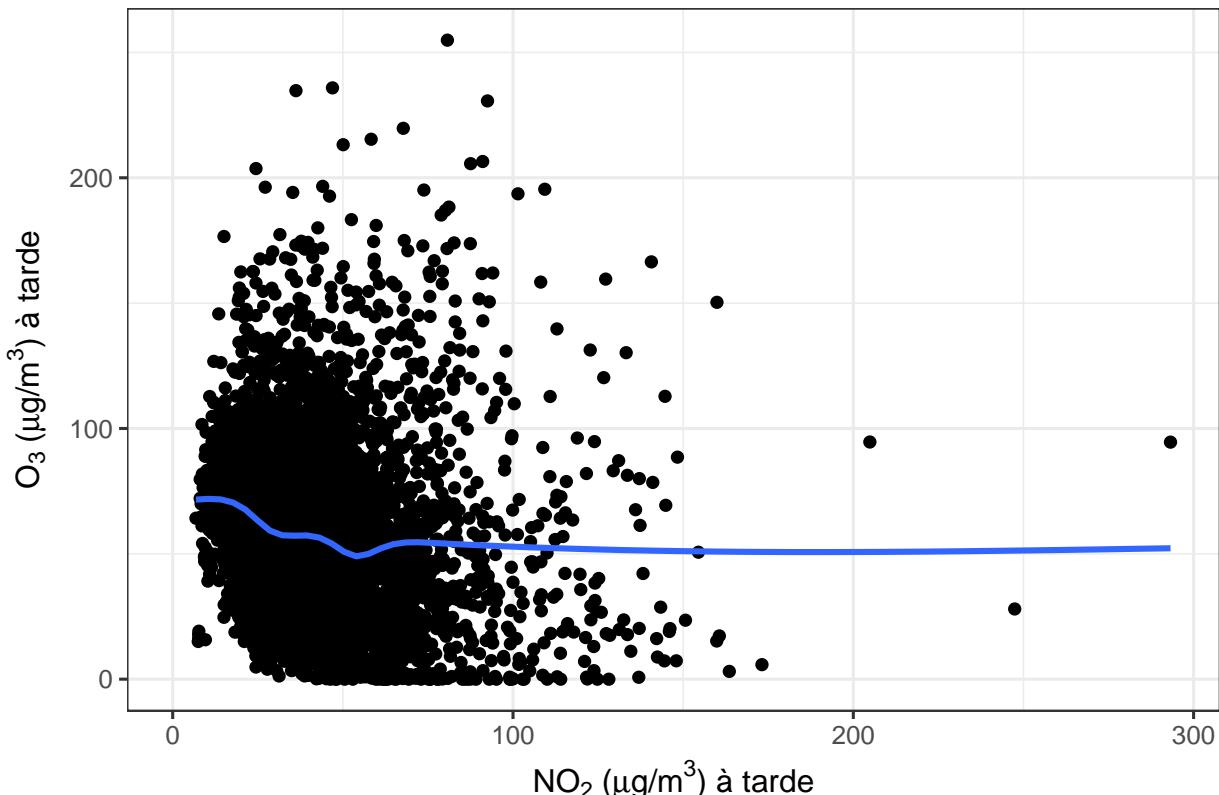


Figura 2.7: Gráfico de dispersão da concentração de ozônio contra a concentração de dióxido de nitrogênio medidas das 12 às 16 horas na estação de monitoramento Dom Pedro II, em São Paulo, de 2008 a 2011.

Uma limitação dos gráficos de dispersão é não levar em conta o efeito de outras variáveis. Muitas vezes a associação entre 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 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 $\mu\text{g}/\text{m}^3$.

Quando estamos interessados em, além de observar a distribuição amostral da variável resposta, compará-la entre os níveis de uma variável explicativa, podemos utilizar os *boxplots*. A partir dos

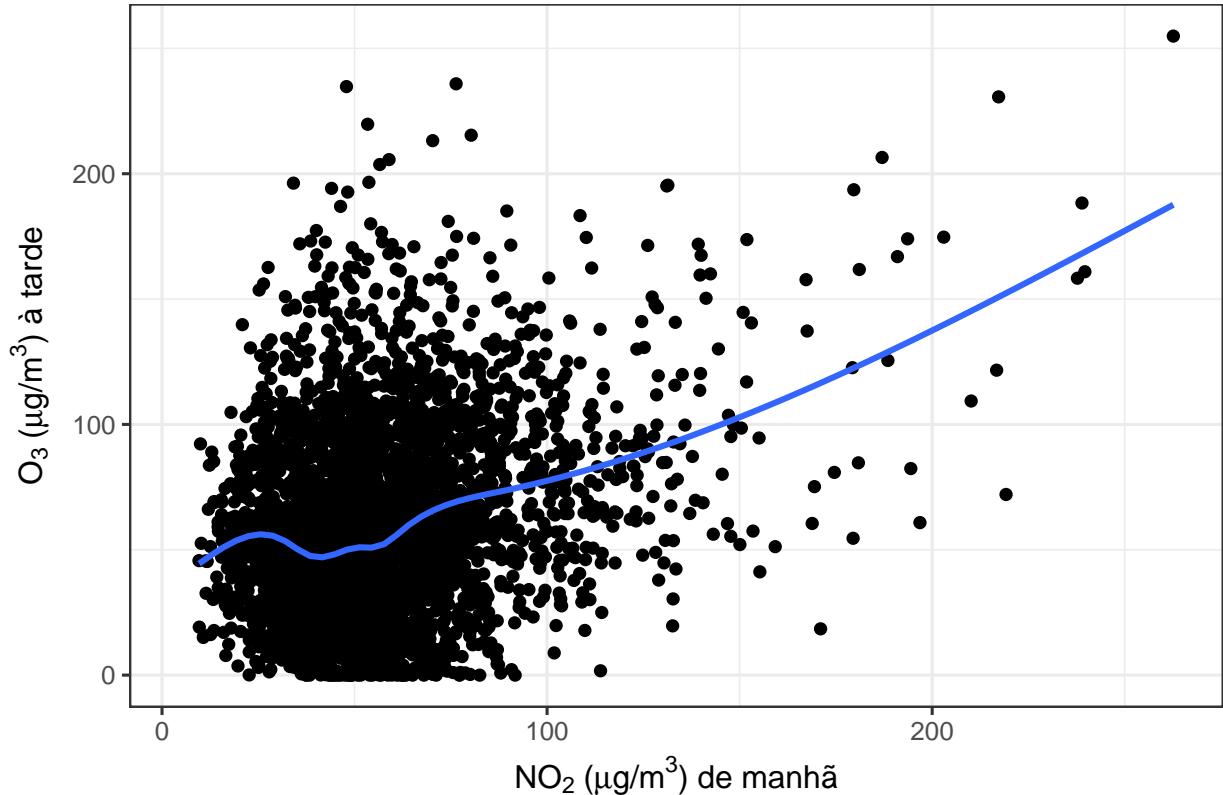


Figura 2.8: Gráfico de dispersão da concentração de ozônio, medida das 12 às 16 horas, contra a concentração de óxido de nitrogênio, medida das 7 às 11 horas, na estação de monitoramento Dom Pedro II, em São Paulo, de 2008 a 2011.

quantis da variável resposta, esses gráficos dão uma ideia geral da sua distribuição, para cada nível da variável explicativa. Eles também nos mostram a presença de pontos atípicos, isto é, observações com valores muito abaixo ou muito acima dos valores medianos. Os chamados *ridges graphs* são outra boa alternativa para comparar a distribuição amostral de uma variável entre os níveis de um preditor. Eles são histogramas suavizados e trazem mais informação sobre a forma da distribuição do que os boxplots. Na Figura 2.10, apresentamos um exemplo desses gráficos. Podemos observar 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.

Os gráficos apresentados até aqui 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, apresentaremos estratégias para conduzir a análise na presença de tendência e sazonalidade.

2.2 Componentes temporais

Séries temporais normalmente sofrem influência de componentes temporais que não são causados pelo fenômeno que estamos estudando ou que estão ligados a variáveis que não puderam ser medidas. Imagine, por exemplo, que estamos investigando a associação da concentração de monóxido de carbono com o número de carros rodando no horário de pico na cidade de São Paulo. Sabemos que,

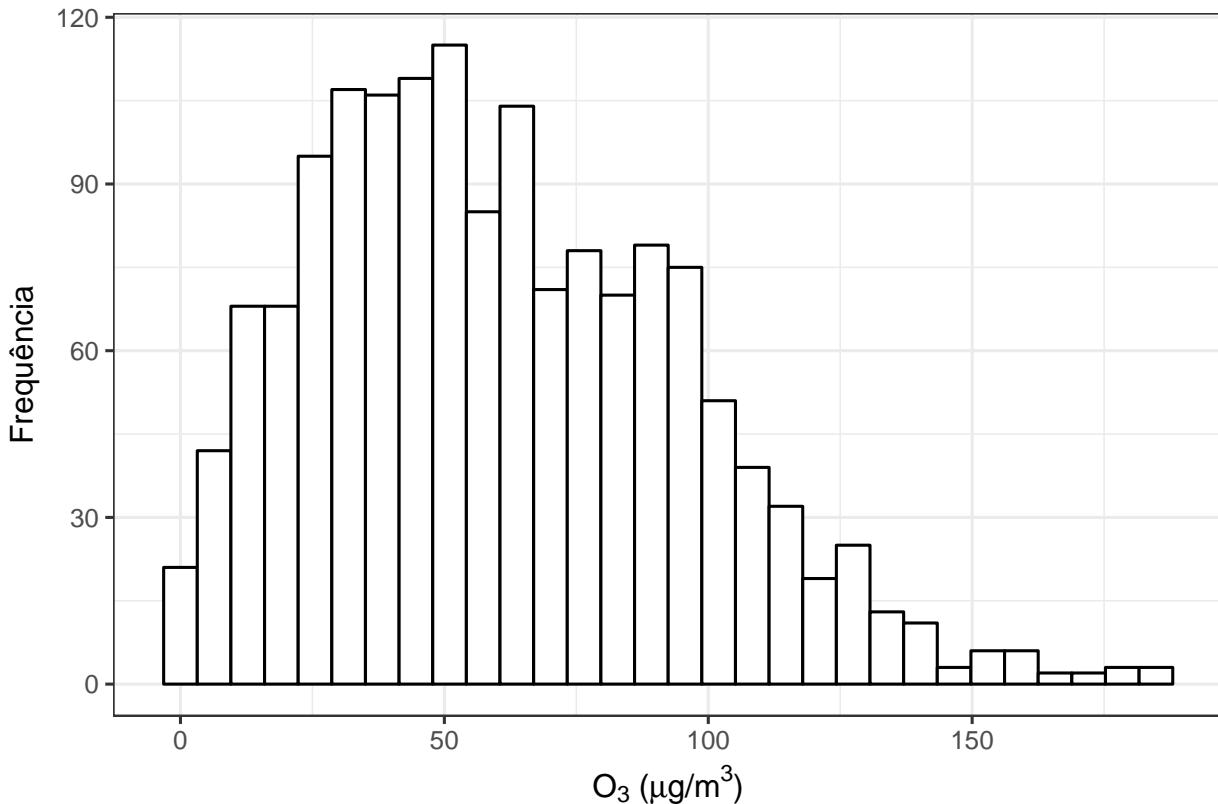


Figura 2.9: Histograma da concentração diária média de ozônio medida das 12 às 16 horas na estação de monitoramento Dom Pedro II, na cidade de São Paulo, de 2008 a 2013.

historicamente, os níveis de monóxido de carbono vêm diminuindo, enquanto o número de carros na cidade de São Paulo tendem a aumentar. Se não controlarmos por nenhuma outra variável, nossa investigação pode concluir que o aumento do número de carros está diminuindo os níveis do poluente, o que não faria sentido prático pois sabemos que existe uma associação positiva entre as variáveis. Acontece que a diminuição histórica do monóxido de carbono se deve a regulamentações feitas em cima dos combustíveis e dos motores veiculares. Como podemos não ter dados disponíveis para incorporar essa informação no modelo, essa *tendência* decrescente da série precisa ser eliminada ou controlada de alguma forma. Só assim conseguiremos quantificar corretamente o efeito do número de carros na concentração de monóxido de carbono.

Tendências são os componentes temporais mais comuns em séries de tempo. Outro componente muito frequente é *sazonalidade*, que representa comportamento cíclicos em intervalos fixos de tempo. Em estudos de poluição, as estações do ano são a principal causa de sazonalidade.

A seguir discutiremos como identificar e eliminar esses componentes de uma série. No Capítulo 3 discutiremos como controlá-las incorporando termos de tendência e sazonalidade no modelo.

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

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

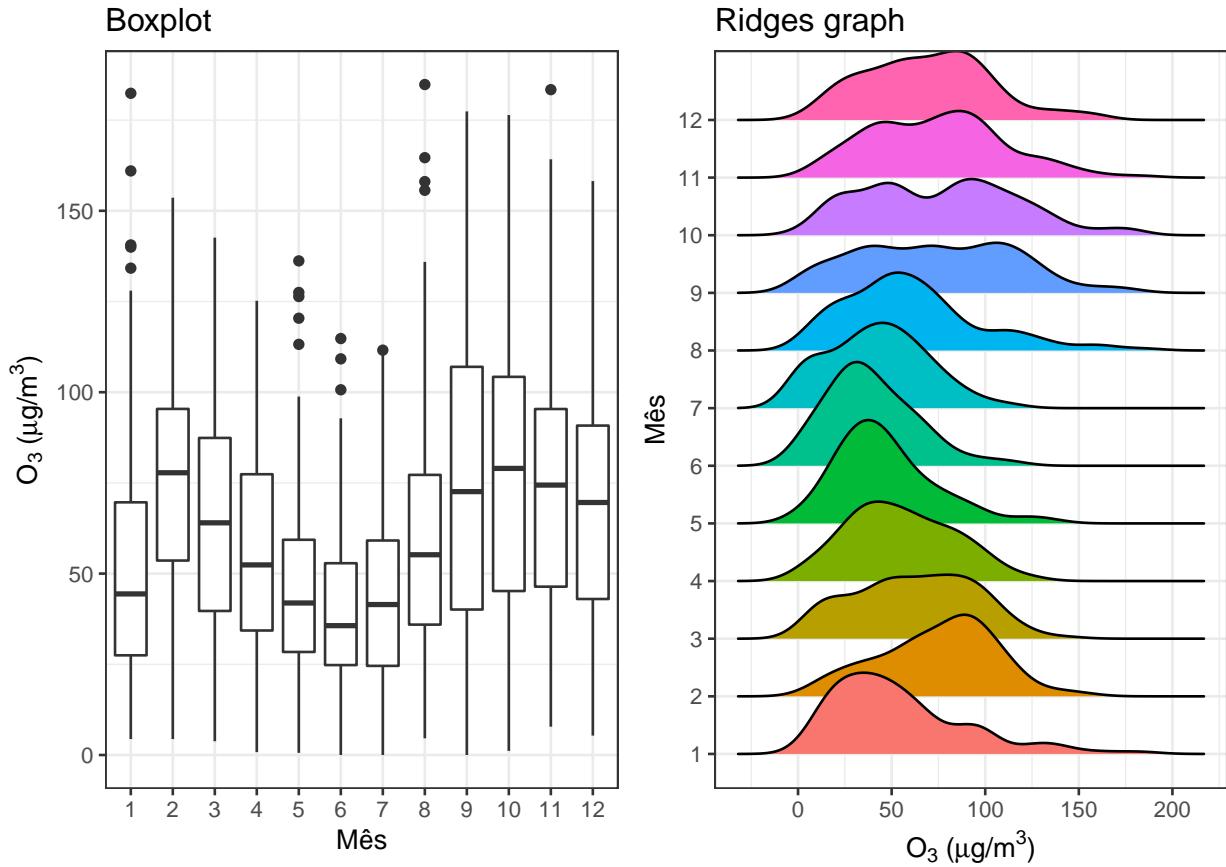


Figura 2.10: Distribuição por mês da concentração diária média de ozônio medida das 12 às 16 horas na estação de monitoramento Dom Pedro II, na cidade de São Paulo, de 2008 a 2013.

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

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

No caso geral, definimos a diferenciação de ordem n como

$$\Delta^n Y_t = \Delta^{n-1} Y_t - \Delta^{n-1} Y_{t-1}, \quad t = 1, 2, \dots \quad (2.1)$$

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.

Uma desvantagem de se utilizar a série de diferenças é a interpretação do modelo, já que a variável resposta ajustada terá sido a diferença entre duas observações consecutivas. As conclusões para essa nova variável nem sempre será interessante.

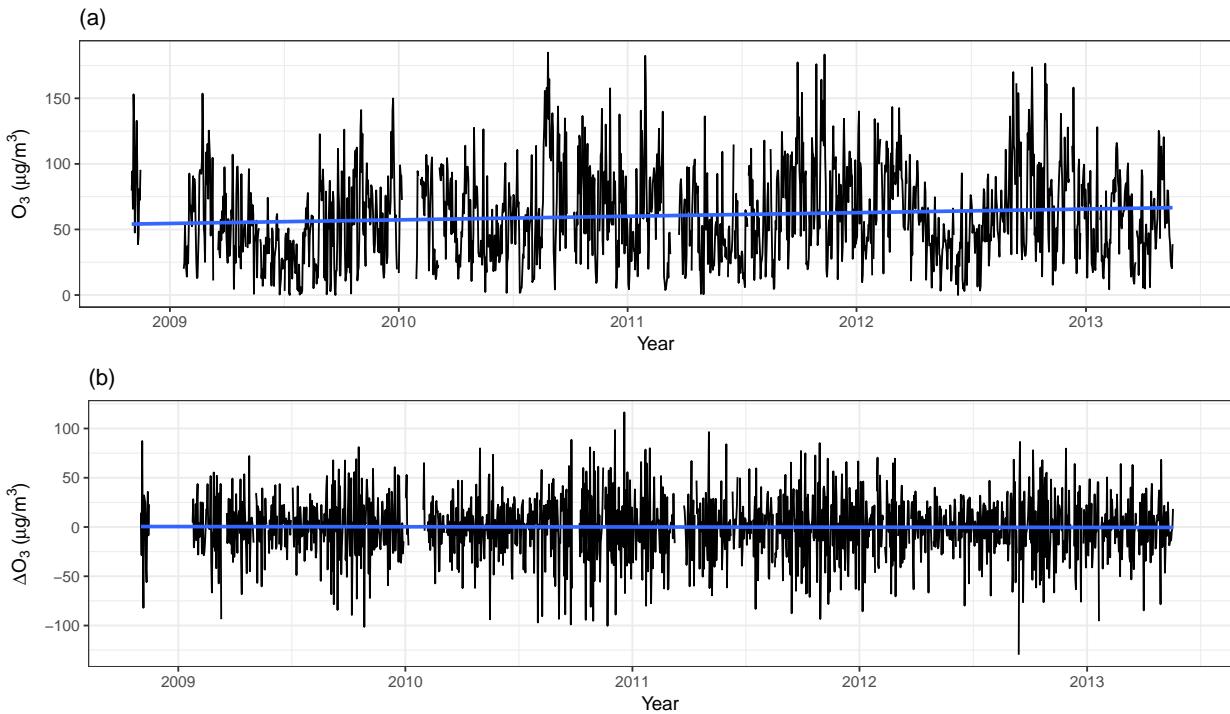


Figura 2.11: Série da concentração de ozônio diária média, medida na cidade de São Paulo (estação de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2011, das 12h às 16h.

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 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 período³ 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 contemplam o componente sazonal, como veremos no Capítulo 3.

³O período é o inverso da frequência.

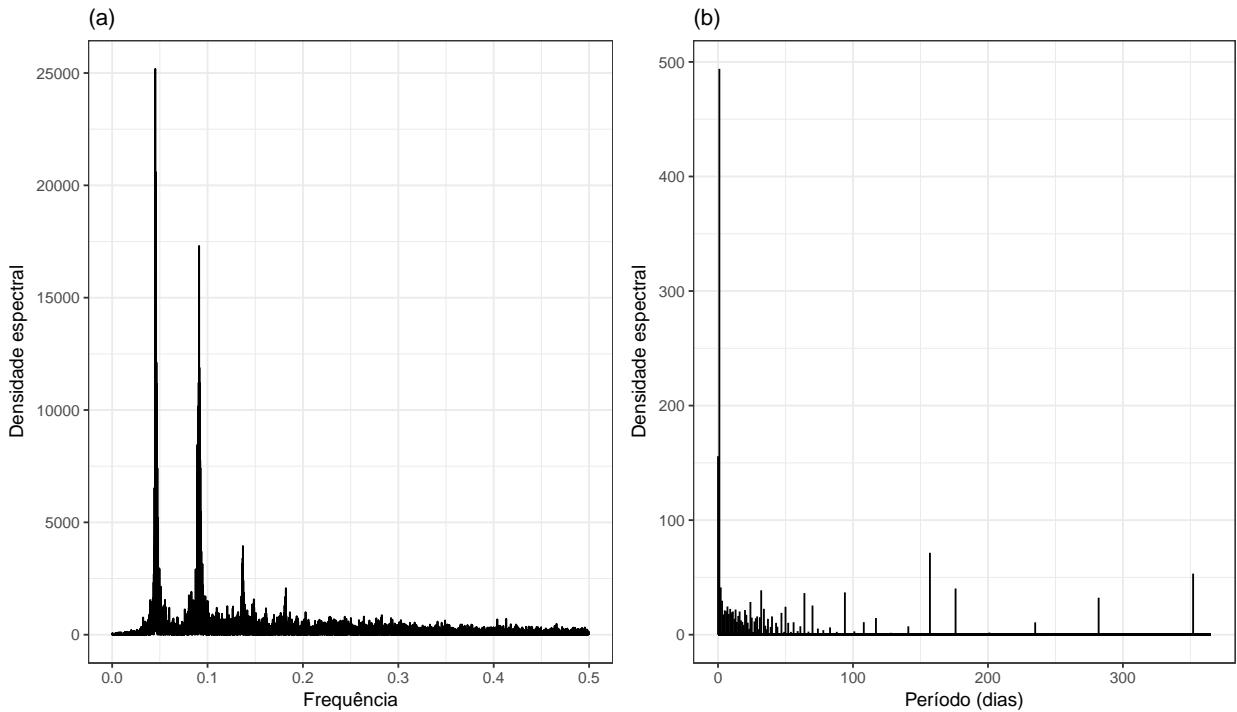


Figura 2.12: Periodogramas para a concentração horária de ozônio medida na cidade de São Paulo (estação de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2013. Dados disponibilizados por Salvo e Geiger (2014). No painel (a), apresentamos a densidade espectral contra a frequência.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

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

2.2.3 Autocorrelação

É natural supormos 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}, \dots , ou das seguintes, Y_{t+1}, Y_{t+2}, \dots , 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*. Ela supõe que a relação entre as observações pode ser descrita por uma função linear, ou seja, invariante ao valor das observações⁴. Quando outro tipo de relação não for especificada, essa será a definição utilizada neste trabalho para descrevermos a correlação temporal entre as observações.

A autocorrelação de uma série pode ser representada pela *função de autocorrelação*, digamos $\rho(s, t)$, que mede a previsibilidade da série no instante t , a partir apenas do valor da variável no instante s . Essa medida varia no intervalo $[-1, 1]$, com os extremos representando uma correlação perfeita entre as observações Y_t e Y_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

⁴Para mais detalhes sobre a interpretação de linearidade, consulte a Seção 3.1.5.

negativa.

Fazendo $h = t - s$, a função de autocorrelação pode ser estimada por

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

sendo

$$\gamma(h) = \frac{1}{n} \sum_{t=1}^{n-h} (y_{t+h} - \bar{y})(y_t - \bar{y}) \quad (2.2)$$

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 autocorrelação da concentração de ozônio medida na estação Parque Dom Pedro II. Podemos observar que a autocorrelação é sempre positiva e não decai para o zero, indicando que a série apresenta tendência. Em caso contrário, esperaríamos que apenas observações próximas fossem correlacionadas, e então a função de autocorrelação convergiria rapidamente para zero conforme aumentássemos o valor de h .

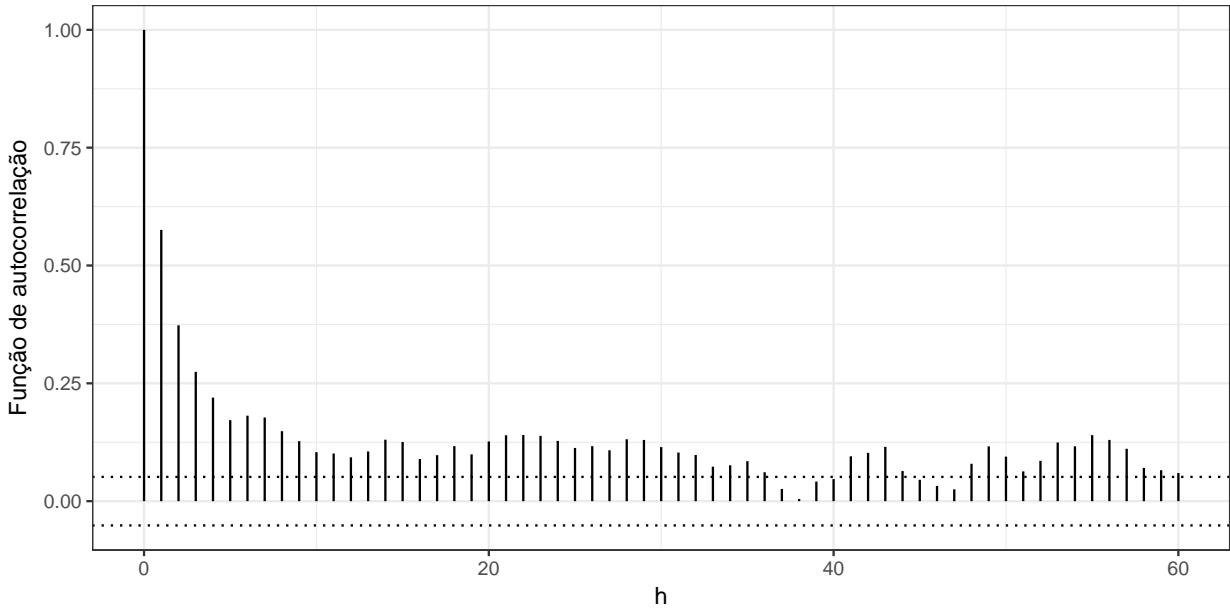


Figura 2.13: Função de autocorrelação da concentração de ozônio diária média, medida na cidade de São Paulo (estação de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2011, das 12h às 16h. Dados disponibilizados por [Salvo e Geiger \(2014\)](#). As linhas pontilhadas representam os limites $\pm 2/\sqrt{n}$, sendo n o tamanho da amostra. Valores fora desse intervalo de confiança (95%) podem ser considerados significativamente 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 $Y_{t-1}, Y_{t-2}, \dots, Y_{t-m+1}$.

Na Figura 2.14, apresentamos a função de autocorrelação 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 significativas, indicando que a série realmente apresenta alguma tendência.

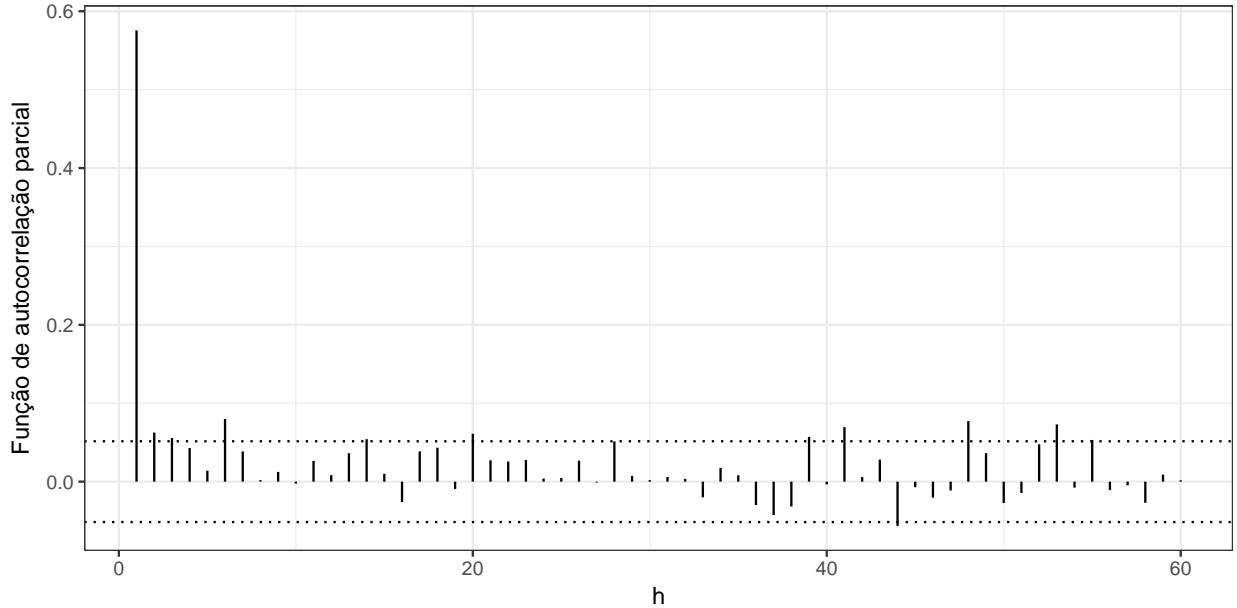


Figura 2.14: Função de autocorrelação parcial da concentração de ozônio diária média, medida na cidade de São Paulo (estaçao de monitoramento Dom Pedro II), no período de outubro de 2008 a junho de 2011, das 12h às 16h. As linhas pontilhadas representam os limites $\pm 2/\sqrt{n}$, sendo n o tamanho 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, utilizamo a *função de correlação cruzada*, que pode ser estimada por

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}},$$

sendo $\gamma_s(h)$ como definido em (2.2) e

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

a função de covariância cruzada amostral. 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 $\rho_{XY}(h) = \rho_{YX}(-h)$.

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 Parque 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.

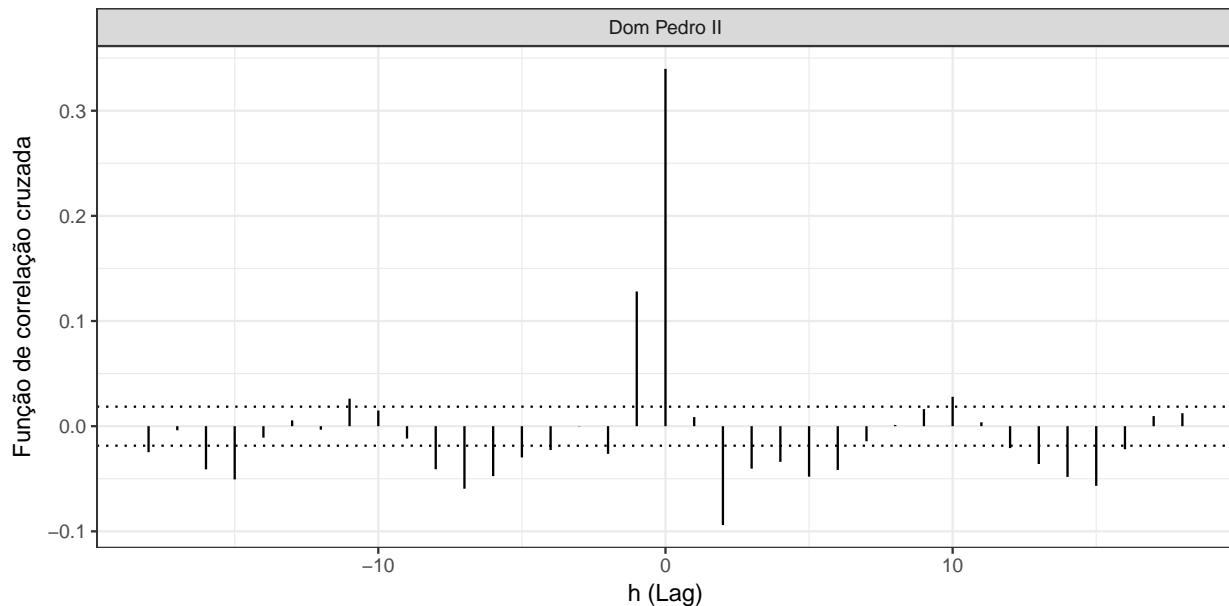


Figura 2.15: Função de correlação cruzada do ozônio em função da temperatura na estação Dom Pedro II (São Paulo) no período de outubro de 2009 a junho de 2011. As linhas pontilhadas representam os limites $\pm 2/\sqrt{n}$, sendo n o tamanho da amostra. Valores fora desse intervalo de confiança (95%)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

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.

2.3 Visualizando dados de poluição durante a greve de caminhoneiros

A greve dos caminhoneiros foi como ficou conhecida a paralisação de caminhoneiros autônomos em todo o território brasileiro em maio de 2018. As manifestações começaram no dia 21 de maio

e duraram até o início de junho. Nesse período, muitas cidades sofreram com desabastecimento, principalmente de combustível, diminuindo não apenas o tráfego de veículos pesados, mas também de automóveis. Como as emissões veiculares são a principal fonte de diversos poluentes em centros urbanos, seria interessante analisarmos o impacto dessas paralisações nos níveis de poluição.

Utilizando os dados disponibilizados pela Companhia de Ambiental do Estado de São Paulo (CETESB), analisamos a concentração de alguns poluentes entre os dias 23 e 30 de maio⁵ na região metropolitana de São Paulo. Os poluentes considerados foram: monóxido de carbono (CO), ozônio (O_3), monóxido e dióxido de nitrogênio (NO e NO_2) e material particulado 10 (MP10). Também consideramos períodos anteriores e posteriores à greve, para avaliar a mudança causada pelas paralisações, e os mesmos dias em anos anteriores, em que não houve greve. O período total analisado foi de 1º de maio a 14 de junho, dos anos de 2016, 2017 e 2018.

As concentrações de cada poluente foram medidas em estações de monitoramento da CETESB: Osasco, Pinheiros, Parque Dom Pedro II e Ibirapuera. O critério para a escolha foi a disponibilidade de dados para os poluentes escolhidos e o perfil do tráfego de veículos na região das estações. As estações Parque Dom Pedro II e Pinheiros ficam em regiões de tráfego intenso, a primeira no centro da cidade e a segunda próxima à marginal Pinheiros, via expressa que liga as zonas sul, oeste e norte. A estação de Osasco também fica numa região de tráfego intenso e relativamente próxima a duas rodovias. A estação Ibirapuera não é muito afetada pelo tráfego pois fica dentro do Parque Ibirapuera e será utilizada como comparação.

Para construir os gráficos das séries, utilizamos o gráfico apresentado na Figura 2.16 para avaliar a média horária de cada poluente em cada dia da semana. Assim, em vez de utilizarmos as séries horárias, que apresentam sazonalidade diária, construímos as séries da média diária nos horários de pico. Esse gráfico mostra, por exemplo, que os picos de CO acontecem de manhã e no começo da noite e que os níveis desse poluente são bem menores nos fins de semana.

Na Tabela 2.1, apresentamos a variação da média dos poluentes em cada estação no período de greve em relação à média nos períodos anterior e posterior à greve. Nas Figuras 2.17, 2.18, 2.19, 2.20 e 2.21 apresentamos, respectivamente, as séries para o monóxido de carbono, o ozônio, o monóxido e dióxido de nitrogênio e para o material particulado. Observamos que, com exceção do ozônio, a concentração média dos poluentes durante o período de paralisação diminuiu. A maior redução foi a de NO, que é diretamente produzido pela queima de combustíveis, principalmente gasolina e diesel.

O ozônio é produto de um complexo processo químico que ocorre ao longo do dia, envolvendo diversos compostos e a radiação solar, sendo que uma explicação para o aumento de sua concentração durante a greve pode ser dada pela diminuição dos níveis de NO. O NO faz parte do balanço diário do ozônio, consumindo-o ao longo da tarde e diminuindo suas concentrações. Como o NO diminuiu devido a redução do tráfego de veículos, menos ozônio era consumido e por isso o aumento na concentração.

Esta análise considera apenas as séries durante o período de greve para explicar a variação da concentração dos poluentes. Uma análise mais completa deveria considerar também os efeitos climáticos (temperatura, precipitação, vento, radiação, entre outros). As conclusões aqui supõem que esses fatores se mantiveram homogêneos durante o período analisado, o que pode não ser razoável. Mais resultados podem ser encontrados em <https://www.rpollution.com/blog/greve-caminhoneiros/>.

⁵Período em que as consequências das paralisações foram mais intensas.

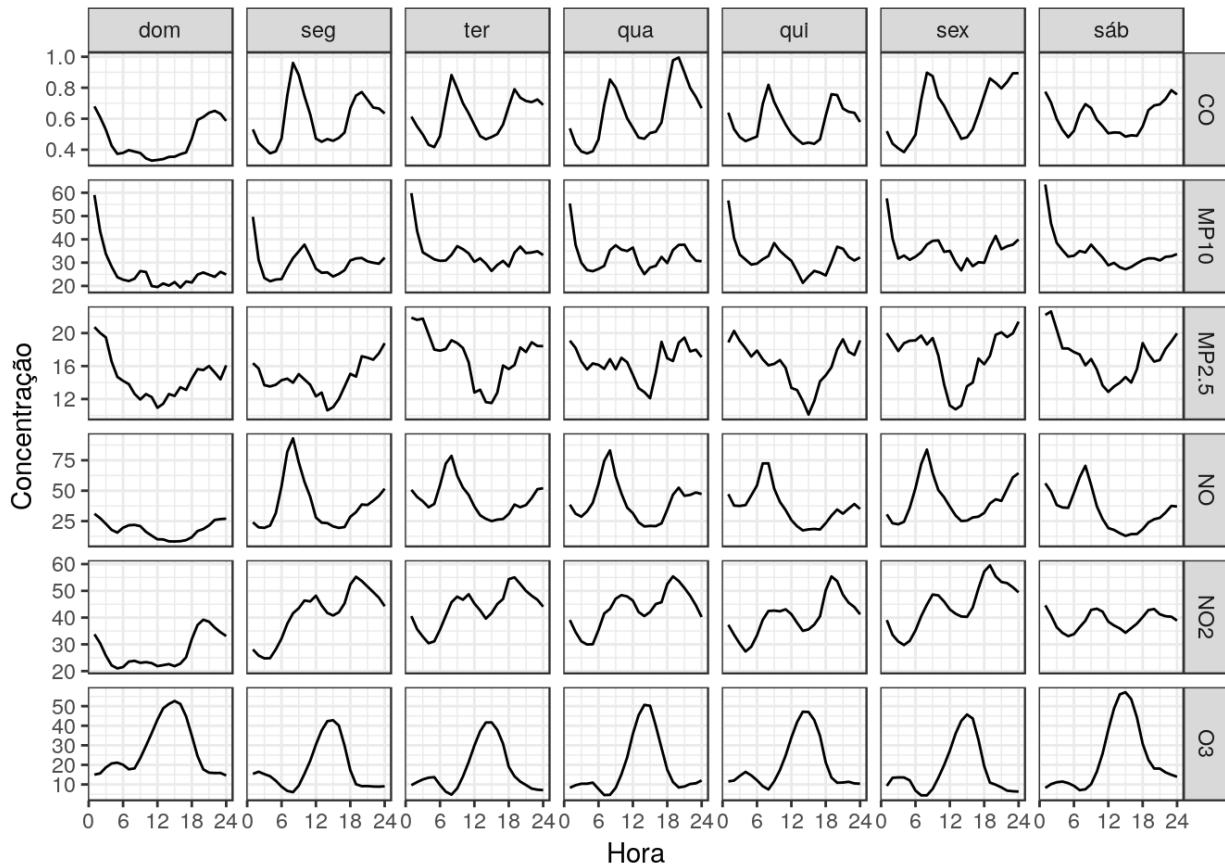


Figura 2.16: Médias horárias por dia da semana durante o período observado dos poluentes considerados na análise.

Tabela 2.1: Variação da média dos poluentes em cada estação no período de greve em relação à média nos períodos anterior e posterior à greve

Poluente	Ibirapuera	Osasco	Parque D. Pedro II	Pinheiros
CO	-48.02%	-32.78%	-51.16%	-65.9%
O ₃	53.7%	N/A	68.43%	126%
NO	-89.66%	-50.9%	-75.38%	-83.09%
NO ₂	-42.49%	-13.24%	-38.73%	-39.45%
MP10	N/A	-19.1%	-19.06%	-20.06%

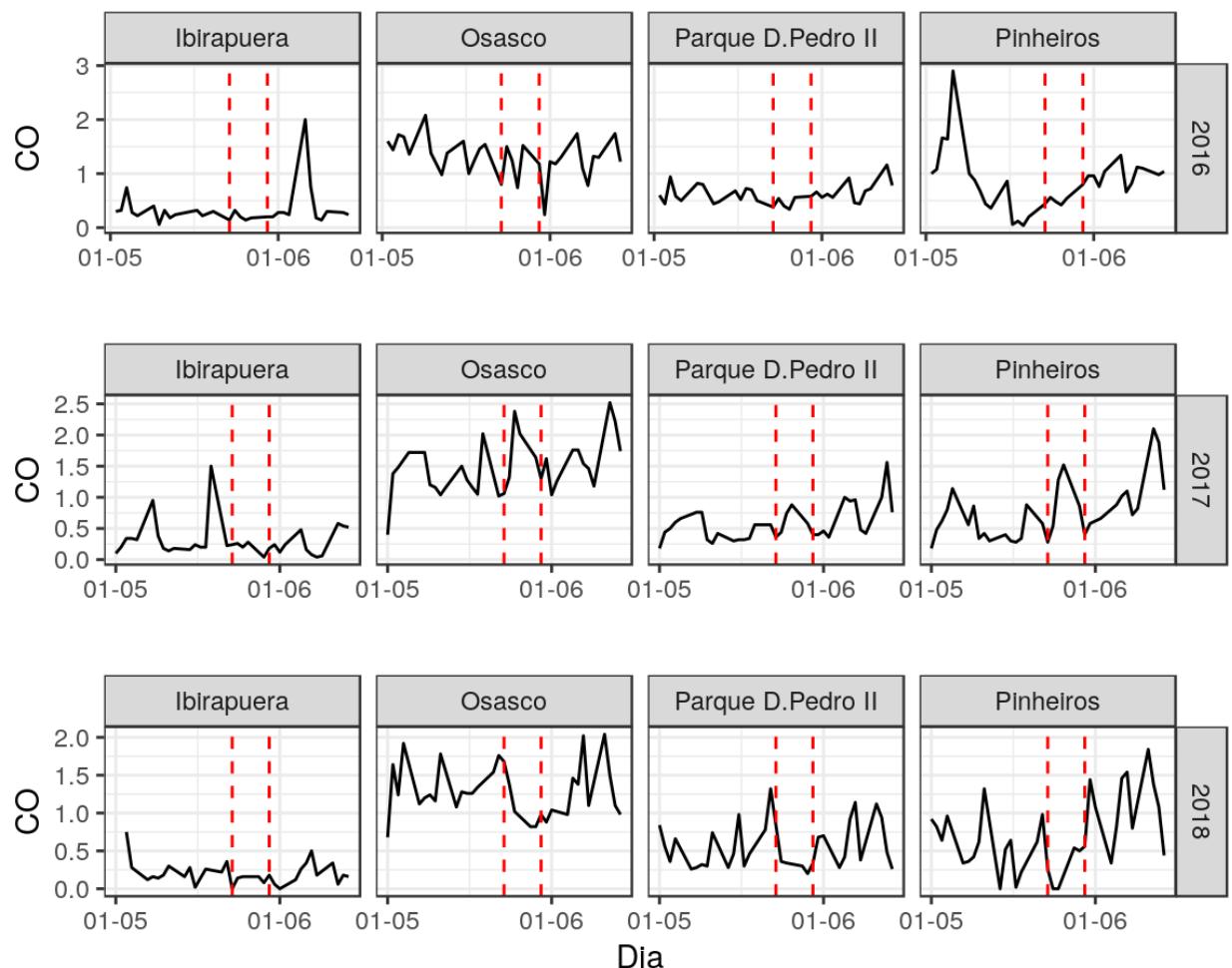


Figura 2.17: Série observada para o monóxido de carbono. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.

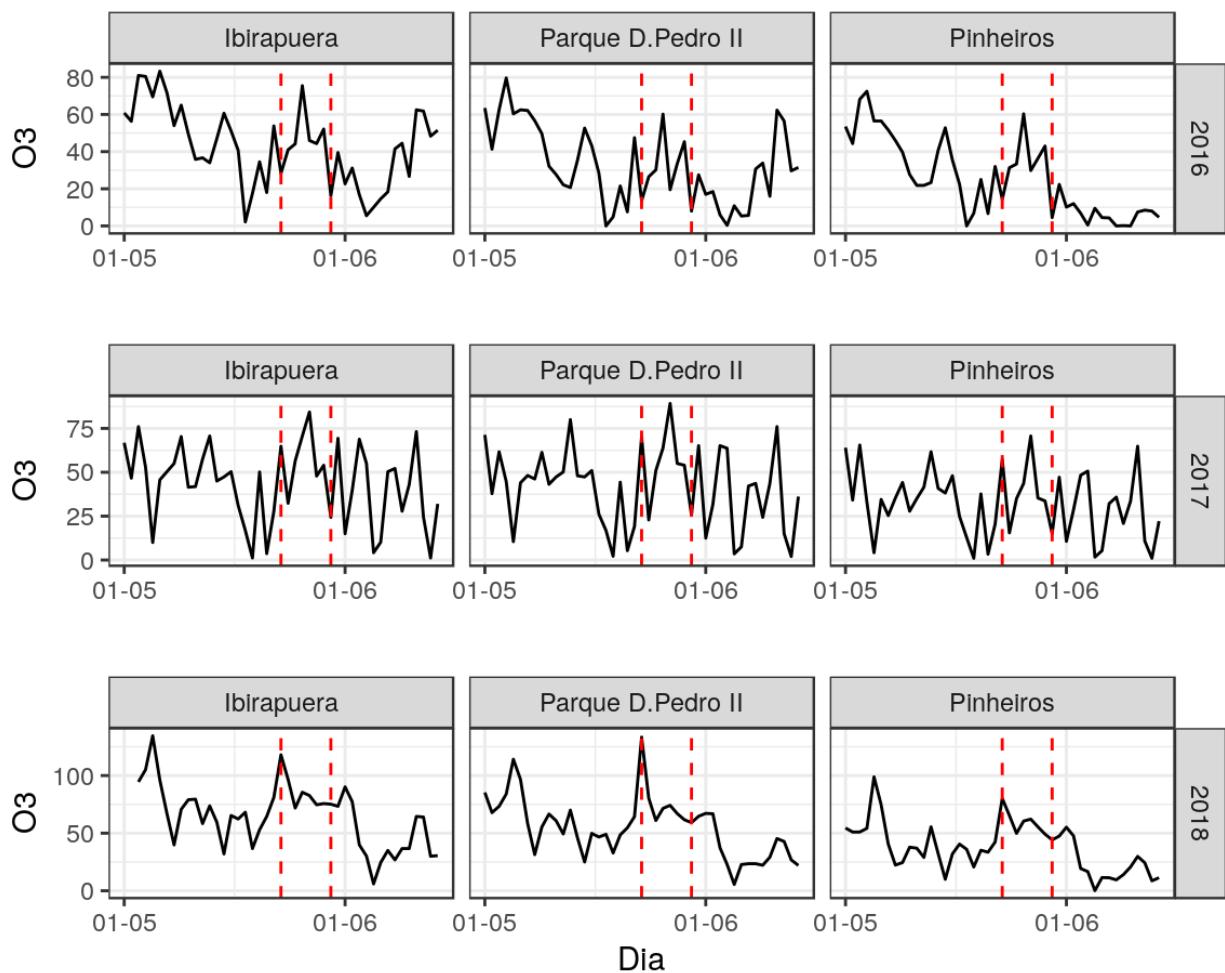


Figura 2.18: Série observada para o ozônio. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.

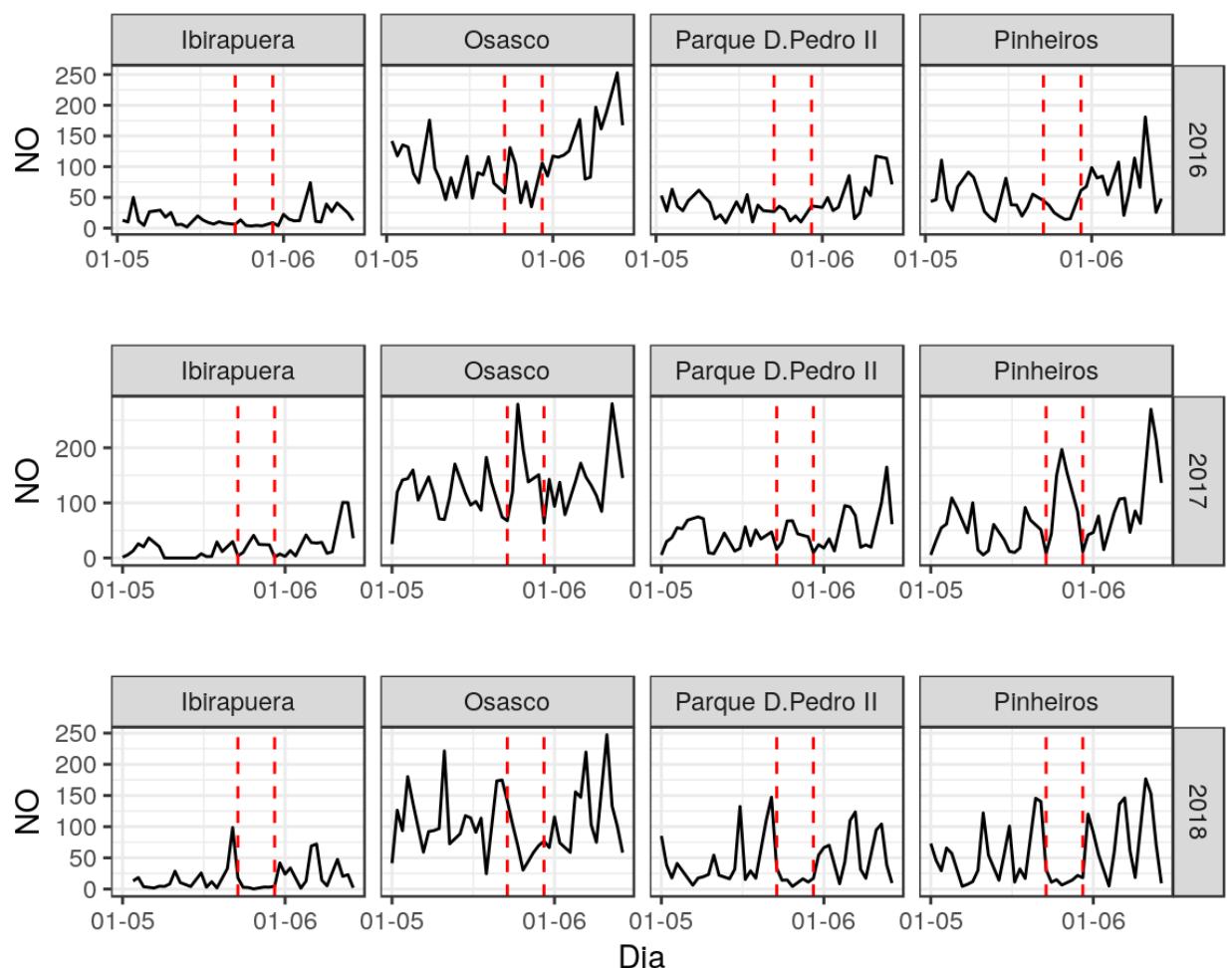


Figura 2.19: Série observada para o monóxido de nitrogênio. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.

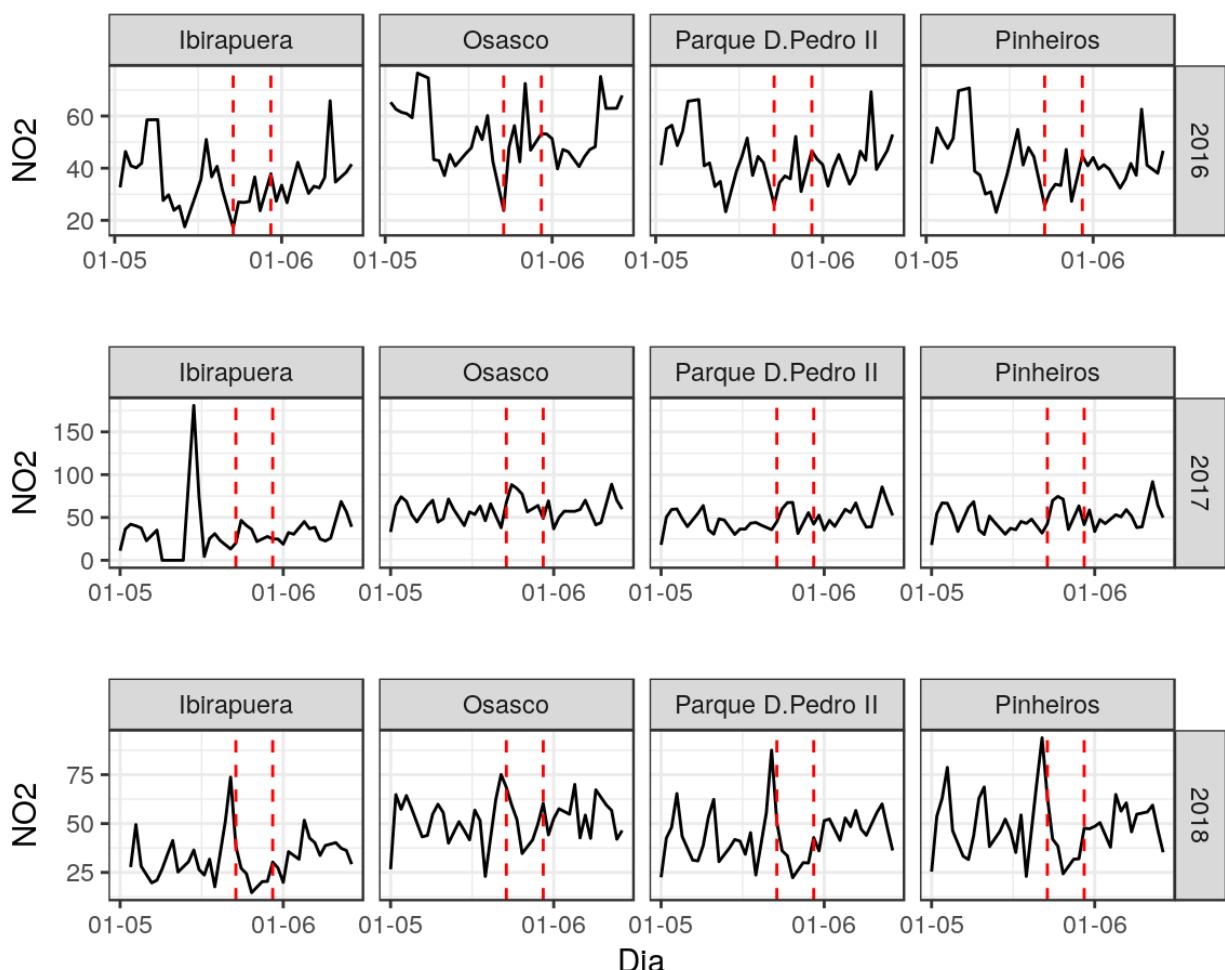


Figura 2.20: Série observada para o dióxido de nitrogênio. O intervalo entre as retas pontilhadas corresponde ao período de paralisações.

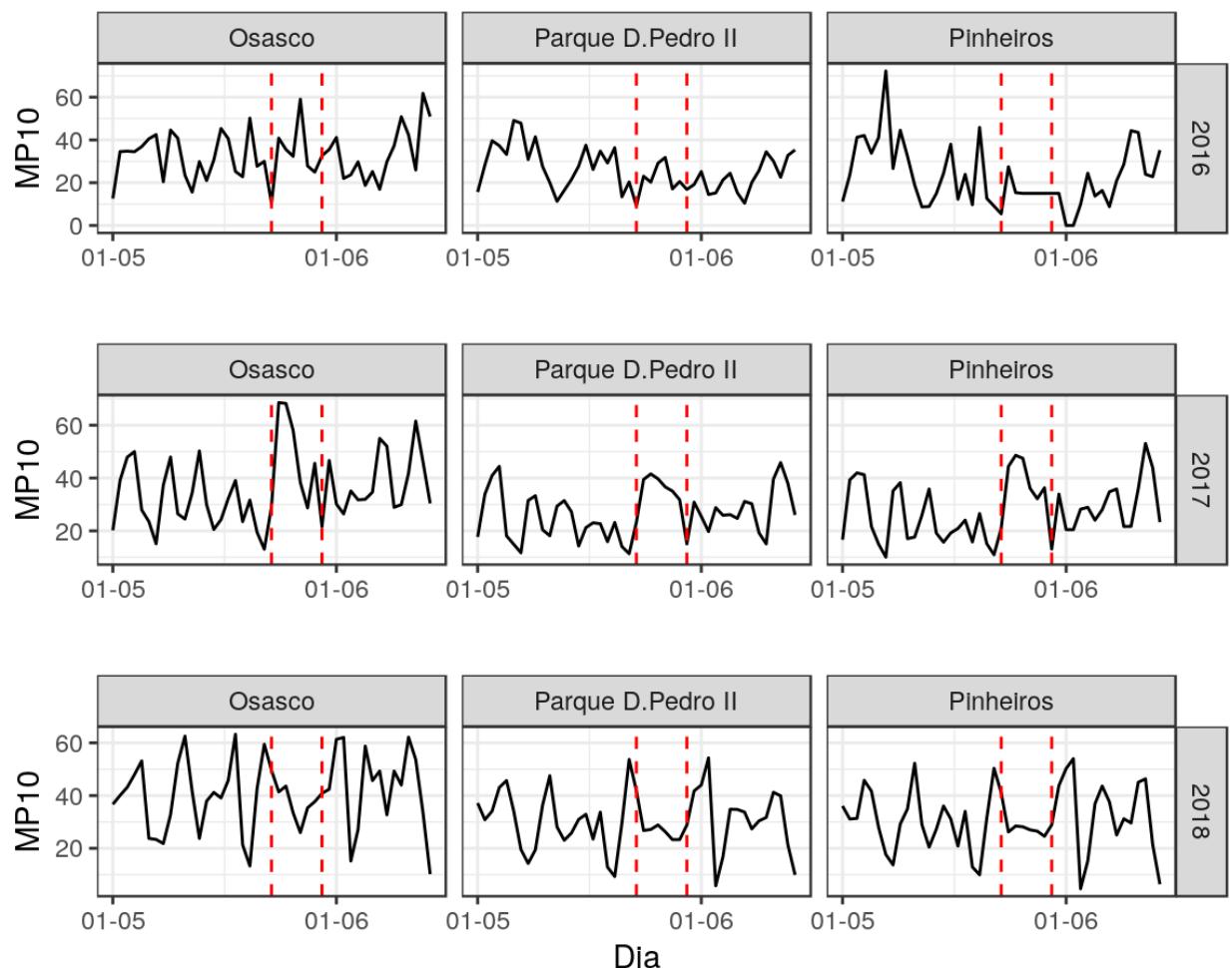


Figura 2.21: Série observada para o material particulado (PM10). O intervalo entre as retas pontilhadas corresponde ao período de paralisações.

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 $\mathbf{X} = (X_1, \dots, X_p)$ são transformadas em uma variável resposta Y^1 (Figura 3.1). Os dados são o resultado desse processo (Breiman, 2001).

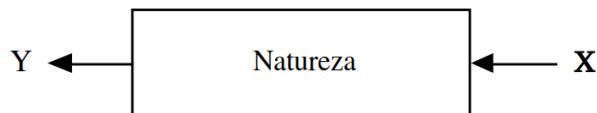


Figura 3.1: Esquematização do mecanismo gerador dos dados.

No contexto da modelagem estatística supervisionada² (Hastie *et al.*, 2008), dada a variável resposta Y e o vetor de variáveis explicativas $\mathbf{X} = (X_1, \dots, X_p)$, queremos encontrar funções f 's tais que

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

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 previsões — 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 qualquer classe de modelo supervisionado, a depender da escolha de $f(\cdot)$. De uma forma geral, modelos estatísticos são simplificações da realidade e, por isso, estão

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

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

sujeitos a erros. Quando modelamos a série de um poluente, por exemplo, estamos supondo que a sua concentração pode ser aproximada por uma função matemática de variáveis explicativas ao longo do tempo. Neste caso, o erro total do modelo quantifica o quanto a nossa função se afasta do verdadeiro mecanismo gerador do poluente. Parte desse erro é irreduzível 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. A parte redutível do erro é minimizada pela escolha adequada do modelo utilizado, o que torna essencial o desenvolvimento de estratégias de modelagem que contemplam as particularidades de cada estudo. Nesse sentido, podemos reescrever (3.1) como

$$Y = f(\mathbf{X}) + \epsilon, \quad (3.2)$$

sendo ϵ o erro irreduzível, que representa toda a informação de Y que não pode ser explicada pelos preditores \mathbf{X} . O nosso objetivo na modelagem estatística é encontrar a função $f(\cdot)$ que melhor se aproxime de Y , minimizando assim o erro redutível. Esse processo de encontrar a melhor $f(\cdot)$ é chamado de estimação.

Na prática, há duas abordagens bastante utilizadas na especificação da função $f(\cdot)$. A primeira consiste em especificar a forma como as variáveis Y e \mathbf{X} se relacionam a partir de parâmetros, que estrutura a função $f(\cdot)$ e reduz o problema de estimação a encontrar os parâmetros que reduzem o erro do modelo. Essa abordagem é conhecida como modelagem *paramétrica* e pode ser *probabilística*, quando o processo de estimação dos parâmetros faz suposições sobre a distribuição da variável resposta, ou não probabilística, quando nenhuma suposição sobre a distribuição de Y é feita. Como exemplo de modelos paramétricos probabilísticos, podemos citar os modelos de regressão (ver próximas seções), e de modelos não probabilísticos, podemos citar o SVM (Hastie *et al.*, 2008).

A segunda abordagem é mais flexível e permite que os próprios dados definam uma estrutura para $f(\cdot)$. Neste caso, os próprios dados definem a forma da função, usualmente a partir de divisões no espaço formado pelas variáveis \mathbf{X} . Essa abordagem é chamada de modelagem não paramétrica, tendo os modelos de árvores como principais representantes (discutiremos esses modelos no Capítulo 4).

Uma outra de se pensar a modelagem estatística foi introduzida por Breiman (2001), que dividiu os modelos em duas classes: *data models* e *algorithmic models*. A primeira classe representa os modelos cuja a forma de $f(\cdot)$ é conhecida e definida a priori. Esses modelos geralmente fazem suposições fortes sobre a relação entre Y e \mathbf{X} e são utilizados em especial para inferência, pois produzem resultados interpretáveis. O maior exemplo dentre dessa classe são os modelos paramétricos probabilísticos. A segunda classe não está preocupada com a forma de $f(\cdot)$, mas apenas em encontrar um algoritmo que encontre a melhor estimativa de Y para novos valores de \mathbf{X} . Esses modelos são quase sempre não interpretáveis, conhecidos como *modelos caixa-preta*, e muito utilizados para predição. Os exemplos mais famosos de *algorithmic models* são os modelos de árvores e as redes neurais (Goodfellow *et al.*, 2016).

Por estarem melhor estabelecidos dentro da comunidade científica e pela facilidade de interpretação, os modelos historicamente mais utilizados em trabalhos científicos são os *data models* ou, mais especificamente, os paramétricos probabilísticos. Neste capítulo, introduziremos os principais modelos paramétricos probabilísticos utilizados na literatura para análise de dados de poluição do ar, como o modelo de regressão linear, os modelos aditivos e os modelos de séries temporais. No

Capítulo 4, apresentaremos alguns modelos caixa-preta e como utilizá-los no contexto de 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 menos restritivos, 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, utilizaram 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) utilizaram para associar os níveis de partículas finas e ozônio com a proporção de carros a álcool e gasolina na cidade de São Paulo.

Apesar da sua popularidade, a complexidade presente nos estudos de poluição atmosférica, como relações não-lineares entre as variáveis, pode desqualificar o modelo de regressão linear como a opção mais adequada para o ajuste dos dados. No entanto, 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, $\mathbf{X}_t = (X_{1t}, \dots, X_{pt})$ um vetor de variáveis explicativas cuja associação com Y_t estamos interessados em avaliar e $t = 1, \dots, n$ o instante no tempo no 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 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$, o modelo de regressão linear pode ser definido por

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

Em geral, supomos que os erros $(\epsilon_1, \dots, \epsilon_n)$ tenham média zero, variância constante (homoscedasticidade) e sejam não-correlacionados³. 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 induzida 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).

³A 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.

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 β_1, \dots, β_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 estimativa por máxima verossimilhança (Casella e Berger, 2001). Sob a suposição de que Y segue uma distribuição Normal, as duas abordagens são equivalentes.

Como o modelo de regressão linear não exige que as observações sejam equidistantes no tempo, podemos utilizá-lo 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 neste caso.

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 $t = 1, \dots, n$, podemos definir os resíduos como

$$r_t = Y_t - \hat{Y}_t, \quad (3.4)$$

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 no contexto de previsão.

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

3.1.2 Incorporando tendência e sazonalidade

Como vimos nos exemplos do Capítulo 2, é comum séries de poluição do ar apresentarem tendência (positiva ou negativa) e diversos tipos de sazonalidade (diária, semanal, anual etc). Fatores como crescimento populacional, industrialização, aumento da frota de automóveis, leis de regulamentação de combustíveis, estações do ano, 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.

Para controlar esses componentes, podemos modelar a tendência e a sazonalidade da série incluindo preditores no modelo de regressão linear⁴. 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.

⁴Em 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.

Para acrescentar um termo de tendência linear ao modelo (3.3), podemos especificar $X_{1t} = t$, $t = 1, \dots, n$. 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.

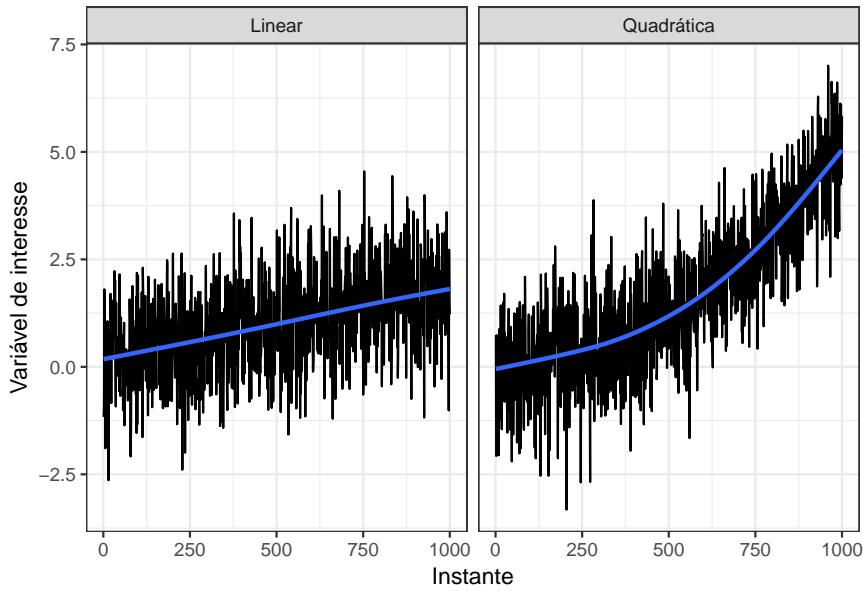


Figura 3.2: Exemplos de séries com tendência linear e quadrática, ambas positivas.

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:

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

e

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

A presença de sazonalidade 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. Quando não conseguimos controlar esses fatores diretamente, precisamos incluir no modelo termos que expliquem a sazonalidade.

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} , $i = 1, \dots, 11$ tais que

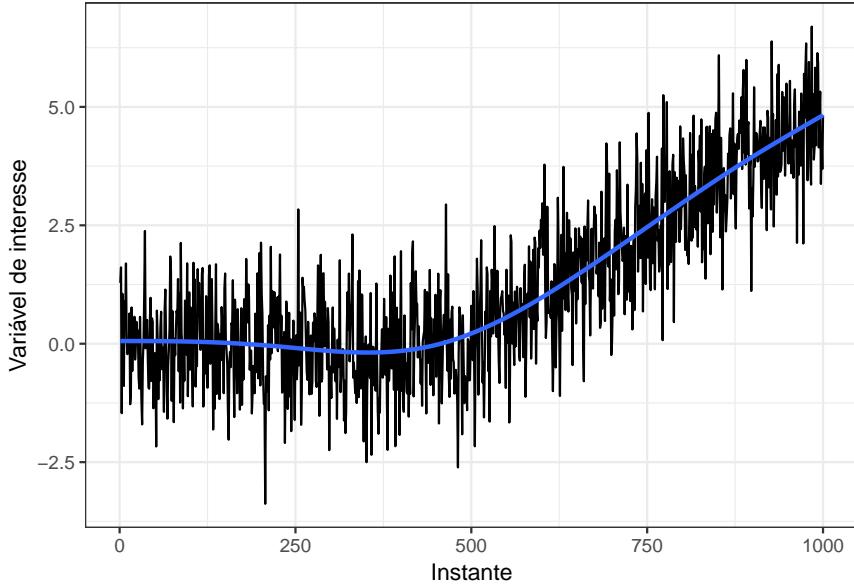


Figura 3.3: Exemplos de uma série com tendência não-constante.

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

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).

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 processo de estimativa do modelo de regressão linear supõe que os erros $(\epsilon_1, \dots, \epsilon_n)$ sejam não-correlacionados. Isso significa que, dados os preditores, as variáveis Y_1, \dots, Y_n devem ser independentes. 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 autocorrelacionados 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.

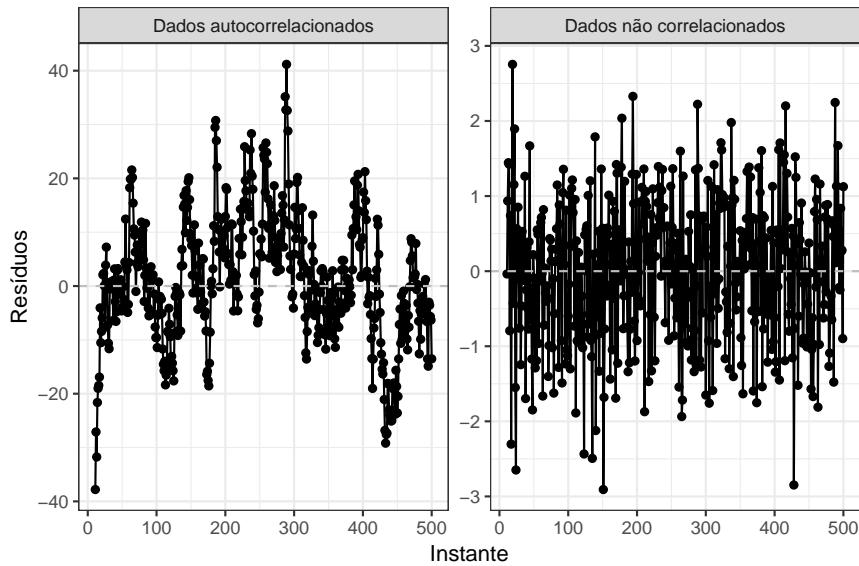


Figura 3.4: Comparação entre os gráficos dos resíduos de um modelo linear contra o tempo para dados auto-correlacionados e dados não correlacionados.

Se as observações são muito correlacionadas, os erros-padrão estimados pelo modelo de regressão linear poderão ser muito diferentes dos verdadeiros erros, o que comprometeria a inferência, já que testes de hipóteses acerca dos parâmetros dependem diretamente dessas estimativas. 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 a 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.

Em estudos de poluição do ar, observações correlacionadas não costumam ser um grande empecilho. Mesmo quando estamos trabalhando com séries horárias, geralmente conseguimos controlar a maioria dos fatores que induzem correlação, como condições climáticas ou trânsito. Quando não temos informação de algum preditor importante ou a correlação dos erros aleatórios ainda é alta, podemos adicionar uma variável indicadora para a hora do dia, por exemplo, para capturar a parte da variabilidade de Y não explicada por \mathbf{X} que está variando por hora e induzindo a correlação. A depender dos objetivos do estudo, também podemos agregar os dados para reduzir o efeito da autocorrelação. 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), pois esperamos que autocorrelação entre as concentrações médias de dois dias consecutivos seja consideravelmente menor. Essa foi a estratégia adotada por Salvo e Geiger (2014) para controlar a autocorrelação nas concentrações horárias de ozônio. Já Salvo *et al.* (2017), para reduzir o efeito da correlação entre medidas de uma mesma estação de monitoramento na estimativa da variabilidade dos coeficientes, 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 contemplam observações correlacionadas, como os modelos mistos (Demidenko, 2013; McCulloch e Searle, 2001). Falaremos brevemente destes modelos na Seção 3.6.1

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.

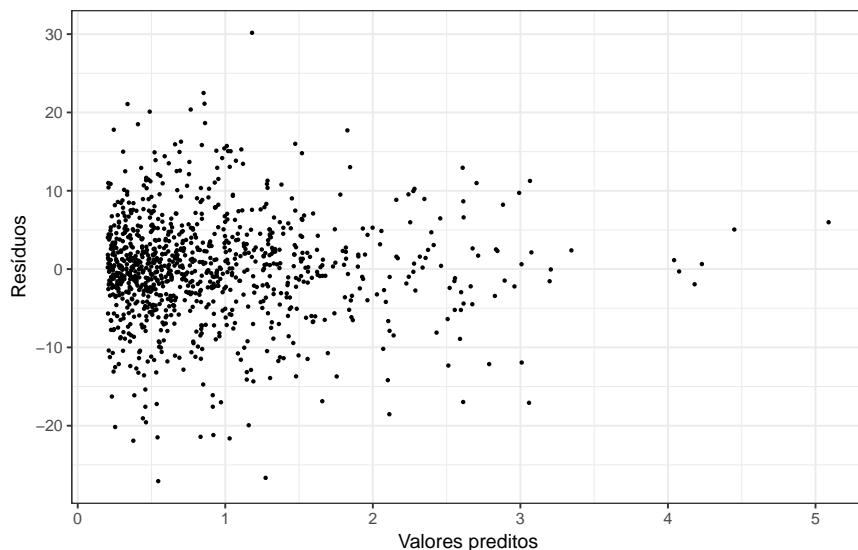


Figura 3.5: Gráfico dos resíduos contra os valores preditos. Exemplo de nuvem de pontos em forma de funil, indicando heteroscedasticidade.

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 a variância das observações.

3.1.5 Contornando a suposição de linearidade

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

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \quad t = 1, \dots, n. \quad (3.6)$$

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

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t, \quad t = 1, \dots, n, \quad (3.7)$$

sendo \hat{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 variarmos o valor de X_t em uma unidade, \hat{Y}_t vai variar $\hat{\beta}_1$ unidades, não importa qual seja o valor de X_t ($\hat{\beta}_1$ determina a inclinação da reta). Essa associação entre \hat{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 (\hat{Y}_t, X_{it}) se mantivermos as outras variáveis fixadas.

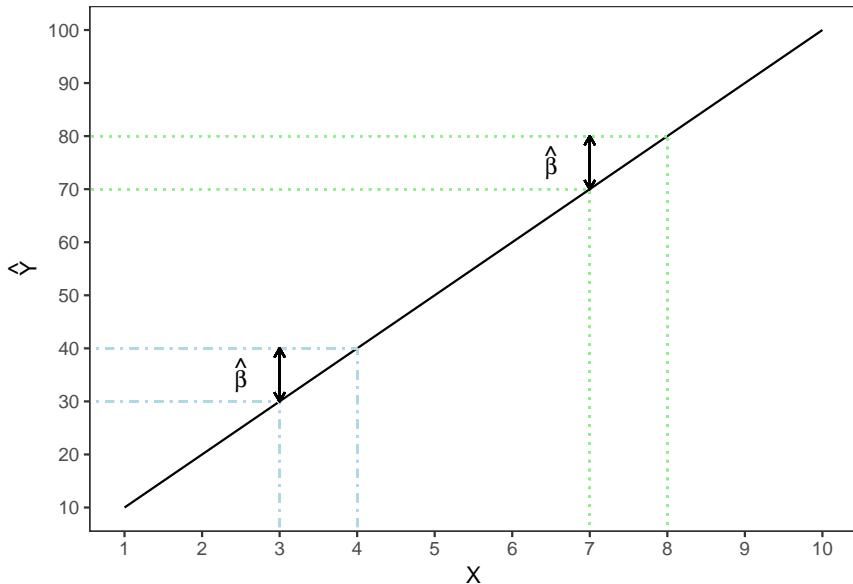


Figura 3.6: A estimativa $\hat{\beta}$ representa a variação em Y quando acrescemos X em uma unidade, não importando o valor de X .

Repare que a suposição de linearidade é mais forte do que apenas monotonicidade. Enquanto a monotonicidade restringe que a associação entre as variáveis seja sempre crescente ou decrescente, a linearidade 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

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).

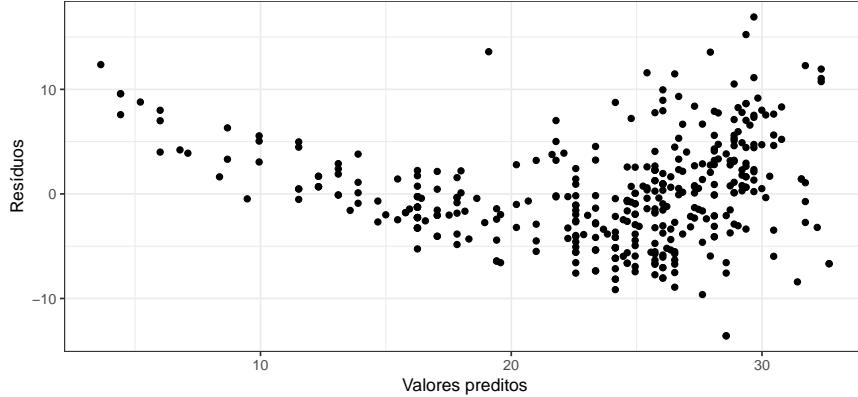


Figura 3.7: Gráfico dos resíduos contra os valores preditos, um exemplo de nuvem de pontos em forma de “U”, indicando não-linearidade.

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

$$Y_t = \beta_0 + \beta_1 T(X_t) + \epsilon_t, \quad t = 1, \dots, n, \quad (3.8)$$

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

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

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.

Modelos de regressão segmentada (Muggeo, 2003) são outra alternativa para ajustar relações não-lineares. Vamos supor que, para um determinado local, as médias diárias da concentração de ozônio e da temperatura sejam relacionadas como na gráfico da esquerda da Figura 3.8. O ajuste de um modelo de regressão segmentada a exemplo consiste em estimar um ou mais pontos de corte para a temperatura e, em cada região formada, ajustar uma reta de regressão com inclinação possivelmente diferente. O gráfico da direita apresenta as retas de regressão segmentada para um ponto de corte na temperatura.

A especificação do modelo, além de permitir que a inclinação das retas ajustadas mude em cada ponto de corte, possui parâmetros interpretáveis, ao contrário do modelo polinomial e dos modelos aditivos (ver Seção 3.3). A ideia por trás do algoritmo de estimação consiste em achar os pontos de corte no preditor que melhor representem mudanças na relação preditor/resposta.

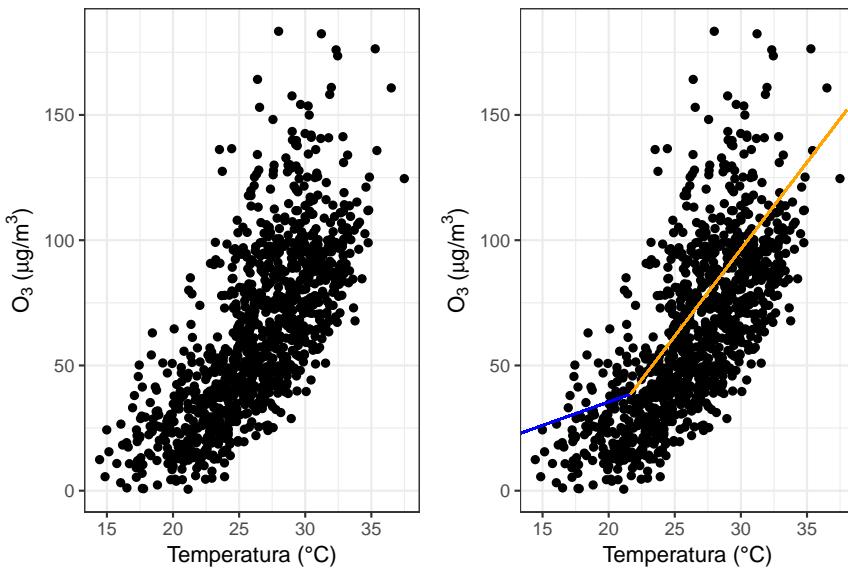


Figura 3.8: Exemplo de regressão segmentada. À esquerda, o gráfico de concentrações médias diárias de ozônio pela temperatura média diária. À direita, um modelo de regressão segmentada ajustada aos pontos com um ponto de corte.

Isso é feito utilizando uma técnica de *linearização* do modelo utilizando a expansão de Taylor de primeira ordem (ver Muggeo (2003) para mais detalhes). Dessa forma, o ponto de corte passa a ser um parâmetro do modelo e a dificuldade computacional é a mesma de um ajuste de um modelo de regressão linear.

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

3.1.6 Contornando 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.

Alguns modelos, como os de árvore (ver Capítulo 4), lidam com interações de forma mais natural, pois fazem divisões no hiperespaço gerado pelas variáveis explicativas. Isso permite que a função estimada capture relações muito mais complexas entre as variáveis sem precisarmos especificá-las.

3.1.7 Avaliando a qualidade do ajuste

Como discutido na introdução deste capítulo, modelos sempre estarão sujeitos a erros. Assim, mesmo quando os dados não violam as suposições estabelecidas pelo modelo, precisamos verificar se o modelo escolhido se ajustou bem aos dados. Para modelos de regressão linear, isso pode ser feito a partir da raiz do erro quadrático médio (RMSE⁵) 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 reta de regressão. Valores baixos de RMSE significam que $\hat{Y}_t \approx Y_t$, para $t = 1, \dots, n$, sugerindo que o modelo está bem ajustado aos dados. Como essa medida depende da magnitude da variável resposta, não podemos definir qual valor configura o que é um RMSE pequeno.

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 análise 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.

Para avaliar se um preditor está ou não melhorando o ajuste o suficiente para mantermos no modelo, podemos utilizar versões do RMSE e do R^2 penalizadas pelo número de parâmetros, conhecidas como RMSE e R^2 ajustados. 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. No entanto, repare que ainda estamos calculando o erro de ajuste, que geralmente é uma estimativa ruim do erro que vamos cometer ao generalizar o modelo para a população. No Capítulo 4, discutiremos os conceitos de *erro de treino*, *erro de teste* e *validação cruzada*, essenciais para evitar o sobreajuste.

Em estudos inferenciais, o coeficiente R^2 costuma ser 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

⁵Sigla 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.

aquele com maior R^2 como o modelo final. O RMSE é bastante utilizado na avaliação de modelos preditivos.

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

Ao assumirmos um modelo probabilístico para os dados, estamos supondo que as observações no mundo real se comportam 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. Quando trabalhamos com amostras pequenas, não podemos garantir a validade dos resultados assintóticos, e então supomos 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.

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

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

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

sendo⁶ \mathcal{D} uma distribuição pertencente à família exponencial⁷, $g(\cdot)$ uma função de ligação, μ_t um parâmetro de posição e ϕ um parâmetro de precisão⁸.

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 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 3.9 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.

Uma alternativa para a distribuição Gama é a Normal inversa. Considere agora $Y \sim \text{NI}(\mu, \phi)$,

⁶ A notação $Y_t | \mathbf{X}_t \stackrel{\text{ind}}{\sim} \mathcal{D}(\mu_t, \phi)$ significa que, conhecido os valores dos preditores \mathbf{X}_t , as variáveis Y_1, \dots, Y_n são independentes e seguem a distribuição \mathcal{D} , governada pelos parâmetros μ_t e ϕ .

⁷ A família exponencial corresponde a uma classe de distribuições de probabilidade que, sob certas condições de regularidade, apresentam certas 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).

⁸ Se ϕ é 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 .

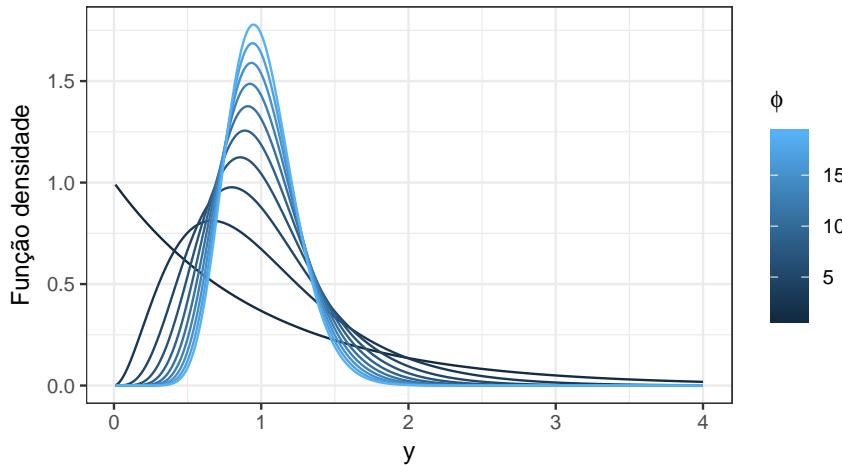


Figura 3.9: Função densidade da distribuição Gama com $\mu = 1$ e diversos valores de ϕ . Conforme ϕ aumenta, a distribuição se torna menos assimétrica, centralizando-se ao redor da média.

novamente sendo $\mu > 0$ a média de Y e $\phi > 0$ um parâmetro de precisão. Podemos ver pela Figura 3.10 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.

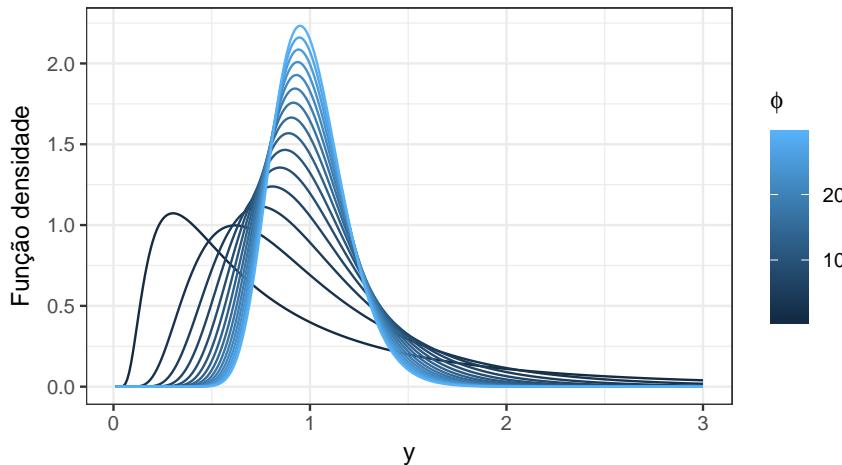


Figura 3.10: Função densidade da distribuição Normal inversa com $\mu = 1$ e diversos valores de ϕ . Conforme ϕ aumenta, a distribuição se torna menos assimétrica, centralizando-se ao redor da média.

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 (1992b), 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, SO₂ 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 fixado⁹. Essa taxa representa o valor médio¹⁰ 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 média¹¹. Uma alternativa nesse caso é a utilização de modelos com resposta Binomial Negativa.

Se $Y \sim BN(\mu, \phi)$, 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 impor-

⁹Esse intervalo de tempo se refere à frequência com que os dados são coletados, isto é, se as séries são diárias, semanais, mensais, anuais etc.

¹⁰A distribuição de Poisson atribui maiores probabilidades aos valores próximos à média λ .

¹¹Para o modelo Poisson, $\phi = 1$.

tante para a implementação de medidas para a redução da poluição do ar pois quantifica de forma objetiva o efeito dos poluentes na saúde pública.

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 resposta é muito melhor representada por curvas senoidais do que por retas.

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 estimativa 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 $t = 1, \dots, n$. O modelo aditivo generalizado pode ser escrito como

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

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

sendo \mathcal{D} uma distribuição pertencente à família exponencial e f_i , $i = 1, \dots, p$, 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, \dots, 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 priori¹². Curvas muito suaves podem ser muito restritivas, enquanto curvas muito *rugasas* 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, \dots, 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.

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

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \quad t = 1, \dots, T. \quad (3.11)$$

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

$$Y_t = \beta_0 + \beta_1 b_1(X_t) + \beta_2 b_2(X_t) + \dots + \beta_k b_k(X_t) + \epsilon_t, \quad t = 1, \dots, T, \quad (3.12)$$

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 $b_j(X_t) = X_t^j$, $j = 1, \dots, k$.

Como uma tentativa para aumentar a flexibilidade da curva ajustada, podemos segmentar X e ajustar diferentes polinômios de grau d em cada um dos intervalos¹³. Cada ponto de segmentação é chamado de *nó*, e uma segmentação com k nós gera $k+1$ polinômios. Na Figura 3.11 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 $b_1(X), b_2(X), \dots, b_k(X)$ 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 3.11 não é

¹²Uma maneira de determinar valores para esses parâmetros é utilizar validação cruzada, que será discutida no Capítulo 4.

¹³Em contrapartida ao modelo polinomial, que ajusta um único polinômio sobre todo o intervalo de variação de X .

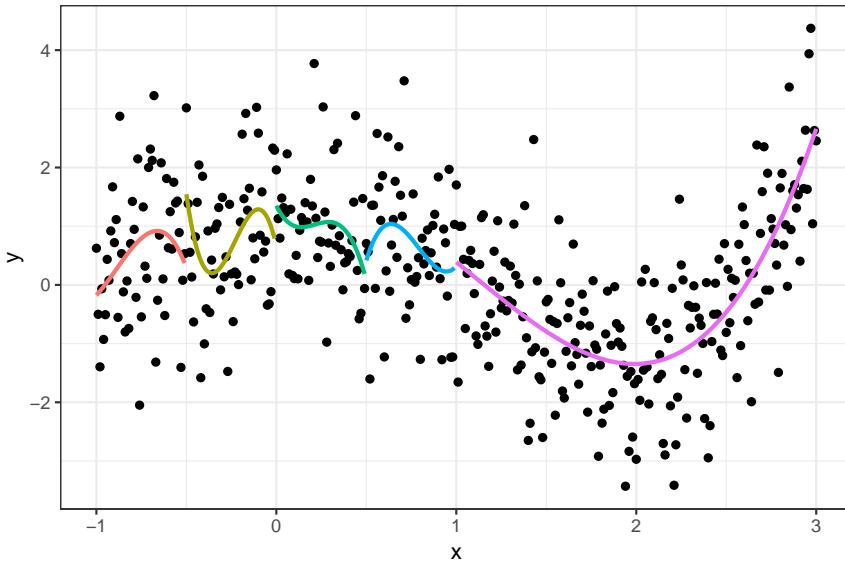


Figura 3.11: Polinômios de terceiro grau ajustados em cada segmentação da variável X . Os nós são os pontos $x = -0.5, x = 0, x = 0.5$ e $x = 1$.

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

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

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

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

sendo c_1, \dots, 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

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

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 quanto 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 séries temporais

Às vezes, queremos explicar a série Y apenas por seus valores defasados no tempo (autocorrelação) ou pelos valores defasados dos preditores X_1, \dots, X_p (correlação cruzada). Isso é feito principalmente em estudos de previsão, nos quais não temos o valor de preditores no instante futuro para alimentar o modelo. Imagine, por exemplo, que temos as concentrações médias de ozônio para os dias $1, 2, \dots, t$ e queremos construir um modelo para prever a concentração média no dia $t+1$. Se incluirmos como preditor nesse modelo a temperatura média contemporânea, vamos precisar da temperatura média do dia $t+1$ para prever a concentração de ozônio, sendo que não teremos essa informação no dia t .

O maior objetivo dos modelos de séries temporais é ajustar os componentes temporais de uma série (autocorrelação, tendência e sazonalidade), gerando um bom modelo para previsão. 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}, \dots, 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

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + w_t, \quad (3.14)$$

sendo ϕ_1, \dots, ϕ_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 zero¹⁴.

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 $w_t, w_{t-1}, \dots, w_{t-q}$. Formalmente, o modelo de médias móveis de ordem q , MA(q), é definido como

$$Y_t = w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}, \quad (3.15)$$

sendo $\theta_1, \dots, \theta_q$ 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

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}, \quad (3.16)$$

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 estimativa dos parâmetros (ϕ_1, \dots, ϕ_p) e $(\theta_1, \dots, \theta_q)$ 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.

¹⁴Se 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 $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$ ao modelo (3.14).

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 $\Delta Y_t, t = 2, \dots, n$, 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 tempo¹⁵.

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.

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.

¹⁵Cuidado 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).

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 <i>lag</i> q	Calda longa
PACF	Desaparece após o <i>lag</i> p	Calda longa	Calda longa

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).

Na linguagem R, uma maneira conveniente de se ajustar um modelo ARIMA é utilizar a função `auto.arima()` do pacote `forecast`. O algoritmo construído nessa função retorna o melhor modelo ARIMA com base em alguma métrica de qualidade de ajuste do modelo¹⁶, ajustando todas as combinações de valores para p , d e q segundo alguma restrição (combinação de todas as ordens menores que 5, por exemplo).

3.5 Modelos não-supervisionados

Nas seções anteriores, discutimos alguns modelos supervisionados, nos quais uma variável resposta *supervisiona* o aprendizado sobre fenômeno de interesse. Em alguns problemas, nós não temos acesso à variável resposta, nos restando buscar informação sobre o fenômeno de interesse apenas a partir da correlação entre os preditores. Essa estratégia é chamada de *análise não-supervisionada* e é geralmente utilizada para realizar agrupamentos e redução de dimensionalidade.

Em estudos de poluição do ar, modelos não-supervisionados são utilizados principalmente para a detecção de fontes de poluentes, isto é, dadas as medidas de concentração de diversos poluentes ao longo de um período, formamos grupos com as emissões mais correlacionadas e, a partir de inventários de poluição e conhecimento teórico, identificamos quais fontes podem ser representadas por cada grupo (Buhr *et al.*, 1992; Chavent *et al.*, 2009; Thurston e Spengler, 1985). Em estudos epidemiológicos, essas técnicas também podem ser utilizados para determinar quais doenças são as principais causas de internação em hospitais (Tecer, 2009).

A seguir, apresentaremos a análise de componentes principais e a análise fatorial, dois modelos não-supervisionados bastante utilizados em estudos de poluição do ar.

3.5.1 Análise de componentes principais

Suponha que queiramos investigar a concentração de dois poluentes, digamos X_1 e X_2 . Dada uma amostra de tamanho n dessas variáveis, para explorar esses dados descritivamente, poderíamos construir um gráfico de dispersão de X_1 contra X_2 e, a partir dele, observar tanto a variabilidade desses poluentes quanto como eles estão correlacionados. Se eles apresentarem correlação positiva, teríamos indícios de que eles são gerados pela mesma fonte ou sob as mesmas condições atmosféricas.

Suponha agora que, em vez de 2 poluentes, tivéssemos 10. Para construir gráficos de dispersão para todas as combinações dois a dois, precisaríamos analisar 45 gráficos, sendo que cada um deles só traria uma pequena parte da informação contida nos dados, pois estaríamos ignorando possíveis interações entre as variáveis.

¹⁶As opções disponíveis são AIC, AICc ou BIC.

Em geral, para p poluentes, gostaríamos de uma maneira de visualizar o máximo possível da informação contida no espaço p -dimensional gerado pelos preditores X_1, \dots, X_p em uma representação (gráfica) com poucas (duas) dimensões. Esse é o objetivo da análise de componentes principais.

Dado um conjunto de preditores X_1, X_2, \dots, X_p , a análise de componentes principais visa encontrar uma projeção ortogonal Z_1, Z_2, \dots, Z_p , tal que

$$VAR(Z_1) \geq VAR(Z_2) \geq \dots, VAR(Z_p).$$

Isso implica que, em geral, com apenas as primeiras variáveis Z_1, Z_2, \dots, Z_p , digamos Z_1 e Z_2 , podemos explicar a maior parte da variabilidade dos preditores X_1, X_2, \dots, X_p . Assim, Z_1 e Z_2 representariam em apenas 2 dimensões a maior parte da informação contida nos dados originais.

Cada variável Z_i , chamada de i -ésima componente principal, é uma combinação linear dos preditores X_1, X_2, \dots, X_p , isto é,

$$Z_i = \phi_{1i}X_1 + \phi_{2i}X_2 + \dots + \phi_{pi}X_p, \quad (3.17)$$

sendo que os elementos $\phi_{1i}, \dots, \phi_{pi}$ são os *pesos* do i -ésimo componente principal. Como esses pesos são normalizados, $\sum_{j=1}^p \phi_{ji}^2 = 1$, temos que $\phi_{ji} < 1$, para todo $j = 1, \dots, p$. Assim, os $\phi_{1i}, \dots, \phi_{pi}$ próximos de 1 indicam preditores positivamente associados e cuja variabilidade está sendo representada por Z_i .

Como Z_1, Z_2, \dots, Z_p representa uma projeção ortogonal, cada par de componentes (Z_i, Z_j) é não-correlacionado. Dessa forma, o componente Z_2 , por exemplo, é a combinação linear de X_1, X_2, \dots, X_p de maior variância entre todas as combinações lineares que são não-correlacionadas com Z_1 . Isso quer dizer que as fontes de variação representadas por Z_2 são não-correlacionadas com as encontradas em Z_1 .

Voltando ao nosso exemplo com os poluentes, se a análise de componentes principais indicasse os poluentes X_1, X_3 e X_5 como aqueles com maiores pesos para o componente Z_1 , então saberíamos que esses são os poluentes que mais contribuem com a variação total dos dados e poderíamos estudar o que causa essa variação. Da mesma forma, se os poluentes X_2, X_3 e X_4 são aqueles com maior peso para o componente Z_2 , então sabemos que a causa da variabilidade desses poluentes é não-correlacionada com a anterior.

Os cálculos por trás da análise de componentes principais envolvem decomposição espectral (Nicholson, 2001), uma técnica de álgebra linear para decompor matrizes em função de seus auto-vetores e autovalores.

Na linguagem R, podemos realizar uma análise de componentes principais utilizando a função `prcomp()` do pacote `stats`.

3.5.2 Análise Fatorial

Assim como a análise de componentes principais, a análise fatorial também pode ser utilizada para redução de dimensionalidade. A segunda técnica difere da primeira em dois pontos principais. Primeiro, a análise fatorial supõe que a variância e covariância contida em um conjunto de variáveis X_1, X_2, \dots, X_p pode ser explicada por um conjunto menor de fatores latentes (não-observáveis). Se esses fatores são definidos a priori, o modelo pode ser utilizado para testar teorias sobre a relação entre os fatores e as variáveis observadas. O segundo ponto diz respeito à inclusão de erros aleatórios.

Enquanto a análise de componentes principais calcula um novo conjunto de variáveis que explica 100% da variabilidade das variáveis originais, a análise factorial considera que parte da variabilidade das variáveis originais pode ser explicada pelos fatores latentes, mas uma outra parte é devida a ruído aleatório.

A análise factorial é geralmente utilizada para avaliarmos fontes de poluição, sendo que cada fator representa uma fonte diferente. Podemos pensar em cada fonte como uma variável latente, que pode ser representada por uma combinação linear das concentrações observadas. A inclusão de erros aleatórios seria justificada, por exemplo, pela variação nas concentrações causadas por condições atmosféricas.

Dada uma amostra de tamanho n dos preditores X_1, X_2, \dots, X_p , a análise factorial procura estimar os pesos l_{ij} , $i = 1, \dots, p$ e $j = 1, \dots, m$, tais que

$$Z_{ik} = l_{i1}F_{1k} + \dots + l_{im}F_{mk} + \epsilon_{ik},$$

sendo que, para a k -ésima observação da amostra, $Z_{ik} = \frac{X_{ik} - \bar{X}_i}{\sigma_i}$ é o i -ésimo preditor normalizado, F_{jk} é o j -ésimo fator e ϵ_{ik} é o erro associado ao i -ésimo preditor.

O número de fatores m deve ser escolhido antes do ajuste, sendo que a interpretação de cada fator é externa ao modelo. Mesmo quando cada fator é construído teoricamente a priori, o modelo não indica qual dos termos F_{jk} representa cada fator. A identificação dos fatores é feita então a partir de conhecimento teórico sobre o fenômeno estudado.

Mais informações sobre análise factorial, como o processo de estimação e interpretação geométrica, podem ser encontradas em Everitt e Hothorn (2011). Na linguagem R, esse modelo pode ser ajustado utilizando a função `factanal` do pacote `stats`.

3.6 Outros modelos

3.6.1 Modelos mistos

Os modelos mistos (ou modelos de coeficientes aleatórios) foram introduzidos por Fisher (1918) para estudar a correlação de traços hereditários em pais, mães e filhos. Eles são particularmente úteis em estudos clínicos, ensaios biológicos e estudos sociais, nos quais variáveis medidas em uma mesma unidade amostral (medidas repetidas) ou unidades amostrais com agrupamentos naturais¹⁷ (dados hierárquicos) geram observações correlacionadas.

Essa classe de modelos utiliza *coeficientes aleatórios* para controlar os efeitos individuais e de grupo que não podem ser observados¹⁸. Ao contrário dos coeficientes fixos dos modelos de regressão vistos até aqui, que modelam a média da variável resposta, os efeitos aleatórios introduzem uma estrutura para a variância de Y , contemplando a correlação entre observações de um mesmo grupo e as diferentes variâncias de observações de diferentes grupos.

Em estudos de poluição do ar, modelos mistos são usados principalmente para controlar variações individuais em estudos epidemiológicos longitudinais. Liao *et al.* (1999), por exemplo, avaliaram a

¹⁷Como membros de uma mesma família, pacientes de um mesmo hospital, animais de uma mesma ninhada ou moradores de uma mesma região.

¹⁸Como carga genética, criação, o efeito do atendimento por um mesmo médico, níveis de poluição do ar de uma mesma região etc.

resposta autonômica cardiovascular a variações diárias de material particulado 2.5 em 26 idosos durante 3 semanas consecutivas, mostrando que o aumento dos níveis de MP2.5 estão associados a um menor controle autonômico cardíaca. Coull *et al.* (2001) utilizaram um modelo aditivo misto para associar mudanças na concentração de poluentes atmosféricos com a função pulmonar de crianças ao longo de 109 dias. Eles mostraram que o modelo misto era o mais adequado para contemplar a heterogeneidade populacional da suscetibilidade à poluição. Chuang *et al.* (2011) avaliaram o uso de um componente aleatório para os graus de liberdade de um modelo aditivo Poisson em estudos epidemiológicos de poluição do ar. Os autores observaram que esse modelo gerava erros-padrão maiores do que o modelo aditivo usual e concluíram que um alisamento variável da função não-linear em conjunto com essa maior variabilidade poderia refletir melhor a realidade. Já Kloog *et al.* (2012) utilizaram modelos mistos para prever concentrações diárias de material particulado (MP2.5) na costa leste dos Estados Unidos a partir de medidas de profundidade óptica de aerossóis feitas por satélites de 2000 a 2008, encontrando um bom modelo para prever exposições humanas a esse poluente tanto a curto quanto a longo prazo.

Mais informações sobre modelos mistos, como especificação do modelo, estimativa e outros exemplos, podem ser encontradas em Singer *et al.* (2012) e Galecki e Burzykowski (2013).

Na linguagem R, modelos mistos podem ser ajustados a partir das funções dos pacotes `nlme` e `lme4`.

3.6.2 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 $t = 1, \dots, n$, o modelo ARCH(q) é definido por

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

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_q \epsilon_{t-q}^2, \quad (3.18)$$

com $\alpha_0 > 0$ e $\alpha_i \geq 0$, $i > 0$, sendo $f(\mathbf{X}, \mathbf{Y})$ uma função dos preditores $\{(X_{1i}, \dots, X_{pi}), i \leq t\}$ e das variáveis defasadas (Y_1, \dots, Y_{t-1}) . Repare a primeira expressão de (3.18) 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

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

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \dots + \beta_p \sigma_{t-p}^2, \quad (3.19)$$

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). Na linguagem R, esse modelo pode ser ajustado usando a função `garch()` do pacote `tseries`.

Em estudos de poluição do ar, modelos GARCH são utilizados principalmente para aprimorar modelos de previsão, como feito por Kumar e Ridder (2010) para dados de ozônio em Bruxelas e Londres.

3.6.3 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 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

$$\begin{aligned} Y_t &= a_t U_t + v_t \\ U_t &= \phi U_{t-1} + w_t, \end{aligned} \tag{3.20}$$

sendo a_t e ϕ parâmetros do modelo e $w_t \sim N(0, \sigma_w)$. A primeira equação em (3.20) 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 $1, \dots, t-1$, 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).

Capítulo 4

Estratégias de machine learning

There are no routine statistical questions,
only questionable statistical routines.

— Sir David Cox

Nos últimos anos, a abordagem conhecida como *machine learning*¹ 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. Por trás de todo o marketing em volta desse termo, existe um conjunto de práticas e 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 maneira geral, essas suposições são feitas a partir de um modelo probabilístico para a variável resposta Y , cuja parametrização dependerá de alguma função dos preditores² \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 $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ (linearidade e aditividade);
- a variância de Y , σ^2 , é constante para todas as observações na população.

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

As técnicas e modelos utilizados para *machine learning* colocam a interpretabilidade em segundo plano e priorizam a produção de previsões o mais precisas possível para a quantidade sob estudo. As estratégias dentro dessa abordagem enfrentam a dualidade entre flexibilidade e sobreajuste, isto é, buscam entre modelos complexos (normalmente não interpretáveis) aquele que melhor se ajuste

¹Também conhecida como modelagem preditiva, aprendizado estatístico, aprendizagem automática ou aprendizado de máquina.

²Muita da literatura sobre *machine learning* 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*.

aos dados, mas que ainda possa ser generalizado para além da amostra. Embora essa visão não seja adequada para inferência, muitas das práticas podem ser incorporadas em estudos inferenciais na tentativa de encontrar modelos melhores ajustados e evitar o *overfitting*.

Neste capítulo, discutiremos com mais detalhes o conceito de sobreajuste, apresentando métodos de reamostragem, seleção de variáveis e regularização. Em seguida, introduziremos alguns modelos de árvores, bastante utilizados para modelagem preditiva devido a sua alta precisão. Por fim, apresentaremos alguns métodos gráficos para interpretar modelos caixa-preta.

4.1 Sobreajuste e o balanço entre viés e variância

Como discutido na introdução do Capítulo 3, nunca vamos encontrar uma função $f(\cdot)$ que relate perfeitamente a variável resposta Y e os preditores \mathbf{X} , pois estamos sempre sujeitos a dois tipos de erros: um erro redutível e outro irredutível. O erro redutível indica o quanto o modelo escolhido representa bem o fenômeno estudado e tem esse nome porque podemos sempre encontrar uma candidata para $f(\cdot)$ que se aproxime mais do processo gerador de Y . Reduzir ao máximo esse erro é o grande objetivo da modelagem.

No entanto, mesmo que conseguíssemos eliminar o erro redutível, nossas previsões não seriam perfeitas devido ao erro irredutível. Esse erro representa a parte da variabilidade de Y que não pode ser explicada pelos preditores \mathbf{X} , comumente chamado de erro aleatório ou ruído. Assim, por construção, todo modelo estatístico tem um erro associado.

O sobreajuste ou *overfitting* acontece quando, na tentativa de eliminar o erro redutível, acabamos eliminando também o erro irredutível. Conforme aumentamos a complexidade do modelo, podemos passar a explicar variações aleatórias, que não estão associadas aos preditores considerados. Assim, um modelo sobreajustado gera conclusões que os dados disponíveis não são capazes sustentar em um contexto mais amplo.

Essa dualidade entre qualidade do ajuste e capacidade de generalização é o maior paradigma da modelagem preditiva, e deveria ter atenção especial em qualquer contexto de modelagem, pois queremos sempre generalizar os resultados do modelo para a população de interesse. Na prática, gostaríamos de encontrar o melhor ajuste que generalize bem os resultados para a população, e essa tarefa pode ser resumida na minimização de duas quantidades: 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,

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

e calcular a raiz do erro quadrático médio (RMSE), 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 maior³, até encontrarmos o menor RMSE.

Na Tabela 4.1, apresentamos o RMSE obtido para os modelos de regressão polinomial até o nono grau. Observe que, conforme aumentamos a complexidade do modelo (grau do polinômio), o

³Esses são os modelos polinomiais apresentados na Seção 3.1.5. O modelo de regressão linear simples é um modelo polinomial de grau 1.

RMSE diminui, até chegar em 0 para o polinômio de grau 9. Se utilizarmos apenas 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 claramente não representa bem o comportamento dos pontos.

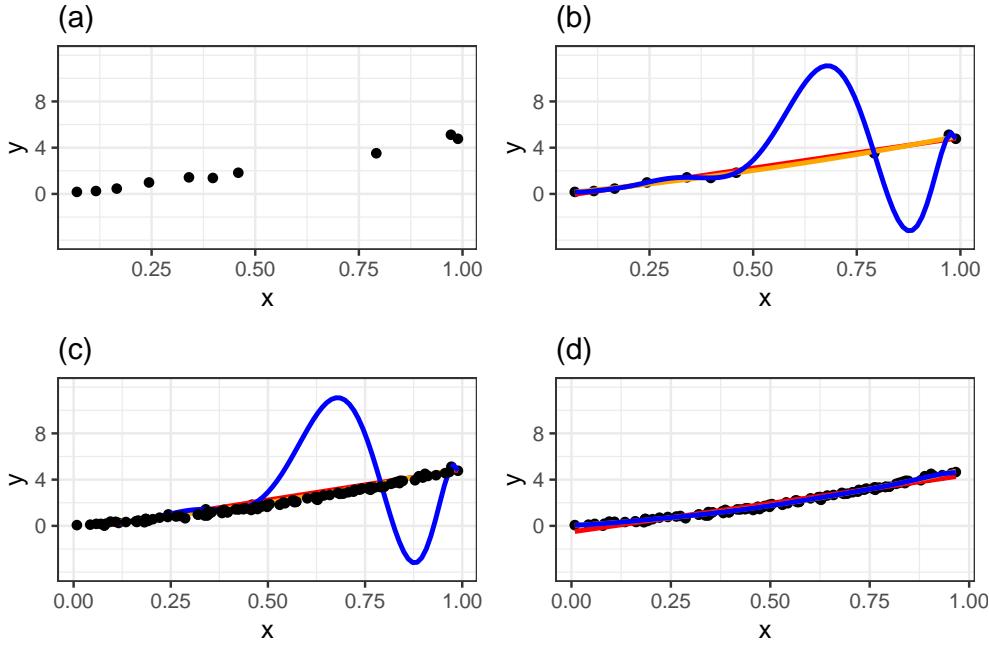


Figura 4.1: Exemplo do trade-off entre viés e variância. (a) Conjunto de 10 pontos que gostaríamos de ajustar. (b) Modelo de regressão linear simples (vermelho), modelo de regressão polinomial de grau 2 (amarelo) e modelo de regressão polinomial de grau 9 (azul), ajustados aos 10 pontos. (c) Amostra de 100 novas observações plotadas juntas das amostras polinomiais ajustadas 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 100 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

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) confirma 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 polinomial de grau 9 calculado nas 100 novas observações⁴ 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 também as 100 novas observações.

Como enfatizado anteriormente, estamos sempre em busca de modelos que se ajustem bem à amostra, mas que também possam ser generalizados para a população. Assim, chamaremos de *viés* o quanto o modelo ajustado está distante das observações da amostra e de *variância* o quanto o modelo mudaria se o ajustássemos em uma nova amostra. O viés representa o erro induzido por aproximar um fenômeno real, que pode ser extremamente complicado, por um modelo muito mais simples. Já a *variância* indica o quanto o nosso modelo erraria se o usássemos para predizer novas observações. Dizemos então que modelos mal ajustados apresentam alto viés e modelos com baixo poder de generalização apresentam alta variância.

É muito comum utilizarmos estratégias que se preocupam apenas com a minimização do viés. Essas estratégias elegem como boas escolhas modelos complexos, visando um ajuste cada vez melhor aos dados, sem levar em conta o quanto isso será representativo na população. No exemplo anterior, isso fica claro com o ajuste de polinômios de grau cada vez maior aos dados. O polinômio de grau 9 ilustra o conceito de sobreajuste, que 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 predores, não é possível visualizar graficamente o sobreajuste, como mostrado no exemplo. Por isso, na prática, pode não ser 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 o RMSE podem ser utilizados para avaliar a qualidade do ajuste de um modelo. Em alguns casos, podemos querer utilizar o erro absoluto médio (MAE, *mean square error*), que, ao contrário do RMSE, não dá mais peso para erros em valores muito altos da variável resposta.

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 calculá-la 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 do erro de teste. Uma alternativa nesses casos é utilizar métodos de reamostragem, que consiste em separar a base disponível em observações utilizadas para *treinar* o modelo e observações para estimar sua

⁴Aqui, os modelos não foram reajustados. Foram considerados os modelos ajustados apenas com as 10 primeiras observações

variância. Na próxima seção, apresentaremos dois métodos de reamostragem bastante utilizados: a *validação cruzada* e o *bootstrapping*.

4.2.1 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, $i = 1, \dots, n$, 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 incorridas ao utilizarmos a amostra de validação. No entanto, uma desvantagem 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.

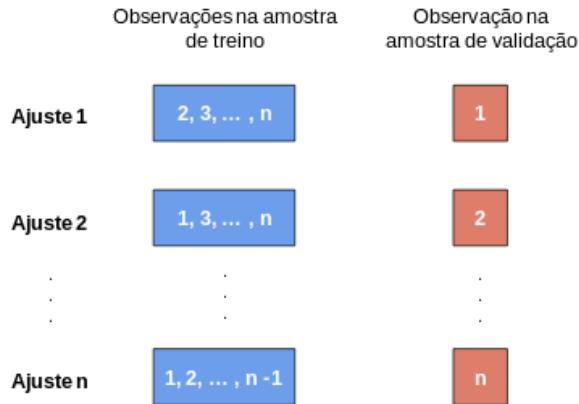


Figura 4.2: Esquematização da validação cruzada leave-one-out.

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 bastante utilizado (Figura 4.3).

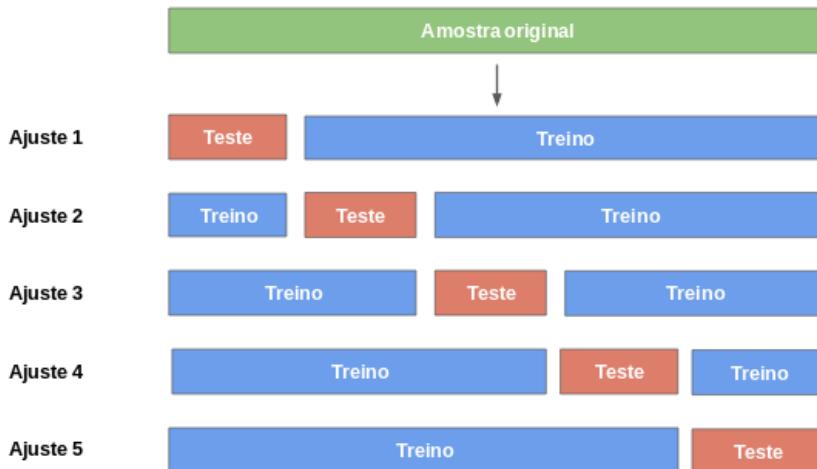


Figura 4.3: Esquematização da validação cruzada k -fold, com $k = 5$.

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 $k \ll n$. 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.

4.2.2 Bootstrapping

O *bootstrapping* é uma poderosa ferramenta 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. Em geral, o tamanho de cada amostra de *bootstrapping* tem o mesmo tamanho da base original.

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

4.3 Seleção de variáveis

Na especificação do modelo, muitas vezes 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, podemos pensar em estratégias para retirá-las da análise, aumentando a interpretabilidade dos resultados. Nesta seção, apresentaremos algumas técnicas de seleção de variáveis que podem ser utilizadas em qualquer classe de modelos estatísticos.

4.3.1 Selecionando o melhor subconjunto de preditores

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

1. Ajustar o modelo nulo, sem nenhum preditor.
2. Para $k = 1, \dots, p$, 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 BIC⁵.

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 e X_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.

4.3.2 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 adicionadas ou retiradas do modelo em cada passo. Os mais utilizados são o *forward stepwise* e o *backward stepwise*.

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

1. Ajuste o modelo nulo (M_0), sem preditores.
2. Ajuste todos os p modelos com 1 preditor e escolha o melhor⁶ (M_1).

⁵O AIC e o BIC são medidas da qualidade do ajuste penalizadas pelo número de parâmetros do modelo. Mais informações, consultar James *et al.* (2013)

⁶Maior R^2 , por exemplo.

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, \dots, M_p).
5. Escolha o melhor modelo entre M_0, M_1, \dots, 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 ($M_{p-1}, M_{p-2}, \dots, M_0$). Ao fim, escolhemos o melhor entre os modelos M_0, M_1, \dots, 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).

4.4 Regularização

Os métodos de seleção de subconjuntos de preditores apresentados na seção anterior diminuem a complexidade do modelo eliminando variáveis que não colaboram significativamente com a diminuição do viés, potencialmente diminuindo a variância. As técnicas de *regularização* apresentam uma ideia similar: diminuir a variância do modelo a partir de suavizações que introduzem um pouco de viés. Essas técnicas 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, sendo uma boa estratégia para evitar o sobreajuste. 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.

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 β_1, \dots, β_p os parâmetros do modelo $f(\cdot)$ e λ um hiperparâmetro⁷ 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 $\lambda \rightarrow \infty$, 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 $i = 1, \dots, p$, $\beta_i = 0$ 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

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

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.

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 cálculo do erro-padrão e o desenvolvimento de testes para 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*.

⁷Hiperparâmetros são parâmetros que não são estimados diretamente pelos dados.

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.

4.5 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. Dada a estimativa de 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 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.

4.6 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 previsão.

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 complexos. Generalizações desse modelo, 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*.

4.6.1 Árvores de decisão

As árvores de decisão se baseiam no particionamento das variáveis explicativas, de tal forma que as regiões formadas gerem previsões 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 espaço 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 $38 \mu/m^3$; dias com temperatura entre 27 e 29° C serão preditos com concentração de ozônio igual a $69 \mu/m^3$; e dias com temperatura maior de 92° C serão preditos com temperatura igual a 92° C. Além de conseguirmos prever 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.

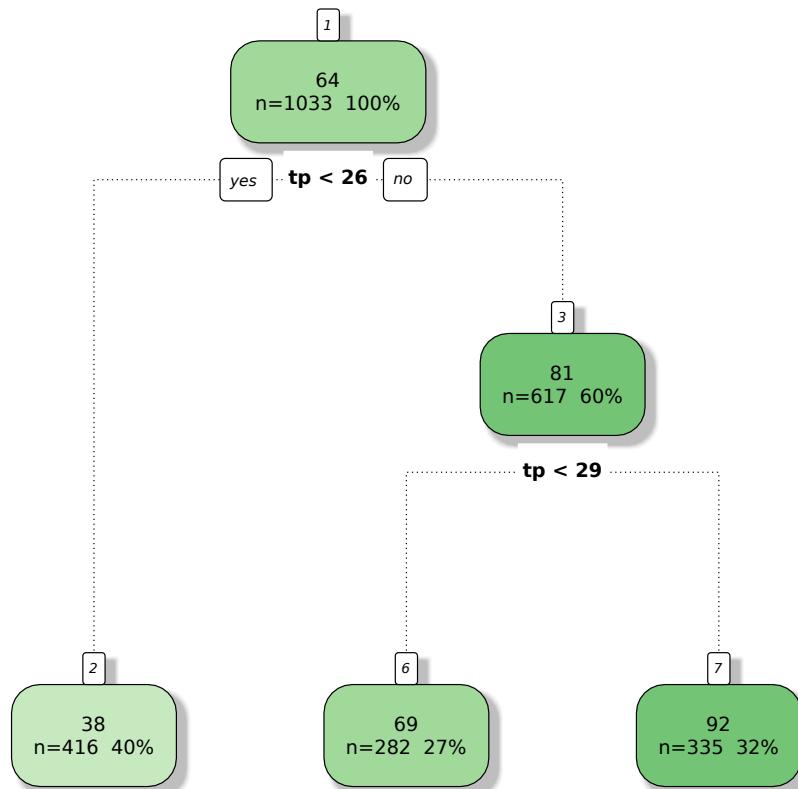


Figura 4.4: Exemplo de uma árvore de decisão para a concentração de ozônio explicada pela temperatura.

Observe que no exemplo temos 3 nós finais, mas, teoricamente, poderíamos continuar participando 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.

Na linguagem R, árvores de decisão podem ser ajustadas a partir da função `rpart()` do pacote `rpart`.

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.

4.6.2 Random Forests

As árvores de decisão apresentadas na seção anterior tendem a ter alta variância, isto é, se dividirmos nossa amostra em duas e aplicarmos o mesmo modelo de árvores de decisão em cada um, há uma grande chance de obtermos divisões diferentes no espaço amostral e, consequentemente, diferentes previsões.

Uma maneira de lidar com esse problema é gerar diferentes amostras de *bootstrapping*, ajustar o mesmo modelo em cada uma delas e utilizar previsões médias como estimativas final do modelo. Em um contexto geral, essa técnica é chamada de *bootstrapping aggregation* ou *bagging*. Ele pode ser utilizada para qualquer modelo e seu objetivo geral é diminuir a variância das estimativas.

As chamadas *random forests* Hastie *et al.* (2008) são uma aplicação do *bagging* para árvores de decisão. Neste modelo, além de utilizarmos uma amostra de *boosstraping* diferente para cada árvore ajustada, também fazemos uma seleção dos preditores para cada ajuste. Isso impede que as árvores ajustadas sejam muito correlacionadas e permite que preditores que seriam preteridos por serem "menos importantes" também ajudem a explicar a variação da variável resposta.

Em cada iteração do algoritmo,

1. sorteamos m dos p preditores, $m \leq p$;
2. geramos uma amostra de *boosstraping* a partir da base inicial;
3. ajustamos uma árvore de decisão utilizando os m preditores escolhidos em (1) e a amostra em (2).

Repetindo esse procedimento M vezes, teremos M estimativas para cada nova observação. A previsão final será a média dessas M estimativas (no caso de regressão) ou a classe mais frequente (no caso de classificação). Geralmente, m é escolhido como \sqrt{p} , mas esse hiperparâmetro pode ser definido utilizando validação cruzada. M por volta de 200 costuma ser suficiente para gerar bons resultados⁸.

Random forests podem ser ajustadas utilizando a função `train()` do pacote `caret` com `method = "ranger"`. Os principais hiperparâmetros que devem ser escolhidos são o número de sorteados em cada ajuste (m) e o número mínimo de observações nos nós finais de cada árvore.

⁸Para 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.

4.6.3 XGBoost

Assim como o *bagging*, apresentado na seção anterior, o *boosting* é uma abordagem geral aplicada a diversos modelos para aumentar seu poder preditivo. No contexto de árvores de decisão, a ideia aqui é construir árvores sequencialmente, sendo que cada árvore utiliza informações obtidas da árvore anterior. Ao contrário do *bagging*, o *boosting* não envolve amostras de *bootstrapping*, sendo que cada passo do algoritmo utiliza uma versão modificada dos dados originais.

A ideia por trás do *boosting* é *aprender devagar*. Em vez de tentarmos ajustar uma grande árvore em toda a base de dados, ajustamos pequenas árvores $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ sequencialmente, de tal forma que a árvore $\hat{f}^i(x)$ é ajustada utilizando os resíduos da árvore $\hat{f}^{i-1}(x)$. Considere $\hat{f}(x)$ a nossa função estimada final do modelo e que, inicialmente, $\hat{f}(x) = 0$. Em cada passo do algoritmo, a nova árvore $\hat{f}^i(x)$ é somada à função estimada $\hat{f}(x)$ utilizando-se um parâmetro de encolhimento $\lambda > 0$, isto é, no passo i , $\hat{f}(x)$ passa a ser $\hat{f}(x) + \lambda\hat{f}^i(x)$. Assim, os resíduos são atualizados em cada passo e $\hat{f}(x)$ melhora lentamente em regiões onde sua performance era ruim. O parâmetro de encolhimento λ , o número de árvores B e o tamanho de cada árvore (controlado pelo número de nós terminais) são os principais hiperparâmetros do modelo e podem ser escolhidos por validação cruzada.

O *gradient boosting* generaliza o *boosting* substituindo o ajuste dos resíduos pela minimização de uma função de custo $L(y, f(x))$. Em cada etapa, ajustamos uma árvore não mais aos resíduos, mas sim às quantidades

$$g_i = - \left[\frac{\delta L(y_i - f(x_i))}{\delta f(x_i)} \right]_{f=\hat{f}^{i-1}}, \quad i = 1, \dots, n.$$

As quantidades g_1, \dots, g_n representam o gradiente da função de custo em relação a cada árvore $f^i(x)$. Portanto, em cada etapa, minimizamos lentamente a função de custo a partir da nova árvore ajustada. Esse é algoritmo de minimização é chamado de *gradient descent* Kiefer e Wolfowitz (1952).

Por fim, o *XGBoost* (ou *extreme gradient boosting*) é uma implementação eficiente do *gradient boosting* que utiliza alguns "truques" para otimizar o processo de estimação, como penalização das árvores para controlar a velocidade de aprendizado, randomização dos parâmetros para diminuir a correlação entre as árvores e encolhimento dos nós terminais para diminuir o sobreajuste. Esse algoritmo é um dos mais utilizados em problemas preditivos complexos hoje em dia, estando entre os três modelos mais vencedores de competições de *machine learning*.

Assim como as *random forests*, esses modelos são não-interpretáveis e precisam de ferramentas como gráficos de dependência parcial ou o LIME para a avaliação da associação entre as variáveis.

Na linguagem R, podemos ajustar o *XGBoost* utilizando a função `train()` do pacote `caret` com o argumento `method = "xgbTree"` ou diretamente as funções do pacote `xgboost`.

4.7 Interpretando modelos caixa-preta

Ao contrário dos modelos apresentados no capítulo anterior, modelos como o *random forest* ou o *XGboost* não são interpretáveis. Isso implica que não conseguimos avaliar diretamente como cada preditor está associado com a variável resposta, impossibilitando que esses modelos sejam usados para inferência ou criando muita desconfiança a respeito do que esses modelos estão fazendo por trás das cortinas.

Neste capítulo, vamos apresentar algumas técnicas que, sob algumas suposições ou restrições, visam elucidar qual é a relação dos preditores com os valores preditos, abrindo um pouco a caixa-preta dos modelos de *machine learning*.

4.7.1 Gráfico de dependência parcial

O objetivo do gráfico de dependência parcial é mostrar o efeito marginal de um preditor no valor predito pelo modelo. A partir dele, podemos investigar qual é a forma e o sentido da relação entre cada preditor e a variável resposta.

Dado um modelo não-interpretável $f(X, Z)$, um preditor sob investigação X e um vetor Z representando os outros preditores do modelo, a ideia por trás da construção do gráfico de dependência parcial consiste em:

1. fixar alguns valores para X , digamos $x = (x_1, \dots, x_M)$;
2. para cada $x_i, i = 1, \dots, M$, repetir:
 - a) para cada uma das n observações na amostra, substituir o valor observado de X por x_i e calcular o valor predito do modelo, $\hat{f}(x_i, z)$;
 - b) calcular a média das n previsões, $\bar{f}_i(x_i) = \frac{1}{n} \sum_{j=1}^n \hat{f}(x_i, z_j)$;
3. plotar o gráfico dos valores x contra as médias das previsões $\bar{f}(x)$.

A função $\bar{f}(x)$ podem ser interpretada como uma estimativa da distribuição marginal das previsões que depende apenas do preditor X , obtida calculando-se a média das previsões sobre os outros preditores. Essa estratégia é razoável apenas se os preditores X e Z são não correlacionados. Caso contrário, o uso dessa média como estimativa da distribuição marginal $\bar{f}(x)$ pode colocar peso em regiões inverossímeis ou de probabilidade zero.

4.7.2 Gráfico da esperança condicional individual

O gráfico da esperança condicional individual é equivalente ao gráfico de dependência parcial, com a diferença que aqui plotamos uma linha para cada observação da amostra, sendo que o gráfico de dependência parcial é a média das linhas do gráfico da esperança condicional individual.

O algoritmo para construção desse gráfico consiste nos passos (1) e (2a) da seção anterior, sendo que plotamos as n curvas formadas pelos pontos $(x_1, \hat{f}(x_1, z_j)), \dots, (x_M, \hat{f}(x_M, z_j))$, $j = 1, \dots, n$. Além de serem ainda mais intuitivo que o gráfico de dependência parcial, a grande vantagem do gráfico da esperança condicional individual é mostrar relações heterogêneas geradas por interações entre os preditores.

4.7.3 Gráfico de efeitos locais acumulados

O gráfico de efeitos locais acumulados é uma alternativa para o gráfico de dependência parcial no caso de preditores correlacionados. A estratégia aqui é, para cada valor x do preditor sob investigação X , estimar a distribuição marginal das previsões $\hat{f}(x)$ utilizando apenas as observações da amostra similares a x . Esse gráfico mostra como a previsão do modelo em uma pequena região ao redor de x para observações da base dentro dessa região.

Dado o preditor X de interesse, o algoritmo para construção desse gráfico é dado por:

1. dividir os valores observados de X em M intervalos (usualmente os quantis de X são utilizados);
2. para cada uma das m_i observações dentro do i -ésimo intervalo, calcular

$$\hat{f}_i^d = \hat{f}(x_i^+, z_j) - \hat{f}(x_i^-, z_j), \quad j = 1, \dots, m_i,$$

sendo x_i^+ e x_i^- , respectivamente, os limites superior e inferior do intervalo i .

3. Para cada valor x de X , calcular a média acumulada

$$\bar{f}_a(x) = \sum_{i=1}^{k(x)} \frac{1}{m_i} \sum_1^{m_i} \hat{f}_i^d,$$

sendo $k(x)$ o intervalo ao qual o valor x pertence.

4. Calcular $\bar{f}_{ac}(x)$, o valor centralizado de $\bar{f}_a(x)$ considerando todas as n observações da amostra:

$$\bar{f}_{ac}(x) = \bar{f}_a(x) - \frac{1}{n} \sum_{i=1}^n \bar{f}_a(x_i).$$

5. Para cada valor observado de X , plotar $\bar{f}_{ac}(x^c) \times x^c$, sendo x_1^c, \dots, x_{M-1}^c os pontos de corte utilizados para construir os M intervalos em (1).

A partir desse algoritmo, podemos fazer algumas considerações sobre o gráfico de efeitos locais acumulados:

- As diferenças calculadas no passo (2) explicam o termo *efeitos locais*. Para cada valor x do preditor X , essas diferenças estimam o quanto a predição é alterada por pequenas mudanças no valor de x .
- Calcular a média das diferenças das predições em cada intervalo (em vez de apenas a média das predições) também garante que o efeito estimado para o preditor X não seja influenciado pelo efeito de algum outro preditor correlacionado com X .
- Como cada $\frac{1}{m_i} \sum_1^{m_i} \hat{f}_i^d$ representa a diferença média nas predições quando X varia localmente, dentro do intervalo i , a soma acumulada no passo (3) é realizada para que $\bar{f}_a(x)$ represente o efeito do preditor X nas predições.
- Ao contrário do gráfico de dependência parcial e do gráfico de esperança condicional individual, que apresentam no eixo y o valor predito, o gráfico de efeitos locais acumulados consideram nesse eixo a diferença em relação ao valor predito médio.

4.7.4 LIME

Uma outra técnica utilizada para interpretar modelos caixa-preta é o LIME (*Local snterpretable Model-Agnostic Explanations*) (Ribeiro *et al.*, 2016). Ela vem sendo muito utilizada para detectar, explicar e corrigir problemas com modelos preditivos, pois permite avaliar para cada observação quais preditores mais influenciaram em sua própria predição.

O LIME assume que todo modelo complexo prevê valores parecidos para duas observações muito próximas, sendo possível encontrar um modelo simples que seja uma boa aproximação do modelo complexo na vizinhança da observação que queremos explicar. 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.
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.

Embora o LIME possa ser utilizado para explicar o modelo globalmente, explicando uma grande quantidade de pontos da base de treino, essa abordagem funciona bem para problemas *bem compostados*, como classificação de imagens, em que cada preditor representa um pixel, e textos, em que cada preditor representa uma palavra. Em problemas muito complexos, principalmente de regressão (variável resposta numérica) com muitos preditores, pode ser difícil encontrar modelos simples que expliquem bem as predições localmente.

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.

4.7.5 Exemplo

Para ilustrar a utilização desses métodos gráficos de interpretação, vamos considerar um problema simples no qual queremos associar a média diária de NO_x com as médias diárias de temperatura e umidade. Vamos então ajustar uma *random forest* para esses dados e tentar interpretar os resultados do modelo.

Na Figura 4.5 apresentamos o gráfico de dispersão do NO_x contra cada uma das variáveis explicativas. Embora os gráficos de dispersão não deixe claro essa relação, podemos observar pela série temporal do NO_x (Figura 4.6) que as maiores concentrações desse poluente tendem a acontecer no inverno, onde é clima⁹ é mais frio e seco.

⁹Dados para a cidade de São Paulo, de 2008 a 2011.

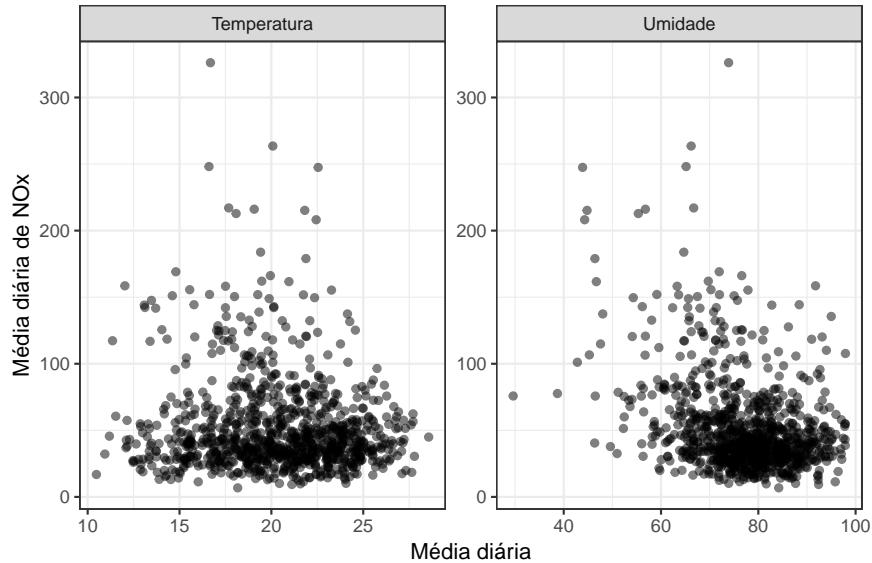


Figura 4.5: Gráficos de dispersão da média diária de NO_x contra as médias diárias de temperatura (Celsius) e umidade relativa do ar (%).

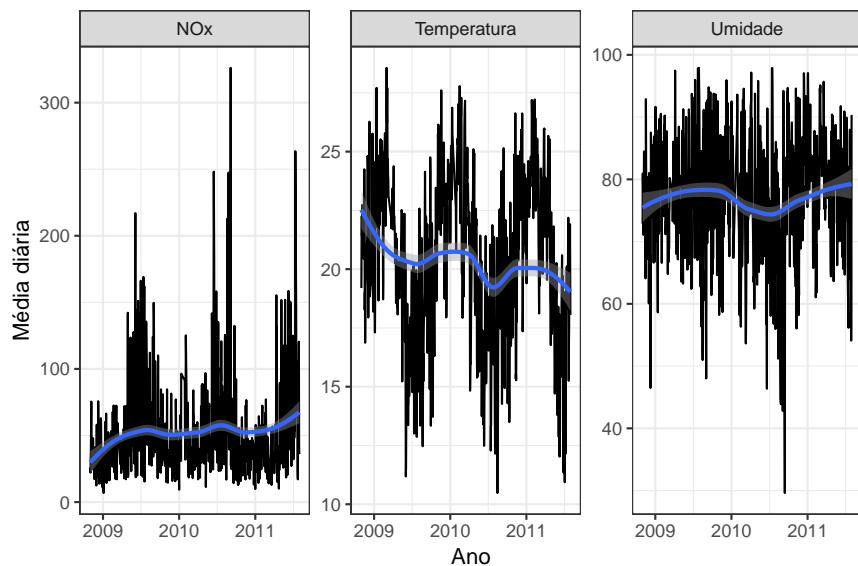


Figura 4.6: Gráficos de dispersão da média diária de NO_x contra as médias diárias de temperatura (Celsius) e umidade relativa do ar (%).

Inicialmente, ajustamos modelos de regressão linear para esses dados, pois eles são interpretáveis e podemos comparar seus resultados com os da *random forest*. Focando o exemplo apenas na interpretação da umidade, ajustamos modelos considerando (1) apenas a umidade como preditor, (2) umidade e temperatura como preditores e (3) umidade, temperatura e a interação entre as duas variáveis como preditores. Na Figura 4.7, apresentamos os gráficos do valor predito contra a umidade para cada um dos modelos. Repare que, para os modelos (2) e (3), avaliamos o efeito de umidade para diferentes temperaturas, que é constante no modelo sem interação e diminui conforme a temperatura aumenta no modelo com interação.

Ajustamos então uma *random forest* para os mesmos dados e construímos os gráficos de dependência parcial (PDP), esperança condicional individual (ICE) e efeitos locais acumulados (ALE) para interpretar os resultados (Figura 4.8). Pelo PDP e ICE, observamos um efeito estranho nos

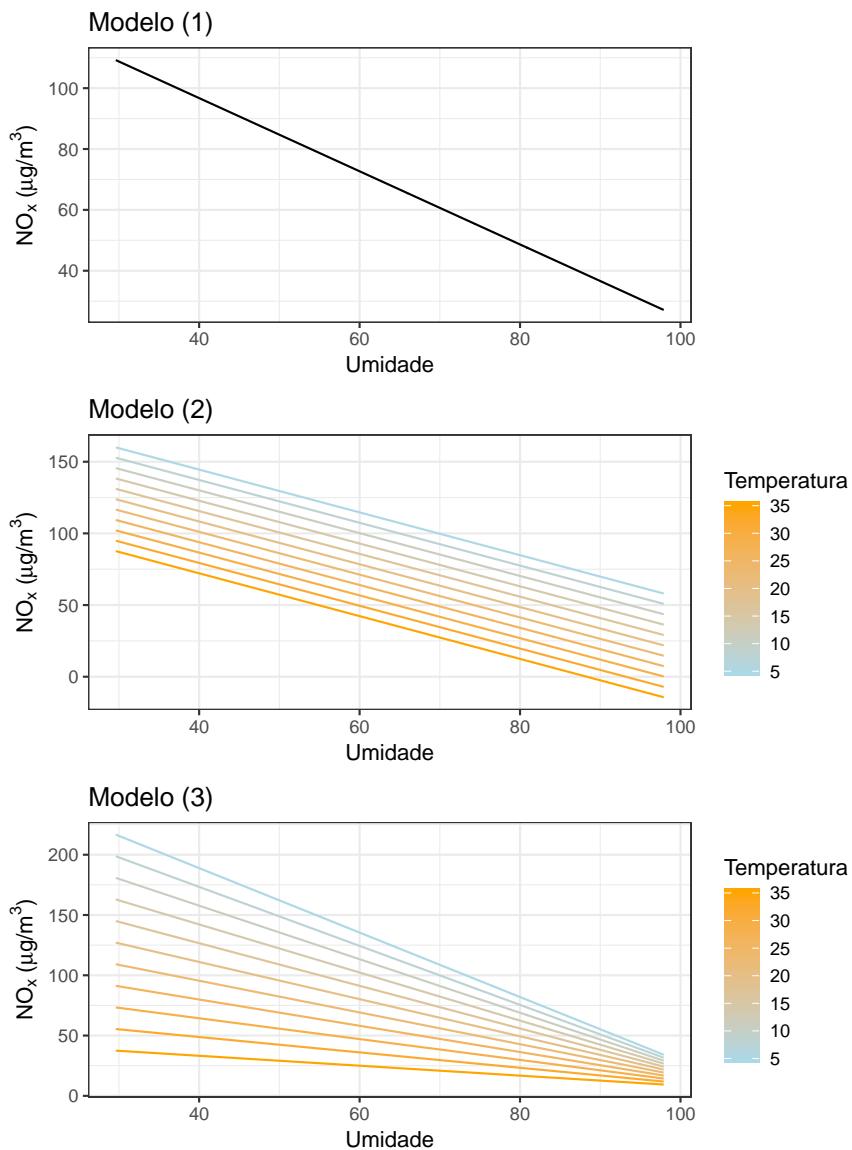


Figura 4.7: Valores preditos do NO_x para cada modelo em função da umidade e da temperatura.

preditores para valores baixos de umidade. Isso acontece porque a temperatura e a umidade são correlacionadas, como podemos observar pelo gráfico de dispersão na Figura 4.9. Como há poucos dias com baixa umidade, esses gráficos acabam apresentando o efeito da temperatura na concentração de NO_x . Nessa situação, o gráfico adequado para analisar o efeito da umidade seria o ALE, pois ele avalia localmente o efeito de cada valor da umidade na predição, desconsiderando regiões com poucas observações e eliminando a influência dos outros preditores.

O ALE também nos mostra que a relação entre umidade e concentração de NO_x pode ser não-linear, dado o aumento na predição quando aumentamos a umidade de 85% para 100%. Essa relação pode estar apenas sendo induzida pela ausência de outros preditores importantes para explicar a variabilidade do NO_x , mas, independentemente se ela tem uma explicação prática ou é uma deficiência do ajuste, é interessante notar que os modelos lineares não são flexíveis o suficiente para detectá-la.

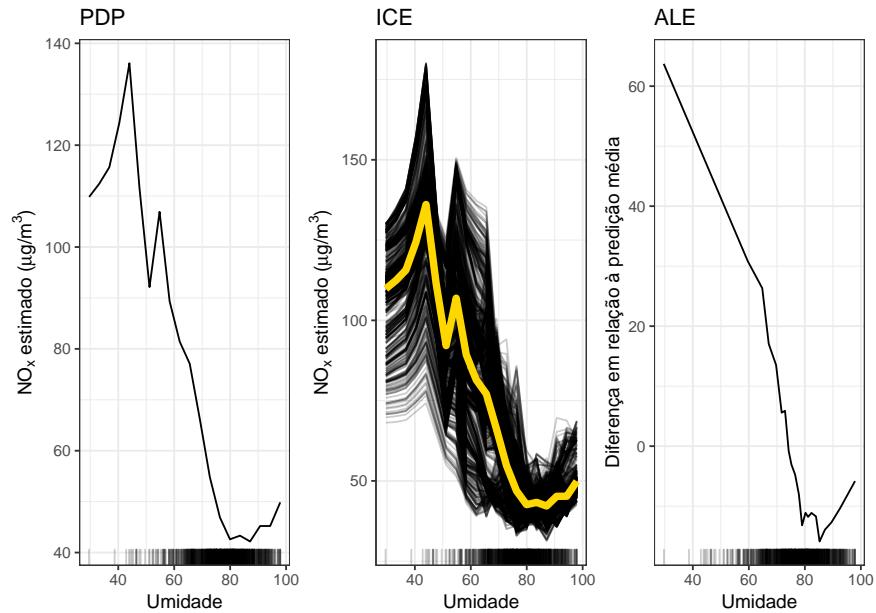


Figura 4.8: Gráficos de dependência parcial (PDP), esperança condicional individual (ICE) e efeitos locais acumulados (ALE) para a random forest. A curva amarela no ICE representa a média de todas as retas individuais, isto é, o PDP.

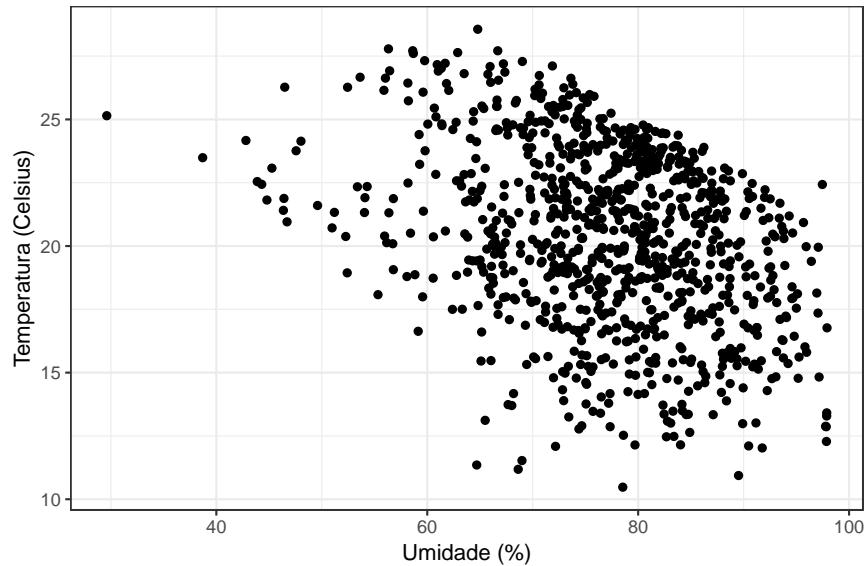


Figura 4.9: Gráfico de dispersão entre as médias diárias da temperatura e umidade.

Capítulo 5

Poluição e uso de combustíveis

Investigar a relação entre o uso de combustíveis e a concentração de poluentes é um problema muito comum em estudos de poluição do ar devido a grande contribuição da fonte veicular nos níveis de diversos gases e partículas. Neste capítulo, vamos analisar a associação entre uso de etanol e a concentração de ozônio na cidade de São Paulo como exemplo de utilização de diversas técnicas apresentadas nos últimos capítulos.

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 e do senso comum considerarem o etanol 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.

Utilizando uma mudança real na preferência por gasolina, ocasionada por flutuações de larga escala no preço do etanol, Salvo e Geiger (2014) analisaram a associação entre a proporção de carros bicombustíveis rodando a gasolina na cidade de São Paulo 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) sumarizaram 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 voláteis¹ e que esses períodos estavam associados a maiores concentrações do poluente.

Neste capítulo, utilizaremos os dados disponibilizados por Salvo *et al.* (2017) para aplicar diversas técnicas de modelagem, tendo como objetivo entender o mecanismo de formação do ozônio e, principalmente, sua relação com o uso de gasolina/etanol.

5.2 Entendendo o problema

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

O primeiro passo de qualquer análise estatística é entender bem o problema a ser resolvido. Embora pareça uma tarefa óbvia, é muito comum nos perdermos durante a modelagem por não termos relacionado com clareza o objetivo do estudo com as informações disponíveis para alcançá-lo. De maneira geral, essa etapa requer três delineamentos: quais perguntas serão respondidas, qual a dimensão do problema e quais métricas serão utilizadas para avaliar o erro associado às nossas conclusões. Definir esses pontos envolve conhecer a fundo tanto as técnicas estatísticas quanto sobre o fenômeno de interesse.

O nosso objetivo neste capítulo é investigar se existe associação entre a concentração de ozônio e a proporção de carros a gasolina rodando na cidade de São Paulo. Assim, qualquer modelo que

¹Um dos principais percursores do ozônio. Sua fonte primária é a queima parcial ou evaporação de etanol.

decidirmos usar deve levantar evidências contra ou a favor dessa pergunta. Em segundo plano, também podemos aproveitar para estudar o mecanismo de formação do ozônio e como os níveis desse poluente se comporta ao longo do ano e como ele se relaciona com variáveis climáticas.

Dimensionar a análise significa entender a complexidade do problema a ser resolvido. Não é raro termos mais ou menos informação do que precisamos para responder a pergunta de interesse, e então devemos avaliar se não estamos gastando muito esforço para responder perguntas que não são relevantes ou então investigando uma pergunta que não pode ser respondida com os dados disponíveis.

Conforme discutido no Capítulo 3, o processo por trás de um fenômeno pode ser muito complexo, sendo que a maioria das vezes não conseguimos nem mesmo identificar todas as variáveis envolvidas no seu mecanismo. 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.

Diante desse complicado panorama, devemos ter cuidado para:

1. não escolher um modelo muito complexo para responder uma pergunta que poderia ser respondida por uma técnica 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, poderíamos obter um modelo pouco interpretável ou, no caso de sobreajuste, que não fosse generalizável para além da amostra. Já no segundo, poderíamos obter um modelo que não respondesse a pergunta de interesse com precisão, podendo levar a conclusões inadequadas ou superficiais.

Dado o objetivo definido anteriormente, queremos descobrir qual é relação entre a variável resposta e um dos preditores, ou seja, 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 controlar a correlação das observações medidas no mesmo dia, o que pode ser feito, por exemplo, usando variáveis indicadoras para hora do dia ou a partir de 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 e menor radiação solar. 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, 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.

Por fim, precisamos definir como medir os erros associados às conclusões dos modelos. Para os modelos de regressão apresentados no Capítulo 3, podemos usar testes de hipóteses para avaliar se o coeficiente referente à proporção de carros a gasolina é ou não estatisticamente significativo para explicar a variação da concentração do ozônio. Para modelos não interpretáveis, nos restam os métodos de descritivos discutidos na Seção 4.7. Para medir a qualidade do ajuste, podemos utilizar o RMSE, que mede o quanto cada predição do modelo está errando em média o verdadeiro valor da concentração de ozônio, e o R^2 , que mede a porcentagem da variabilidade da concentração de ozônio explicada pelos preditores incluídos no modelo.

As estratégias apresentadas até aqui consideram apenas o conhecimento prévio sobre o fenômeno estudado e as variáveis disponíveis. Podemos continuar buscando informação sobre o fenômeno para melhorarmos nossa intuição sobre o modelo mais adequado a partir de uma análise exploratória.

5.3 Análise exploratória

O objetivo desta análise exploratória é investigar o comportamento dos preditores considerados por Salvo *et al.* (2017) 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 Parque Dom Pedro II. A análise exploratória completa pode ser visualizada nas postagens disponíveis em <https://rpollution.com/categories/ozônio>.

Na Seção 2.1.1, vimos como a concentração 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 à 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 a 50%. Repare que os dois picos acontecem no começo do ano, o que gera um desbalanço deste preditor quando observamos sua distribuição em cada mês (Figura 5.3). Se não controlarmos bem as outras variáveis envolvidas nesse mecanismo, como a concentração de ozônio é naturalmente maior no verão, esse desbalanço pode gerar uma correlação espúria, isto é, uma possível associação entre as

variáveis ser apenas coincidência.

Analizando agora 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.

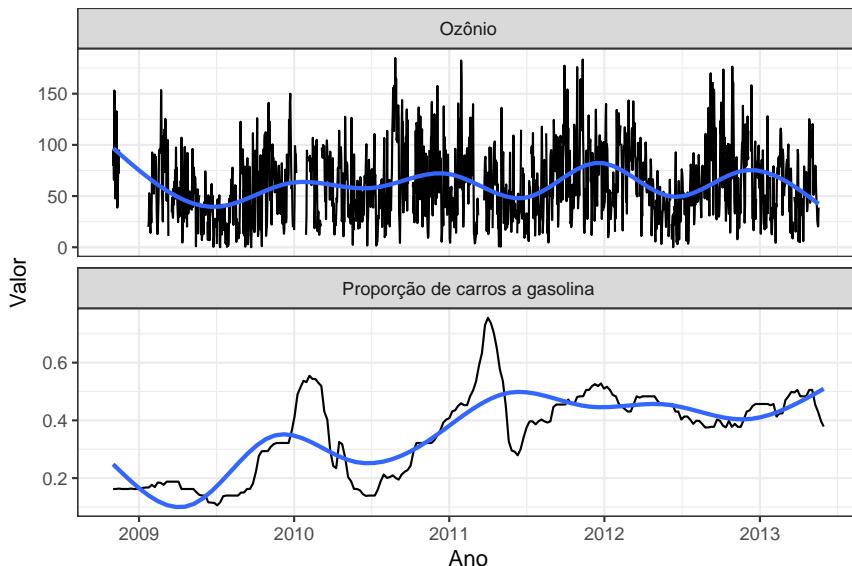


Figura 5.1: Séries da concentração de ozônio diária média e da proporção estimada de carros a gasolina rodando na cidade. Dados da estação Dom Pedro II, de 2008 a 2013.

Nas Figuras 5.4 e 5.5, 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 5.6, 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 climáticas², podemos concluir:

- dias com maior radiação estão associados a maiores níveis de ozônio;
- categorizando a variável precipitação em “Choveu no período” e “Não choveu no período”³, 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.

Analizando agora o trânsito diário médio na região da estação de monitoramento (Figura 5.7), não parece clara qual a relação com a concentração de ozônio. No entanto, se observamos agora a

²Os gráficos para os outros preditores podem ser encontrados em <https://www.rpollution.com/flexdashboards/ozonio-clima-sp/dash-ozonio-clima-sp.html>.

³A escolha pela categorização se deve ao alto número de zeros (dias sem chuva) que essa variável possui.

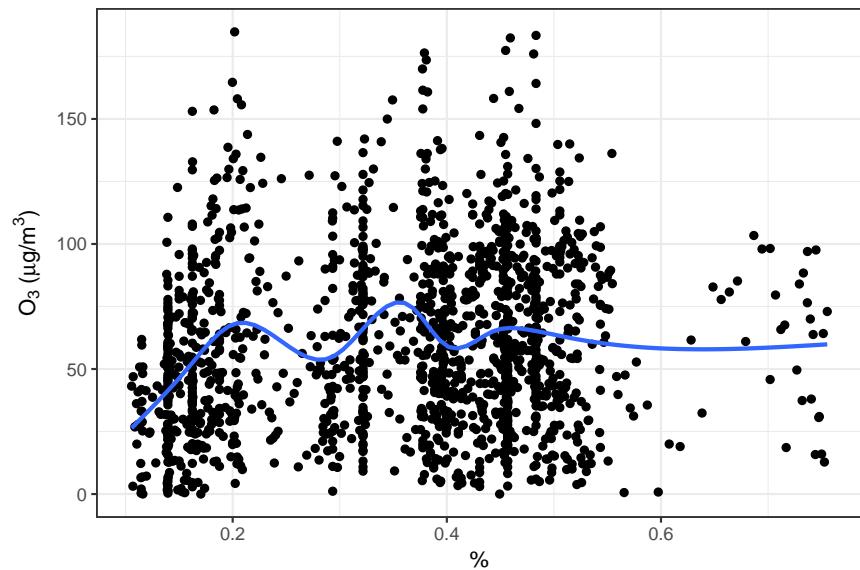


Figura 5.2: Gráfico de dispersão da concentração de ozônio contra a proporção estimada de carros rodando a gasolina na cidade. Dados da estação Dom Pedro II, de 2008 a 2013.

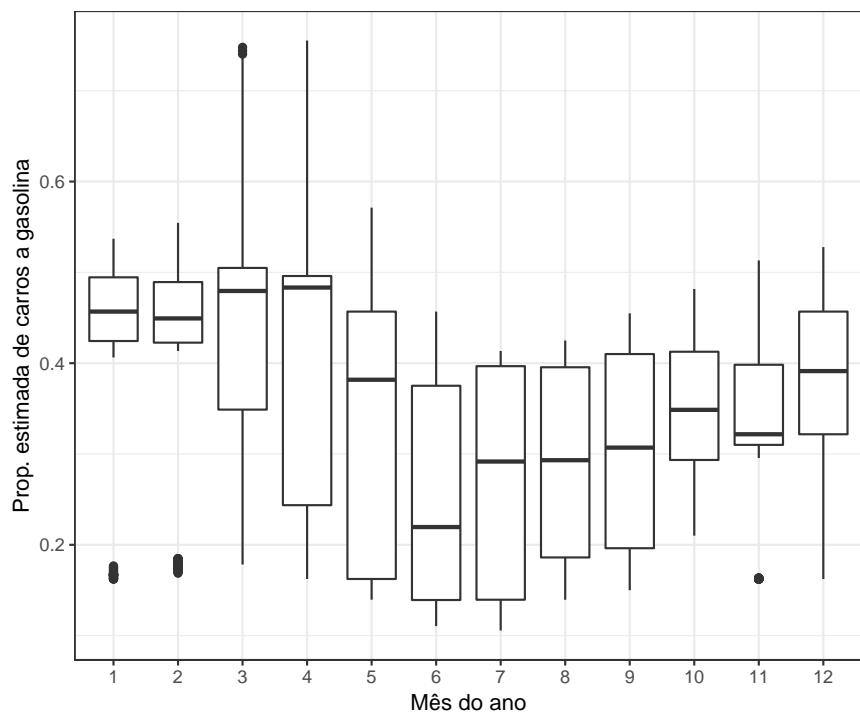


Figura 5.3: Boxplot da proporção estimada de carros a gasolina para cada mês.

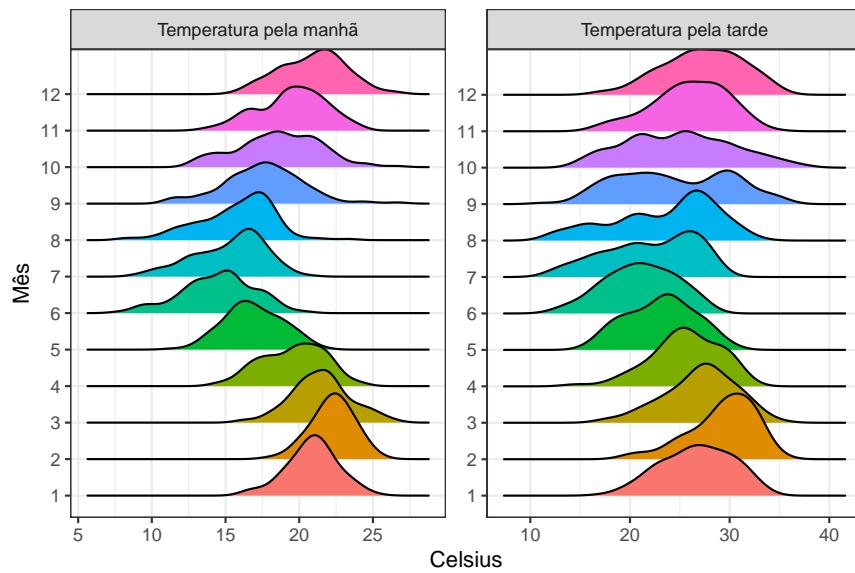


Figura 5.4: Gráficos ridge da temperatura diária média nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.

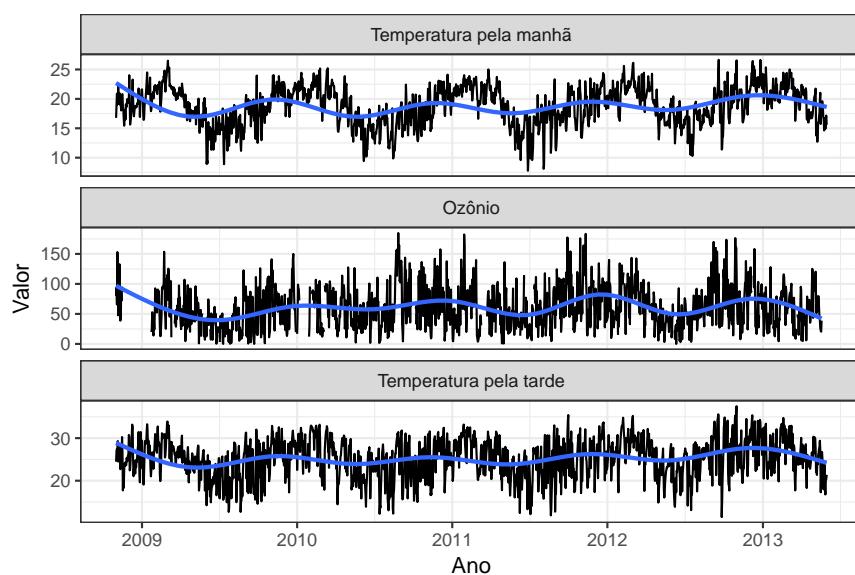


Figura 5.5: Gráficos das séries da concentração de ozônio e da temperatura diária média nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.

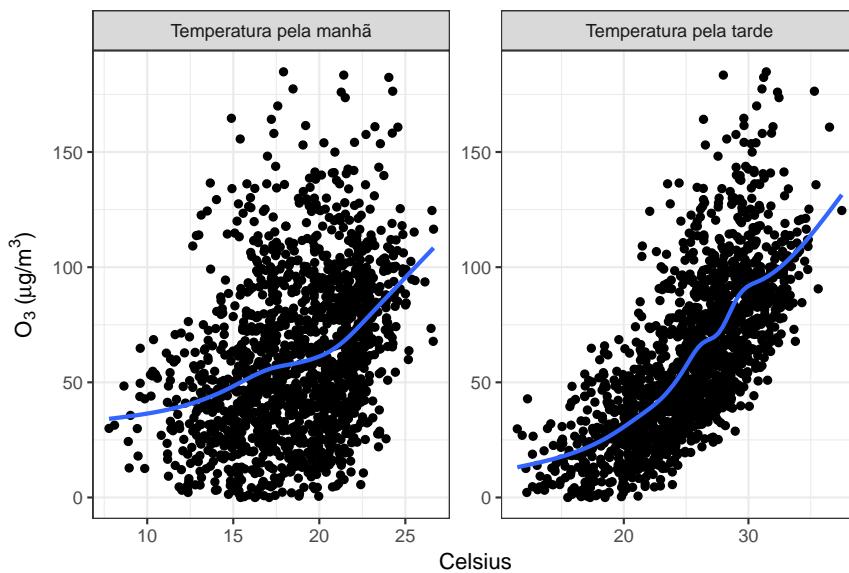


Figura 5.6: Gráficos de dispersão da concentração de ozônio pela temperatura diária média nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.

Figura 5.8, 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 e 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 5.9, a concentração de ozônio é maior em dias de menor tráfego independentemente de qual combustível está sendo mais utilizado na cidade.

Devido à complexidade do problema e ao grande número de variáveis, essa análise exploratória poderia 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. Nas próximas seções, vamos prosseguir com a análise discutindo 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 pontuado anteriormente, os autores optaram por remover da análise os meses de junho a setembro, optando por uma amostra mais homogênea em relação às condições climáticas. Além

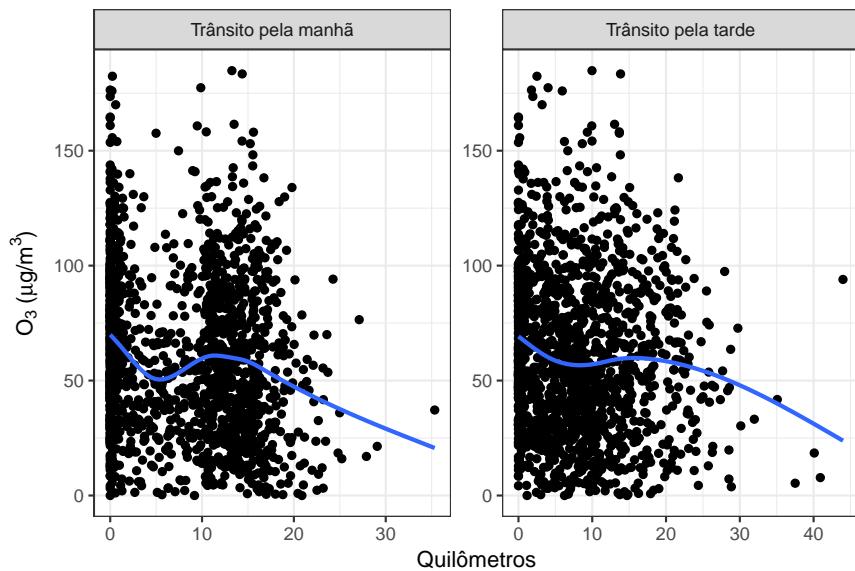


Figura 5.7: Gráficos de dispersão da concentração de ozônio pelo congestionamento diário médio, na região da estação de monitoramento, nos períodos da manhã (das 8 às 11 horas) e no período da tarde (12 às 17 horas). Dados da estação Dom Pedro II, de 2008 a 2013.

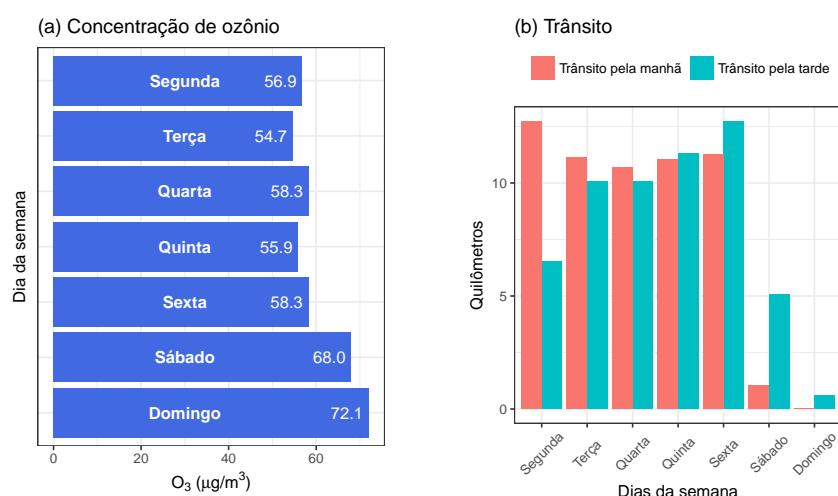


Figura 5.8: Relação entre a concentração de ozônio e o congestionamento na região da estação de monitoramento ao longo da semana. (a) Concentração de ozônio diária média ao longo da semana. (b) Congestionamento diário médio, no período da manhã e da tarde, na região da estação de monitoramento ao longo da semana. Dados da estação Dom Pedro II, de 2008 a 2013.

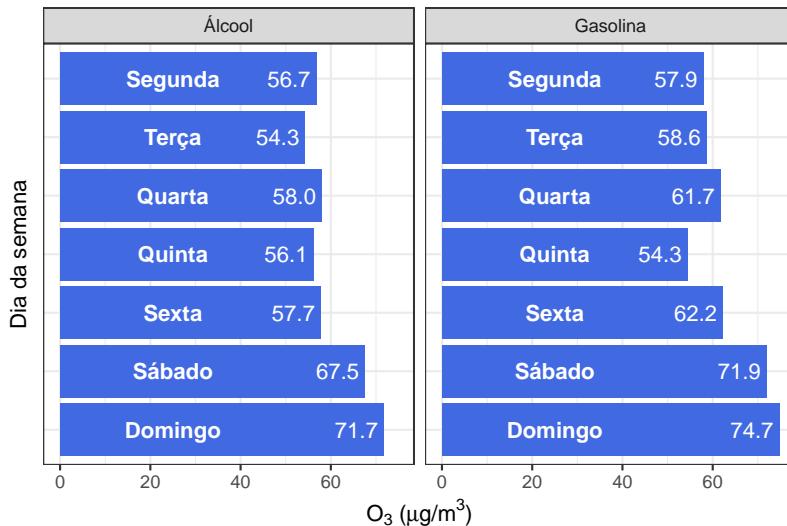


Figura 5.9: Concentração de ozônio diária média ao longo da semana em dias com maior proporção estimadas de carros rodando a álcool e em dias com maior proporção estimada de carros rodando a gasolina. Dados da estação Dom Pedro II, de 2008 a 2013.

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 diárias médias de uma estação. 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)⁴, o que indica que o aumento da proporção

⁴Esse 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.

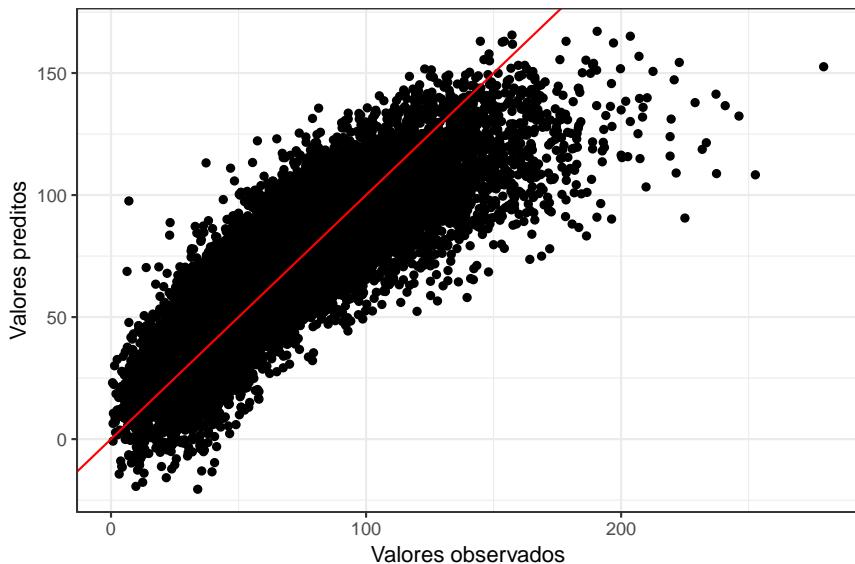


Figura 5.10: Valores da concentração de ozônio preditos pelo modelo de regressão linear ajustado por Salvo et al. (2017) contra os valores observados.

estimada de carros rodando a gasolina na cidade está associada com a diminuição da concentração 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 5.10, 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 próximas seções, vamos comparar o resultado obtido pelos autores com outras estratégias de análise, visando obter um ajuste mais preciso e avaliar se as conclusões encontradas pelos autores sem mantêm. Inicialmente, vamos utilizar as mesmas variáveis para explorar modelos mais flexíveis que o modelo de regressão linear. Em seguida, vamos testar novas estratégias de análise, alterando as variáveis e o dimensionamento do problema.

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 fortes sobre a relação entre as variáveis. 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, modelos de regressão segmentada, uma *random forest* e um *XGBoosting*. A variável resposta e os preditores serão os mesmos utilizados por [Salvo et al. \(2017\)](#). As performances serão comparadas a partir do RMSE calculado usando validação cruzada 5-fold. Também discutiremos como interpretar cada modelo e se as interpretações obtidas corroboram ou não as conclusões dos autores.

5.5.1 Modelos aditivos generalizados

Com o objetivo de ajustar um modelo mais flexível em relação a suposição de linearidade, consideraremos primeiramente modelos aditivos generalizados com as distribuições Normal, Gama e Normal Inversa. Embora a distribuição Normal gere resultados mais simples de serem interpretados, as distribuições Gama e Normal Inversa são, em teoria, mais adequadas à natureza da variável resposta, uma medida positiva e assimétrica. As funções não-lineares foram atribuídas a todos os preditores numéricos: 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 escolhido 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: Resultados dos modelos aditivos generalizados em comparação com o modelo utilizado por [Salvo et al. \(2017\)](#).

Modelo	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
Salvo et al. (2017)	19.74	70.65	Temperatura, velocidade do vento, radiação, umidade e var. ind. estação São Caetano do Sul

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 à 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 e nos modelos Gama e 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 estimada⁵. Na Figura 5.11, apresentamos esse gráfico para o preditor referente à proporção estimada de carros a gasolina utilizando o modelo com distribuição Normal. A partir dele, observamos que

- a função estimada sugere uma relação não linear entre as variáveis;
- para valores extremos, menores que 20% e maiores que 60%, um aumento da proporção estimada de carros a gasolina está associada com um aumento na concentração de ozônio;
- para valores intermediários, entre 20% e 60%, um aumento da proporção estimada de carros a gasolina está associada, de uma forma geral, com uma diminuição da concentração do ozônio;
- mesmo na faixa intermediária, onde a gasolina parece assumir um papel protetor, a curva não é monótona, isto é, ela decresce com diferentes velocidades e chega a ser crescente próximo ao 50%;
- por esse modelo, o valor da proporção de carros a gasolina que geraria o menor impacto no nível de ozônio seria aproximadamente 58%;
- como há relativamente poucos dias com proporção estimada de carros a gasolina muito baixa ou muito alta⁶, devemos tomar cuidado ao tirarmos conclusões sobre a relação entre as variáveis nesses intervalos;
- os indícios trazidos por esse gráfico não contradizem os resultados encontrados por Salvo *et al.* (2017), mas sim os complementam, levantando outras hipóteses para um problema inerentemente complexo.

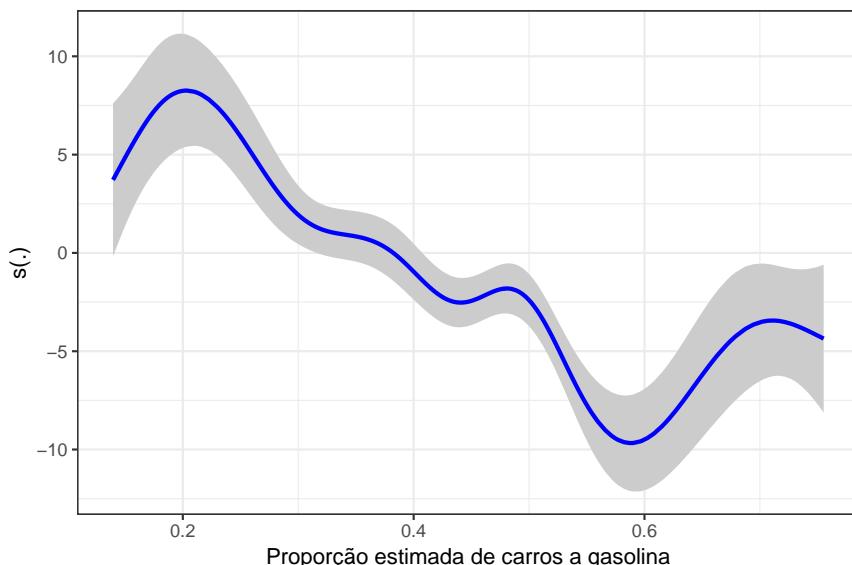


Figura 5.11: Função não-linear estimada pelo modelo aditivo generalizado com distribuição Normal para a proporção estimada de carros rodando a gasolina. A área cinza em volta da curva representa o intervalo de confiança com 2 erros-padrão para cima e para baixo.

⁵Os preditores que entraram no modelo de forma linear podem ser interpretados de maneira análoga a um modelo de regressão linear.

⁶Como pode ser observado pela amplitude do intervalo de confiança dado pela área cinza em torno da curva.

A Figura 5.12 apresenta o gráfico dos valores preditos contra os valores observados para o modelo com distribuição Normal (a) e para o modelo com distribuição Gama (b). O modelo com distribuição Normal possui o mesmo problema 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.

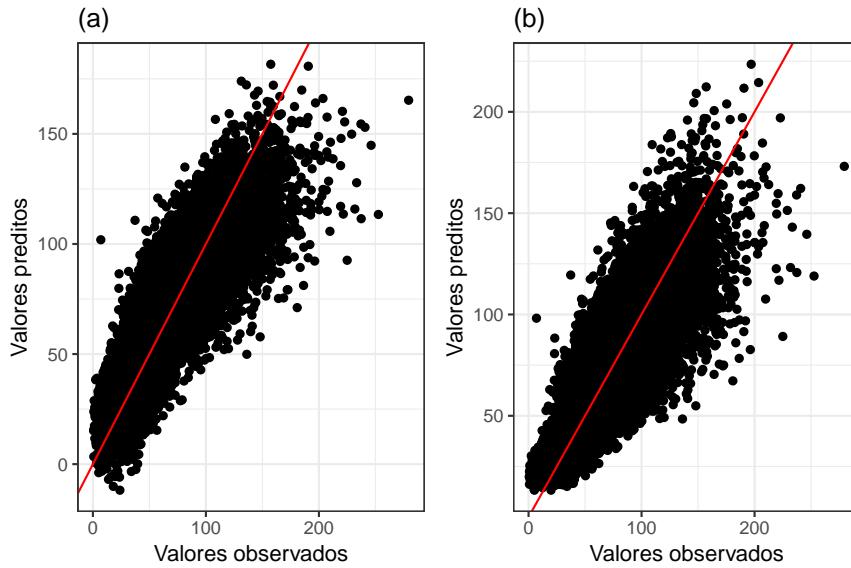


Figura 5.12: Valores da concentração de ozônio preditos pelo modelo com distribuição Normal (a) e pelo modelo com distribuição Gama (b) contra os valores observados.

Para avaliar a variabilidade das estimativas, criamos 200 amostras de *bootstrapping* e ajustamos o modelo aditivo generalizado com distribuição Normal (que apresentou o melhor desempenho) para cada uma delas. Na Figura 5.13, 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 5.11. A curva azul é a curva suavizada por *splines* cúbicos. Podemos notar que o modelo parece robusto quanto as interpretações obtidas para a Figura 5.11. Esse gráfico também ressalta a maior variabilidade nos extremos do preditor.

Embora o modelo aditivo tenha levantado outras hipóteses sobre a relação entre a concentração de ozônio e a proporção estimada de carros a gasolina, ele não trouxe ganho de precisão se compararmos com o modelo de regressão linear. Vamos então continuar explorando a não linearidade começando com um modelo de regressão segmentada, mais simples do que o modelo aditivo, e, em seguida, ajustando uma *random forest*, um modelo não-linear mais flexível para tentar aumentar a precisão do ajuste.

5.5.2 Modelo de regressão segmentada

Para explorar um modelo não linear mais simples do que o modelo aditivo generalizado, ajustamos o modelo de regressão segmentada discutido na Seção 3.1.5. Com base na Figura 5.11, optamos por segmentar a reta de regressão primeiro em dois pontos, com valores iniciais por volta de 0.2 e 0.6, e em seguida em três pontos, com valores iniciais por volta de 0.2, 0.5 e 0.6. Esses pontos

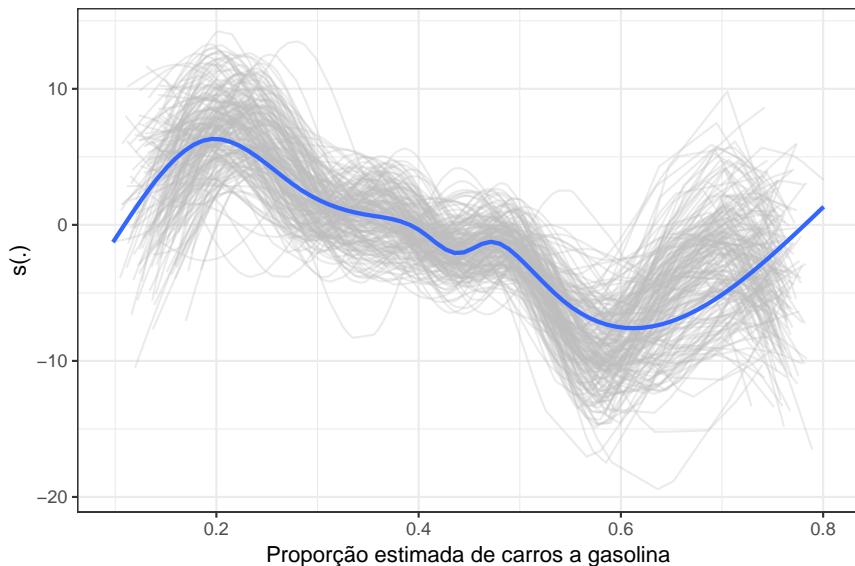


Figura 5.13: Em cinza, as funções estimadas da variável referente à proporção estimada de carros a gasolina para cada uma das 200 amostras de bootstrapping. Em azul, a curva suavizada por splines cúbicos.

representam as maiores inflexões da curva estimada no modelo aditivo Normal para a proporção estimada de carros a gasolina. Como o modelo é altamente sensível ao ponto de segmentação inicial, testamos diversas combinações de pontos com valores ao redor desses números. A combinação que levou ao menor RMSE foi $(0.21, 0.6)$ para o modelo com dois pontos e $(0.21, 0.51, 0.59)$ para o modelo com três. Os resultados para estes modelos estão apresentados na Tabela 5.3.

Tabela 5.3: Resultado do modelo de regressão segmentada.

Pontos de segmentação estimados	RMSE	% var. explicada	Variáveis mais importantes
$(0.51, 0.58)$	19.6	70.50	Temperatura, vento, radiação, umidade e var. ind. para a estação São Caetano do Sul
$(0.26, 0.51, 0.58)$	19.45	71.23	Temperatura, vento, radiação, umidade e var. ind. para a estação São Caetano do Sul

Observamos que os modelos apresentam performance ligeiramente melhor do que a dos modelos lineares e aditivos (menor RMSE e maior R^2). As variáveis mais importantes para explicar a concentração de ozônio continuam as mesmas do modelo de regressão linear. A proporção de carros a gasolina foi considerada estatisticamente significante (valor-p < 0.001) em ambos os modelos, sendo que a Figura 5.14 apresenta as retas de regressão segmentada estimada para o efeito desse preditor na variação da concentração de ozônio. Quando comparamos esses gráficos com os gráficos do modelo aditivo generalizado (Figuras 5.11 e 5.13), notamos que ambos os modelos capturam relações não-lineares semelhantes, principalmente a queda acentuada do efeito por volta do valor 0.5 e o aumento próximo ao valor 0.6. A diferença entre o modelo com 2 pontos e o com 3 pontos é que o primeiro não captura o vale existente em torno do 0.28 que o segundo modelo sugere.

Embora as funções não-lineares dos modelos de regressão segmentada e do modelo não-linear apresentem algumas diferentes, ambos sugerem que um modelo linear não é o mais adequado para

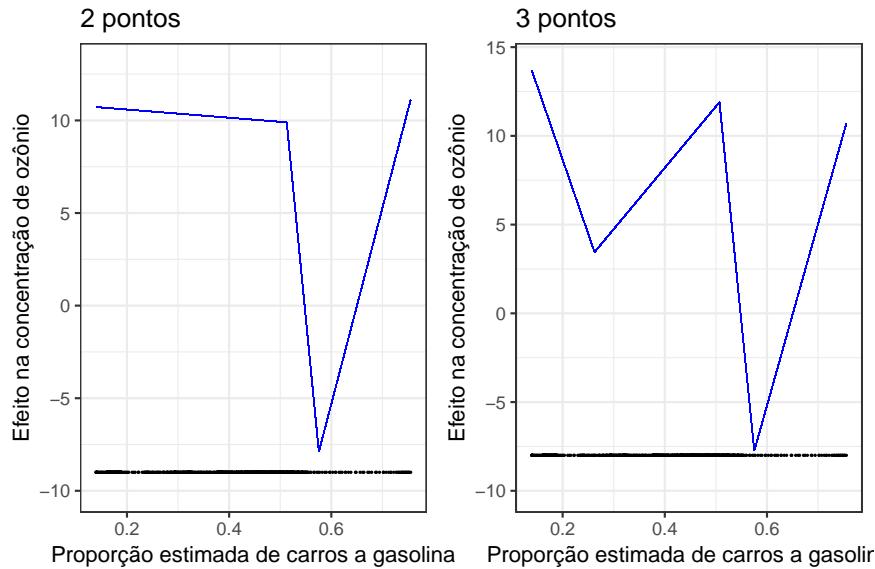


Figura 5.14: Retas de regressão segmentada para a proporção estimada de carros a gasolina. O efeito representa o valor da concentração de ozônio para cada valor da proporção estimada de carros a gasolina se todos os outros preditores tivessem valor igual a 0. Essa medida não tem interpretação prática, mas pode ser utilizada para calcular a variação no ozônio quando variamos o a proporção de carros a gasolina.

investigar a associação entre a concentração de ozônio e a proporção de carros rodando a gasolina na cidade. Ambos os modelos também estimam que a proporção ótima de carros rodando a gasolina para a redução dos níveis de ozônio estaria em torno de 58%. A seguir, ajustaremos um modelo mais flexível, na tentativa de melhorarmos a precisão das estimativas.

5.5.3 Random Forest

Em busca de resultados mais precisos, ajustamos também uma *random forest* aos dados. Os resultados estão resumidos na Tabela 5.4.

Tabela 5.4: 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 ($\text{RMSE} = 14.11$) do que o modelo de regressão linear (19.74), além de explicar uma maior porcentagem da variação da concentração de ozônio (85.72% contra 70.65% do modelo de regressão linear). 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 5.15 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 valores altos da concentração de ozônio

tendem a ser subestimadas pelo modelo.

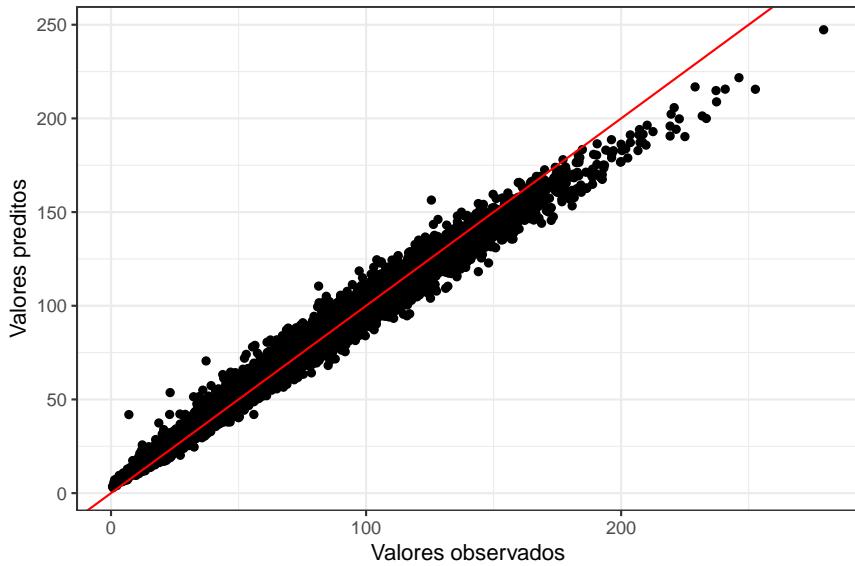


Figura 5.15: Valores da concentração de ozônio preditos pelo modelo random forest contra os valores observados.

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 associado à resposta. Sem essa interpretação, não conseguimos responder à pergunta de interesse do estudo.

Para transpor esse problema, vamos utilizar as técnicas gráficas apresentadas na Seção 4.7 para investigar o efeito da proporção estimada de carros a gasolina na concentração de ozônio. Na Figura 5.16, apresentamos os gráficos de dependência parcial (PDP) e de efeitos locais acumulados para esse preditor. Observamos que o efeito é muito parecido com o encontrado no modelo aditivo generalizado. A semelhança entre as curvas dos dois gráficos sugere que a proporção de carros a gasolina não é correlacionada com outros preditores.

Os resultados encontrados aqui mostraram que a *random forest* ratificou as conclusões encontradas pelo modelo aditivo generalizado. Na próxima seção, vamos dar mais um passo na direção da precisão e ajustar um *XGboost* aos dados.

5.6 XGBoost

Graças a sua acurácia e eficiência computacional, o *XGBoost* é um dos modelos mais utilizados hoje em dia em tarefas de predição. Com o objetivo de validar os resultados encontrados até então, vamos ajustar esse modelo aos dados e avaliar se, primeiro, realmente obtemos um melhor ajuste e, segundo, se as conclusões apontam na mesma direção que os modelos anteriores.

A performance do *XGBoost* ajustado estão na Tabela 5.5. Podemos observar que, de fato, o ajuste foi mais preciso do que para os modelos de regressão linear, aditivo generalizado e a *random forest*.

Agora, assim como fizemos para a *random forest*, vamos utilizar os gráficos de dependência parcial (PDP) e de efeitos locais acumulados (ALE) para avaliar o efeito da proporção estimada de

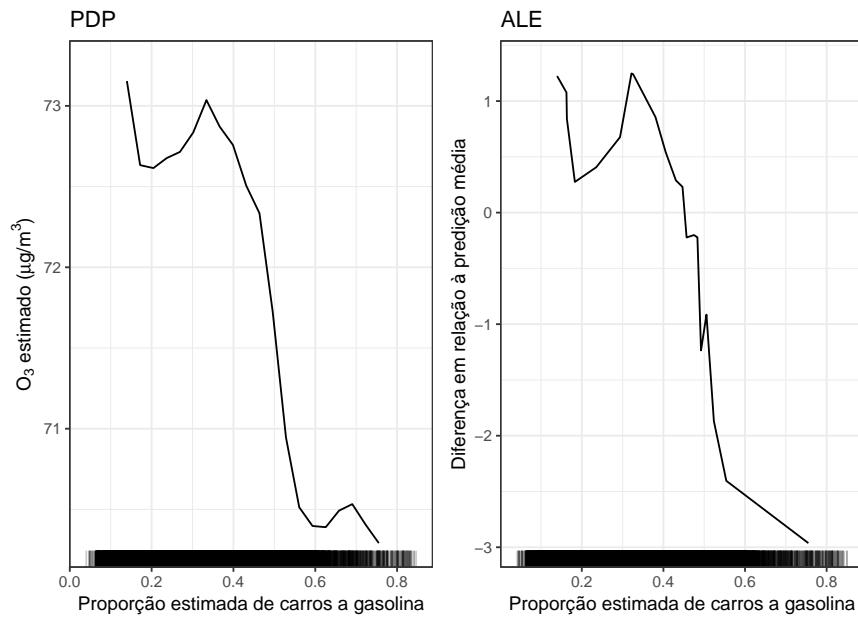


Figura 5.16: Gráficos de dependência parcial (PDP) e de efeitos locais acumulados para o modelo random forest.

Tabela 5.5: Performance do modelo XGBoost aplicado aos dados de Salvo et al. (2017) em comparação com os outros modelos ajustados.

Modelo	RMSE	% var. explicada	Variáveis mais importantes
XGBoost	12.24	88.56	Temperatura, umidade, radiação, tendência e vento
Random forest	14.11	85.72	Temperatura, umidade, radiação, vento e tendência
GAM (Normal)	19.82	70.50	Temperatura, vento, umidade, radiação e tendência
Salvo et al. (2017)	19.74	70.65	Temperatura, velocidade do vento, radiação, umidade e var. ind. estação São Caetano do Sul

carros a gasolina na concentração de ozônio predita pelo modelo. Podemos observar (Figura 5.17) que a proporção de carros a gasolina não parece exercer um efeito claro na concentração de ozônio estimada, variando a depender do valor desse preditor.

Como comparativo e também para avaliarmos a qualidade das interpretações feitas pelos gráficos de dependência parcial e de efeitos locais acumulados, apresentamos na Figura 5.18 o ALE para a temperatura, umidade, radiação e velocidade do vento. Podemos observar que a interpretação encontrada é razoável com o conhecimento sobre a geração do ozônio: temperatura e radiação estão positivamente associadas e velocidade do vento e umidade negativamente associadas.

Os resultados desta seção levantaram algumas dúvidas sobre o verdadeiro efeito da proporção estimada de carros a gasolina na concentração de ozônio e também sobre a capacidade dessa variável realmente estar representando o real uso de gasolina/etanol na cidade. Nas próximas seções, utilizaremos novas estratégias de análise para continuar testando a robustez das conclusões obtidas nesse estudo.

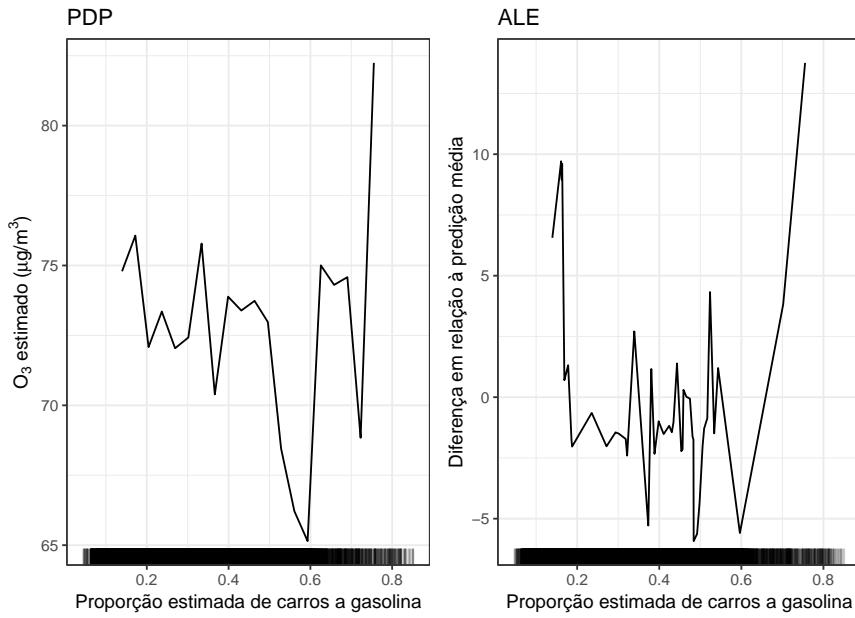


Figura 5.17: Gráficos de dependência parcial (PDP) e de efeitos locais acumulados para a proporção estimada de carros a gasolina do modelo XGBoost.

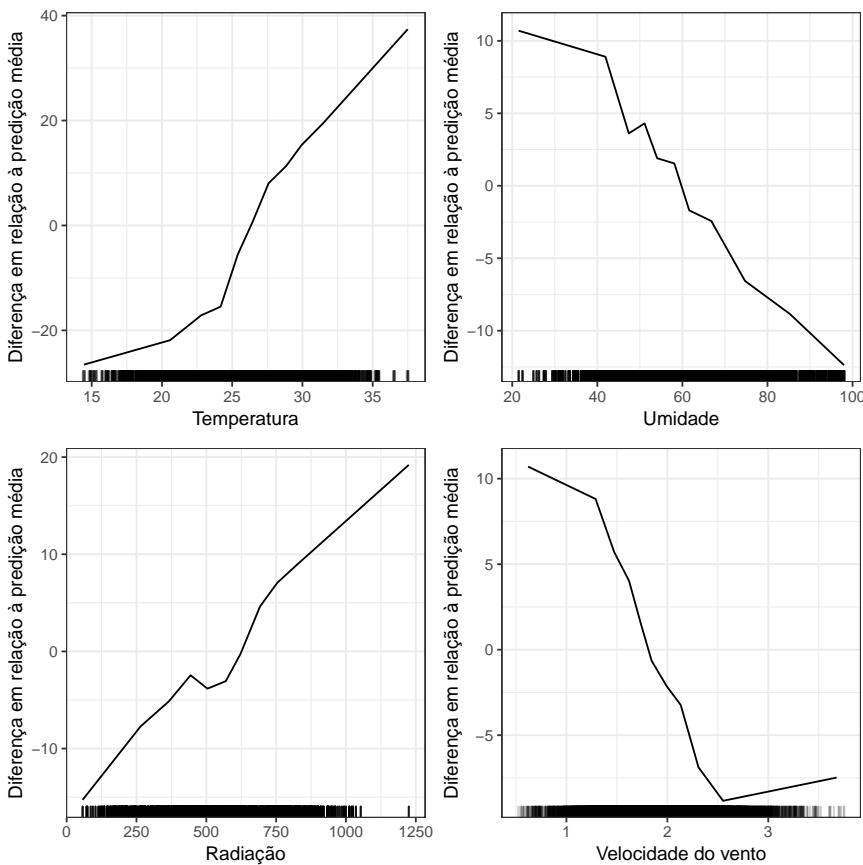


Figura 5.18: Gráficos efeitos locais acumulados para as variáveis climáticas do modelo XGBoost.

5.7 Outras estratégias de análise

Como forma de avaliar a robustez dos indícios encontrados até então, vamos apresentar nesta seção algumas alternativas para as estratégias de análise adotadas anteriormente.

5.7.1 Seleção de variáveis

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 4.4, 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 estimativa 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.

5.7.2 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 desses níveis elevados de ozônio que não foi inserida na análise. No entanto, existem situações em que esse problema de ajuste se deve às 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 5.19).

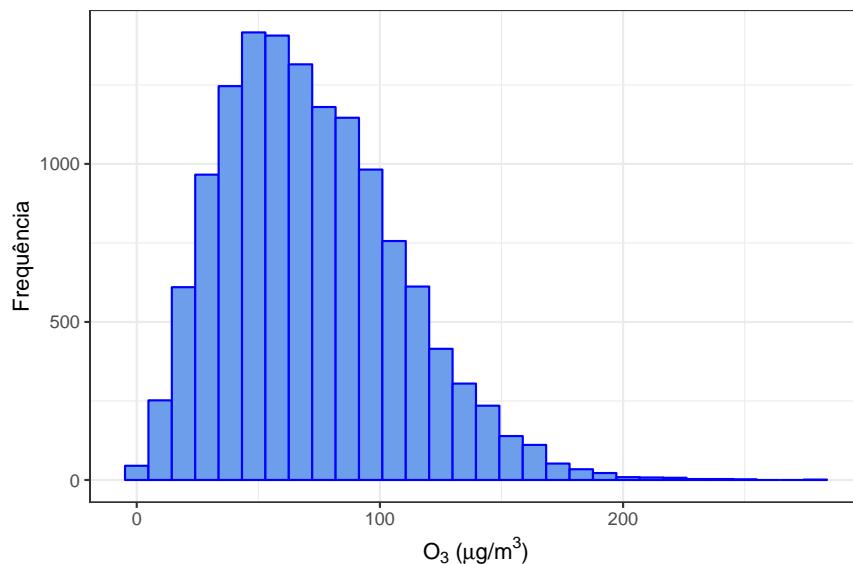


Figura 5.19: Distribuição da concentração de ozônio na amostra considerada por Salvo *et al.* (2017).

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 5.6, apresentamos os resultados do modelo de regressão linear ajustado com as transformações *log* e Box-Cox (com lambda = 0.51). Com a ajuda da Figura 5.20, 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 aumentando a porcentagem da variância explicada. Como a *random forest* não faz restrições sobre a distribuição da variável resposta, o ganho encontrado ao transformarmos a variável foi muito pequeno.

Tabela 5.6: Resultado dos modelos ajustados com a variável resposta transformada.

Modelo	Transformação	RMSE	% var. explicada	Variáveis mais importantes
Regressão linear	Sem transformação	19.74	70.65	Temperatura, vento, radiação, umidade e var. indicadora da estação São Caetano
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	Sem transformação	14.11	85.72	Temperatura, umidade, radiação, vento e tendência
Random Forest	Box-Cox	14.04	86.87	Temperatura, umidade, radiação, vento e tendência

Embora esses modelos não tragam novas informações sobre o fenômeno sob estudo, seus resultados são importantes como validação da análise, mostrando que os indícios encontrados até então são robustos em relação a distribuição da variável resposta.

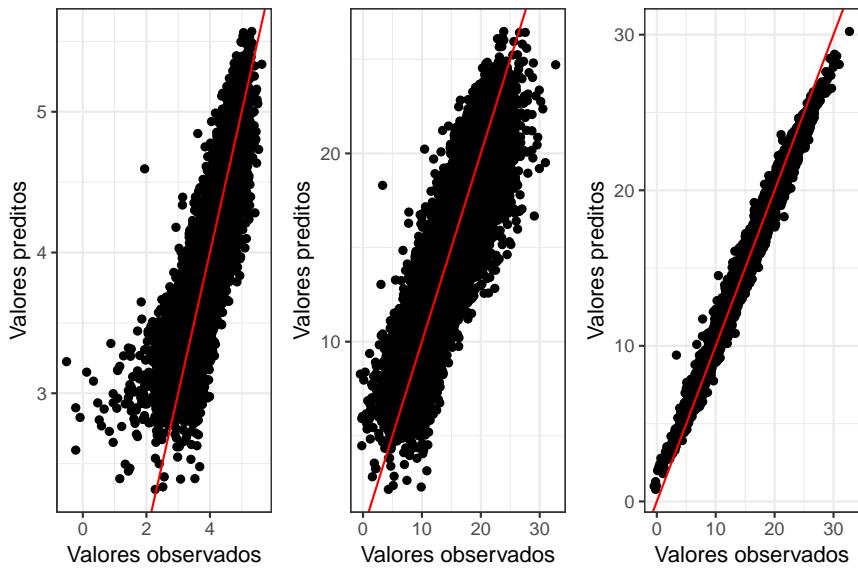


Figura 5.20: Gráficos dos valores da concentração de ozônio preditos pelos modelos contra os valores observados. No painel da esquerda, modelo de regressão linear com transformação log. No painel do meio, modelo de regressão linear com transformação Box-Cox. No painel da direita, random forest com transformação Box-Cox.

5.7.3 Ajustando a máxima diária

Para verificar se os resultados são robustos quanto a maneira escolhida para agrregar as observações de um mesmo dia, vamos ajustar o modelo de regressão linear e a *random forest* agora para a máxima diária da concentração de ozônio.

Na Tabela 5.7, 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.

Tabela 5.7: 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

Também observamos pela Figura 5.21 que os modelos para essa nova variável também subestima valores altos da máxima diária.

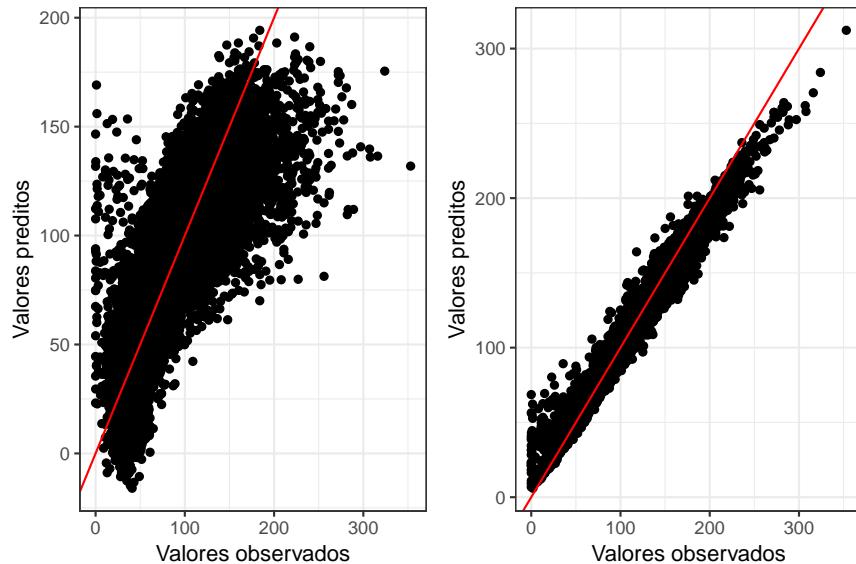


Figura 5.21: Gráficos dos valores da máxima diária da concentração de ozônio preditos pelos modelos contra os valores observados. No painel da esquerda, modelo de regressão linear e, no painel do meio, a random forest.

Apesar da redução na precisão dos modelos, os resultados obtidos apontam na mesma direção: a proporção estimada de carros rodando a gasolina é estatisticamente significante para explicar variações na concentração de ozônio. Além disso, de uma maneira geral, o aumento dessa proporção tende a diminuir as concentrações do poluente.

5.7.4 Ajustando cada estação separadamente

A estratégia adotada até aqui para tratar as diferentes estações de monitoramento foi juntar as informações de todas elas em uma única amostra, analisando os dados conjuntamente e incluindo variáveis indicadoras para controlar o efeito de cada estação.

O mecanismo gerador do ozônio é altamente dependente da química atmosférica e das condições climáticas do local. Dada a grande área em que as estações estão espalhadas⁷, podemos refazer a análise ajustando um modelo para cada uma das estações para avaliar se os resultados encontrados são robustos em relação à posição geográfica de cada estação.

Na Tabela 5.8, 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 de monitoramento. 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 Sul⁸, 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 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?

Mesmo nem sempre sendo possível responder todas as perguntas levantadas com os dados disponíveis, levantar dúvidas é importante tanto para ponderar as conclusões do estudo em questão quanto para motivar a realização de novos estudos. Discutiremos melhor os pontos em torno dessas perguntas nos comentários da próxima seção.

5.8 Comentários

A análises realizadas neste capítulo indicaram que a associação entre a concentração de ozônio e a proporção estimada de carros a gasolina não é linear. Para valores muito baixos ou muito altos da proporção estimada de carros a gasolina, a relação com a concentração de ozônio não é muito clara. Embora os modelos aditivos e de regressão segmentada indiquem uma associação positiva nesses intervalos, o relativamente pequeno número de dias com proporções muito baixas ou muito altas

⁷São Paulo ocupa uma área de mais de 1500 quilômetros quadrados, estando entre as 10 maiores cidades do mundo.

⁸Para a estação IPEN, o coeficiente é marginalmente significativo.

Tabela 5.8: 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

nos impede de tirar conclusões mais concretas. A não-linearidade sugerida pelos modelos reflete bem a complexidade do fenômeno sob estudo.

Os modelos também apontaram que a direção dessa associação é, em geral, negativa, mas a forma variou consideravelmente a depender do modelo escolhido. Também mostramos que os resultados

dependem da estação de monitoramento, indicando a importância da química atmosférica local na formação do ozônio ou a impossibilidade da estimativa da proporção de carros a gasolina representar toda a cidade.

Se a associação negativa entre essas variáveis for verdadeira, ela pode ser explicada pelo mecanismo gerador do ozônio troposférico. Dentre outros fatores, esse processo depende da presença de NO₂, 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 NO₂ tendem a gerar mais ozônio quando mais NO₂ 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, em geral, negativa entre congestionamento e concentração do poluente. Essa relação, no entanto, não parece ser linear, ou seja, ter uma frota inteiramente formada por carros a gasolina não levaria aos menores índices de ozônio. Encontramos indícios de uma proporção ótima de carros a gasolina por volta de 60%.

Os resultados contrários encontrados para as estações IPEN e São Caetano do Sul podem indicar que esses locais são limitadas em NO₂, sendo preciso estudar melhor a química atmosférica nessas regiões.

Na maioria dos modelos, observamos um maior erro para valores altos da concentração de ozônio. Isso indica que possivelmente estamos deixando fora da análise alguma variável que explique os picos desse poluente.

Neste tipo de estudo, também é importante considerarmos 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 $\mu\text{g}/\text{m}^3$. Considerando a concentração média de ozônio em toda a amostra, uma redução média de 72.2 $\mu\text{g}/\text{m}^3$ para 63.9 $\mu\text{g}/\text{m}^3$ pode não ter relevância para a criação de políticas públicas para a redução de emissões de etanol.

Assim, 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.

Capítulo 6

Poluição e saúde pública

Embora o estudo da poluição do ar só tenha passado a ser tratada como ciência no século 20, relatos milenares de problemas ambientais e de saúde pública envolvendo queima de compostos e derretimento de metais foram encontrados em cidades da antiguidade, como Grécia e Roma (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 (1992b), por exemplo, concluíram que a concentração de partículas suspensas no ar estava positivamente associada com a mortalidade no dia seguinte em Steubenville, Ohio. Saldíva *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 $5 \mu\text{g}/\text{m}^3$ 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.

Neste capítulo, utilizaremos dois problemas como exemplo. Primeiro, vamos expandir a análise do capítulo anterior e investigar 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, vamos comparar esses resultados com os de uma análise utilizando a própria concentração de ozônio como preditora da mortalidade.

6.1 Uso de etanol e mortalidade

No último capítulo, utilizamos os dados disponibilizados por *Salvo et al.* (2017) para estudar a relação entre a proporção estimada de carros bicombustíveis rodando a gasolina na cidade de São Paulo e a concentração troposférica de ozônio. As análises conduzidas tanto pelos autores quanto neste trabalho levantaram indícios de um efeito protetor da gasolina, isto é, o aumento do uso de gasolina estaria associado com uma diminuição da concentração de ozônio. Enquanto *Salvo et al.* (2017) estimaram uma relação linear entre as variáveis, os modelos do capítulo anterior sugeriram que essa associação é não-linear, sendo que sua direção e intensidade podem depender do valor da proporção de carros rodando a gasolina na cidade.

Se esse indício for verdadeiro, dado o impacto do ozônio na saúde pública, é razoável pensarmos em políticas públicas para controlar o uso do etanol, visando diminuir os níveis do ozônio. Contudo, *Salvo et al.* (2017) mostraram também que o aumento da proporção estimada de carros a gasolina está associado com o aumento de partículas finas, enquanto *Salvo e Geiger* (2014) mostraram que o aumento dessa variável está associada com maiores concentrações de monóxido de carbono e de nitrogênio. Assim, é interessante avaliarmos a associação da proporção estimada de carros a gasolina diretamente na mortalidade, investigando o impacto do uso de etanol na saúde pública independentemente dos poluentes aos quais ele está associado. Além disso, podemos utilizar a concentração de ozônio agora como preditor para explicar a variação da mortalidade e comparar se o uso de etanol e a concentração de ozônio impactam a mortalidade da mesma forma.

Para cumprir esses objetivos, vamos utilizar os dados climáticos e de poluição disponibilizados por *Salvo et al.* (2017) e dados de mortalidade extraídos do *Sistema de Informações de Mortalidade* (SIM) do DATASUS, a plataforma de tecnologia da informação do Sistema Único de Saúde. O período avaliado será novamente de 2008 a 2013. Segundo diversos estudos (*Chang et al.*, 2017; *Conceição et al.*, 2001a; *Garrett e Casimiro*, 2011; *Saldiva et al.*, 1995; *Schwartz e Dockery*, 1992a), controlaremos as condições climáticas, a tendência e a sazonalidade por meio das seguintes variáveis: temperatura (mínima, média, máxima ou variação), umidade, dia da semana, mês (ou estação do ano), indicador de dia útil e termo tendência. Para os modelos com a concentração de ozônio como preditor, vamos considerar a concentração média na cidade, calculada nas 12 estações de monitoramento consideradas em *Salvo et al.* (2017).

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. Focaremos a análise nesses dois grupos de maior risco.

A Figura 6.1 apresenta as séries da mortalidade diária para crianças e idosos. Observamos que

no período observado há uma leve tendência negativa para as crianças e forte tendência positiva para os idosos. Também notamos o comportamento sazonal da série de idosos, com maiores picos acontecendo nos meses de inverno, ao contrário da concentração de ozônio, cujos picos tendem a acontecer no verão.

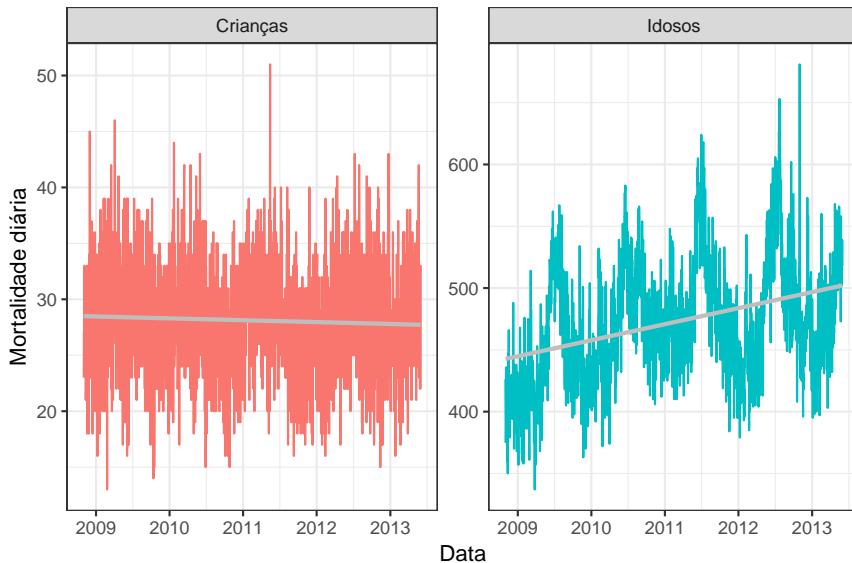


Figura 6.1: Séries da mortalidade diária para crianças e idosos.

A seguir, ajustaremos alguns modelos para avaliar a associação da mortalidade com a proporção estimada de carros a gasolina. Em seguida, ajustaremos modelos utilizando a concentração de ozônio como preditor para comparmos os resultados obtidos.

6.1.1 Modelo linear Poisson

Como a mortalidade diária é um dado de contagem, ajustamos inicialmente um modelo linear generalizado com distribuição Poisson (ver Seção 3.2.3) para avaliar a relação dessa quantidade com a proporção estimada de carros a gasolina rodando na cidade de São Paulo.

A Tabela 6.1 apresenta os resultados dos modelos ajustados para idosos e crianças. Observamos que a proporção estimada de carros a gasolina é estatisticamente significante para explicar a mortalidade geral, sendo que um aumento de 10% na proporção de carros a gasolina está associado a um aumento médio de 108% na mortalidade em idosos. Analisando os resultados para as crianças, os preditores considerados não foram suficiente para explicar a mortalidade, já que apenas 2.64% da variação foi explicada. Neste caso, a proporção de carros a gasolina não foi considerada estatisticamente significativa.

Tabela 6.1: Resultados do modelo Poisson para mortalidade geral em crianças, idosos e na população como um todo.

Grupo	RMSE	% var. explicada	Valor-p para a proporção de carros a gasolina	Variação na mortalidade
Idosos	31.04	59.30	< 0.001	108%
Crianças	5.13	2.64	0.19	-

Na Figura 6.2, apresentamos o gráfico dos valores preditos contra os valores observados para

cada modelo. Para as crianças, a associação entre essas duas quantidades é bem baixa, ilustrando o baixo poder preditivo do modelo. Para os idosos, o modelo erra mais para valores baixos da mortalidade.

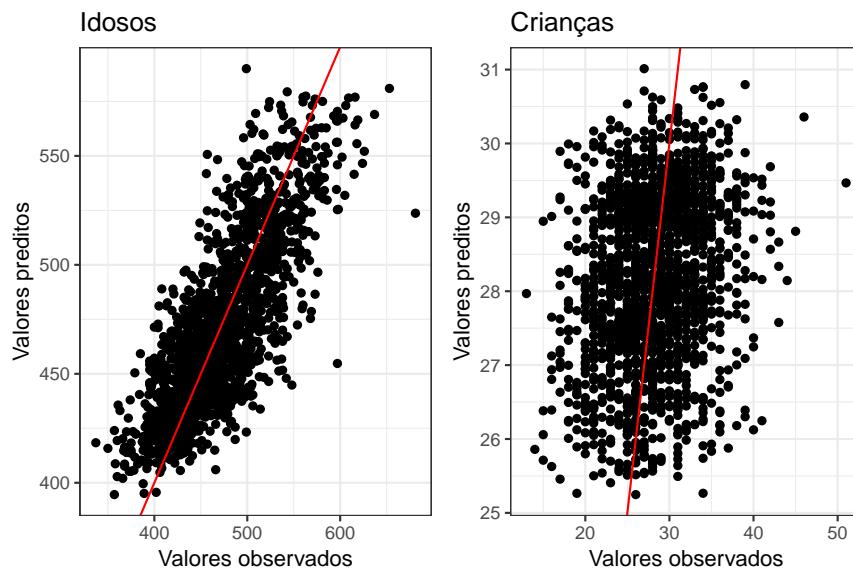


Figura 6.2: Gráfico dos valores preditos versus os valores observados para o modelo linear generalizado com distribuição Poisson para cada um dos grupos.

Para avaliar a melhor escolha de preditores, ajustamos modelos com mês ou estação do ano para controlar a sazonalidade e temperatura média, mínima, máxima ou variação de temperatura ao longo do dia para controlar a temperatura. Utilizando o RMSE como métrica de performance, o melhor modelo para os 2 grupos foi aquele com os preditores mês e temperatura média.

A seguir, vamos ajustar um modelo aditivo generalizado para verificar possíveis relações não lineares são melhores ajustadas, gerando resultados mais precisos.

6.1.2 Modelo aditivo Poisson

Na tentativa de obtermos estimativas mais precisas do que as obtidas pelos modelos lineares generalizados, caso exista relações não lineares entre as variáveis, ajustamos agora modelos aditivos generalizados com distribuição Poisson (ver Seção 3.3) para associar a mortalidade geral com a proporção estimada de carros a gasolina.

Os resultados para cada um dos grupos considerados se encontram na Tabela 6.2. Observamos aumento na performance do modelo (menor RMSE) e poder preditivo (maior R^2) para os idosos. Além disso, a proporção de carros a gasolina continua estatisticamente significativa para explicar a variação da mortalidade.

Tabela 6.2: Resultados do modelo aditivo Poisson para mortalidade geral em crianças, idosos e na população como um todo.

Grupo	RMSE	% var. explicada	Valor-p para a proporção de carros a gasolina
Idosos	29.21	64.78	< 0.001
Crianças	5.10	3.4	0.50

Observando agora os gráficos das funções não-lineares ajustadas para a proporção de carros a

gasolina (Figura 6.3), notamos que a associação com a mortalidade, antes positiva, agora parece ser negativa. De uma forma geral, esses gráficos apontam que, para os idosos, quando aumentamos a proporção de carros a gasolina, a mortalidade tende a diminuir, com valor mínimo quando a proporção está próxima a 68%. Esses indícios contrariam os encontrados no modelo linear da seção anterior. No entanto, uma ressalva sobre essa conclusão deve ser feita. As regiões do gráfico onde as curvas decaem são aquelas com menos observações (isso pode ser observado pela maior amplitude dos intervalos de confiança). Além disso, pela Figura 5.3, notamos que a distribuição da proporção de carros a gasolina apresenta uma sazonalidade: os maiores valores tendem a acontecer no verão e os menores valores no inverno. Com a mortalidade, essa relação é inversa: a mortalidade é maior no inverno. Assim, se essa associação entre mortalidade e proporção de carros a gasolina for espúria, isso explicaria o comportamento das funções estimadas pelo modelo aditivo.

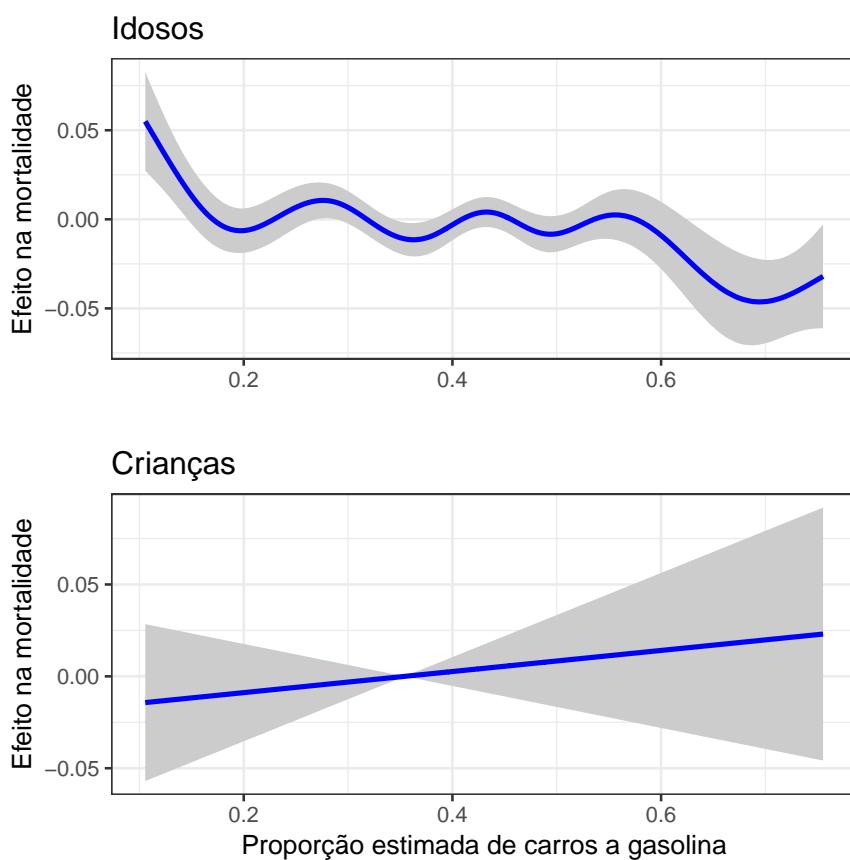


Figura 6.3: Funções estimadas para a proporção estimada de carros a gasolina pelo modelo aditivo Poisson para cada grupo.

Na próxima seção, ajustaremos uma *random forest* para avaliarmos qual conclusão podemos obter de um modelo precisamente mais preciso.

6.1.3 Random forest

Como o modelo linear e aditivo apontaram conclusões diferentes sobre o efeito da proporção estimada de carros a gasolina na mortalidade de idosos, vamos agora ajustar uma *random forest* na tentativa de obter um ajuste mais preciso para compararmos os resultados. A partir daqui, ajustaremos os modelos apenas para o grupo de idosos, tendo vista a impossibilidade de explicar a mortalidade geral a partir dos preditores considerados.

A *random forest* ajustada obteve um RMSE de 27.69 (contra 31.04 do modelo linear e 29.21 do modelo aditivo) e uma variação da proporção explicada igual a 67.43% (contra 59.30% do modelo linear e 64.78% do modelo aditivo). Na Figura 6.4 apresentamos os gráficos de dependência parcial e de efeitos acumulados para a variável proporção estimada de carros a gasolina. A associação deste preditor na mortalidade parece ser, de uma forma geral, negativa: quanto maior a proporção de carros a gasolina, menor a mortalidade. Um padrão semelhante foi indicado pelo modelo aditivo (Figura 6.3). Analisando o gráfico com mais cuidado, observamos que o efeito na mortalidade estimada tem duas quedas: uma quando variamos o preditor aproximadamente de 10 para 20% e outra quando variamos aproximadamente de 40 para 50%. O efeito também volta a subir para valores maiores de 50%. É difícil encontrarmos uma explicação prática para essa relação estranha (altamente não-linear), sendo que ela pode indicar dois problemas: a relação entre as variáveis é, de fato, complexa e difícil de ser modelada e interpretada ou, assim como observamos no capítulo anterior, a distribuição desbalanceada da proporção de carros a gasolina pode estar gerando correlações espúrias.

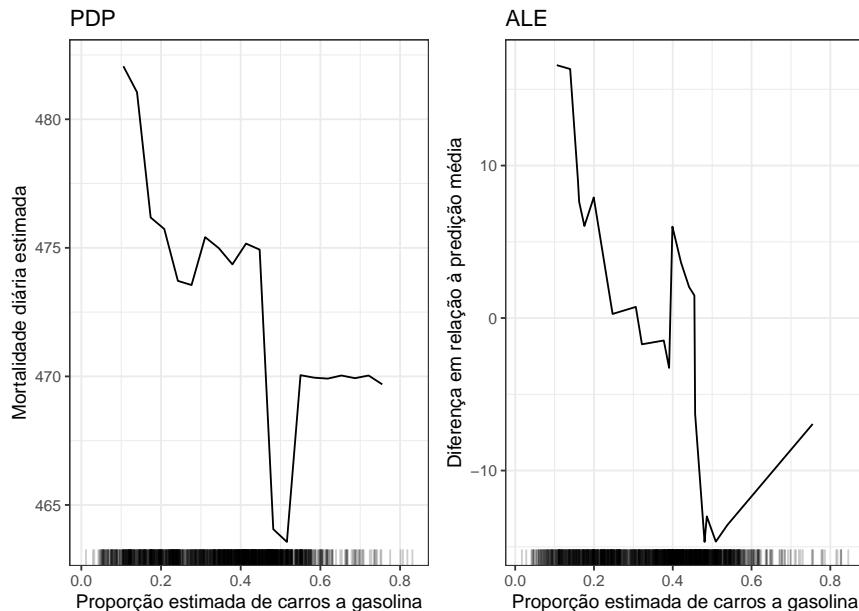


Figura 6.4: Gráficos de dependência parcial (PDP) e de efeitos acumulados (ALE) da random forest para a variável proporção estimada de carros a gasolina.

A seguir, apresentaremos algumas análises adicionais realizadas na tentativa de entender melhor a relação da proporção de carros a gasolina com a mortalidade.

6.1.4 Análises complementares

Para testar a sensibilidade dos resultados obtidos pelo modelo anterior, nós realizamos análises complementares, variando algumas variáveis mantidas fixas e testando outros modelos. Focamos essas análises apenas no grupo de idosos pois nenhum dos modelos considerados sugeriu associação entre os preditores a mortalidade infantil.

Inicialmente, avaliamos se o efeito da proporção estimada de carros a gasolina na mortalidade pode estar defasado no tempo, isto é, se o valor deste preditor em um determinado dia está associado com o número de mortes em idosos no futuro. Para isso, construímos o gráfico da função de correlação

cruzada entre essas variáveis com defasagem até 60 dias (Figura 6.5) e observamos indícios de correlação para defasagens maiores que 20. Embora a forma desse gráfico sugira apenas um efeito sazonal, vamos ajustar a *random forest* considerada na seção anterior com a proporção estimada de carros a gasolina defasada em 20, 25, 30, 35, 40, 45, 50, 55 e 60 dias. Os resultados estão apresentados na Tabela 6.3. Observamos que não houve redução substancial do erro do modelo em comparação o ajuste sem defasagem (RMSE igual a 27.69) e não há consistência na ordem das melhores defasagens.

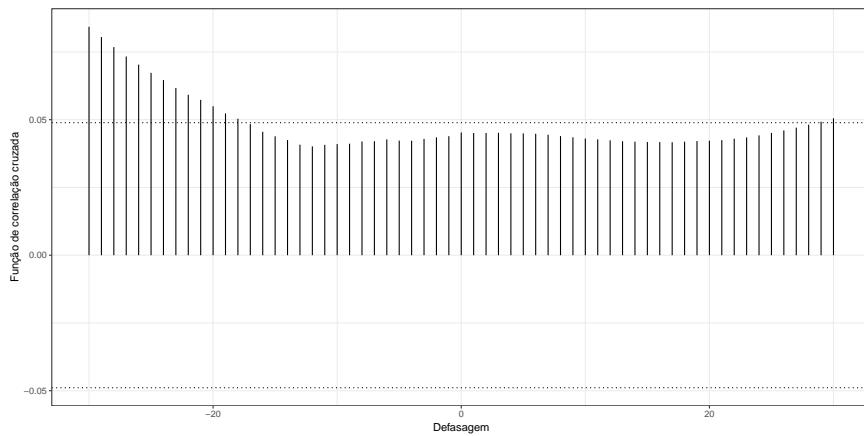


Figura 6.5: Gráfico de correção cruzada da mortalidade diária para idosos contra a proporção estimada de carros a gasolina.

Tabela 6.3: Resultados do modelo da Random forest com proporção estimada de carros a gasolina defasada no tempo. Os valores estão ordenados do menor ao maior RMSE.

Grupo	Defasagem	RMSE	% var. explicada
Idosos	30	27.30	68.55%
	20	27.39	68.52%
	45	27.45	68.18%
	25	27.49	68.29%
	35	27.59	67.88%
	60	27.62	67.49%
	50	27.64	67.59%
	55	27.64	67.64%
	40	27.78	67.48%

O gráfico de dependência parcial e de efeitos locais acumulados para o modelo com proporção estimada de carros a gasolina defasada em 30 dias (Figura 6.6) indica a mesma relação do que o modelo ajustado sem defasagem.

Os modelos linear e aditivo com distribuição Poisson, assim como a *random forest*, também foram ajustados utilizando a semana do ano (em vez do mês do ano) para o controle da sazonalidade e máxima, mínima e variação diária da temperatura (em vez da temperatura média), mas as conclusões obtidas para cada modelo foram as mesmas.

Também ajustamos um *XGBoost* para a mortalidade em idosos, mas os resultados não foram melhor do que os da *random forest*: RMSE igual a 28.13 e 67.45% da variação total da mortalidade explicada. A interpretação do efeito da proporção estimada de carros a gasolina foi similar ao encontrado na *random forest*.

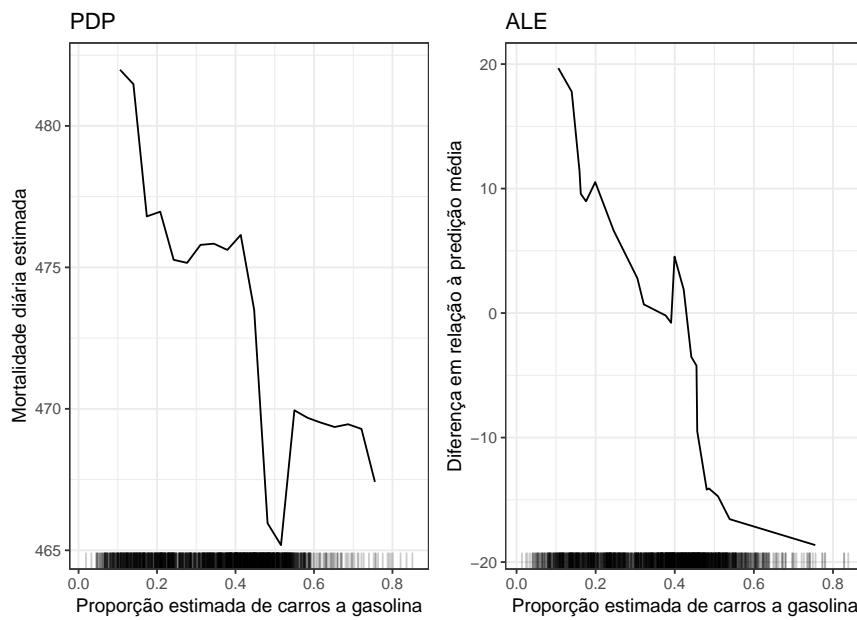


Figura 6.6: Gráficos de dependência parcial (PDP) e de efeitos acumulados (ALE) da random forest para a variável proporção estimada de carros a gasolina defasada em 3 dias.

6.2 Concentração de ozônio e mortalidade

Nesta seção, apresentaremos os resultados dos modelos ajustados para a mortalidade diária utilizando a concentração de ozônio diretamente como preditor. O objetivo é comparar esses resultados com aqueles obtidos nas seções anteriores para avaliar se a proporção estimada de carros a gasolina estaria representando bem o efeito da poluição por ozônio na mortalidade.

Os modelos ajustados aqui consideraram como controle os seguintes preditores: temperatura, umidade, tendência, mês do ano, dia da semana e variável indicadora de dia útil. A concentração de ozônio utilizada foi a média diária para toda a cidade, calculada a partir das medidas das 12 estações consideradas por [Salvo et al. \(2017\)](#).

A Tabela 6.4 apresenta os resultados dos modelos ajustados. Observamos que todos os modelos indicam uma associação positiva entre mortalidade em idosos e concentração de ozônio. Para o modelo linear, um aumento de $10 \mu\text{g}/\text{m}^3$ está associado com um aumento médio de 0.44% na mortalidade de idosos. A relação para o modelo aditivo pode ser observada pelo gráfico da função não linear apresentado na Figura 6.7 e a interpretação do efeito da concentração de ozônio para a *random forest* se encontra na Figura 6.8. Mais uma vez, as variáveis consideradas não explicam a variabilidade da mortalidade infantil.

Tabela 6.4: Resultados dos modelos ajustados para a mortalidade diária utilizando a concentração de ozônio como preditor.

Modelo	Grupo	RMSE	% var. explicada	Relação com a mortalidade (valor-p)
Linear Poisson	Idosos	31.10	59.10	Positiva (< 0.001)
	Crianças	5.13	2.56	Não há (0.59)
Aditivo Poisson	Idosos	29.29	64.60	Positiva (0.05)
	Crianças	5.11	2.90	Não há (0.97)
<i>Random forest</i>	Idosos	27.85	68.00	Positiva
	Crianças	5.11	2.80	Não há

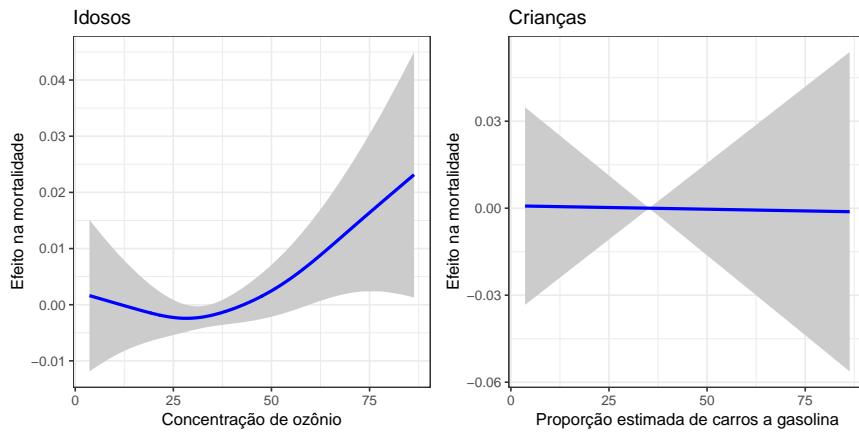


Figura 6.7: Gráfico da função não-linear estimada pelos modelos aditivos generalizados para a proporção estimada de carros a gasolina.

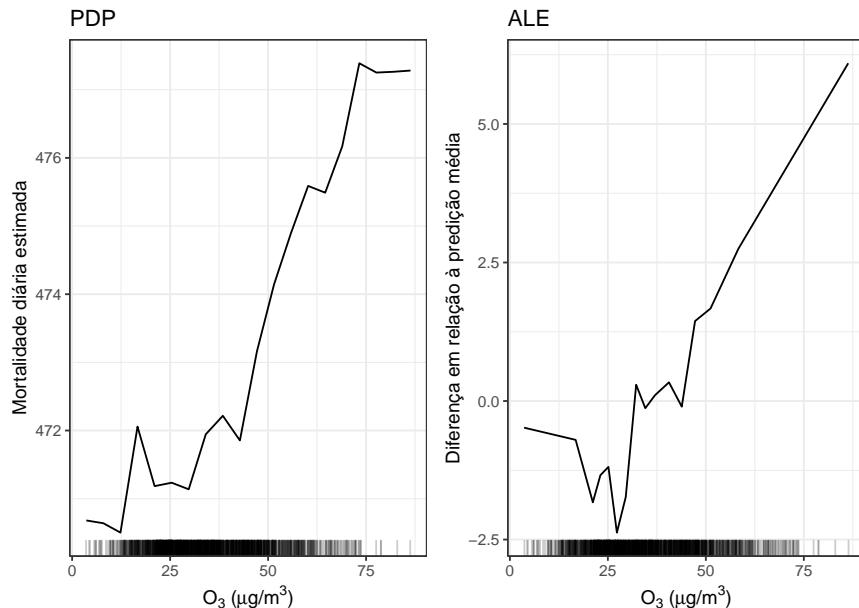


Figura 6.8: Gráficos de dependência parcial (PDP) e de efeitos acumulados (ALE) da random forest (idosos) para a concentração de ozônio.

Os resultados apresentados aqui indicam uma relação bem clara entre a concentração de ozônio e mortalidade. Isso sugere que, se a proporção de carros a gasolina e a concentração de ozônio tivessem uma relação simples (linear, por exemplo) esperaríamos que os modelos ajustados na seção anterior reproduzissem uma relação próxima a encontra aqui (mesmo que contrária) entre a proporção estimada de carros a gasolina e a mortalidade em idosos.

Podemos concluir que a concentração de ozônio está associado com um aumento na mortalidade em idosos e, embora o modelo aditivo Poisson e a *random forest* indiquem uma associação negativa entre proporção de carros a gasolina e a mortalidade (principalmente por meio do modelo com o preditor defasado em 3 dias), a forma dessa relação não é muito clara, o que deixa difícil encontrar uma explicação prática e distinguir esse possível efeito de uma correlação espúria.

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

Duas etapas cruciais da análise de dados são a coleta e a estruturação dos dados. Na maioria dos estudos, a obtenção de dados requer a realização de experimentos, medições ou aplicação de questionários. Não é raro encontrarmos trabalhos comprometidos por falhas na coleta, seja por ausência de variáveis importantes, por má especificação da população alvo, por falta de randomização ou questionários mal construídos. Na Estatística, as áreas de amostragem, planejamento de experimentos e teoria da resposta ao item dão atenção especial à coleta de dados, criando delineamentos amostrais a depender do objetivo do estudo.

A estruturação dos dados consiste na transferência dos registros obtidos na coleta para uma base de dados retangular¹. Para diminuir o tempo e esforço gastos nessa etapa, que muitas vezes chega a ser a parte mais demorada da análise estatística, é essencial estar claro como a base deve estar estruturada para a análise e dispor de ferramentas que auxiliem a execução dessa tarefa. Na linguagem R, os pacotes `janitor`, `tidyverse` e `dplyr` possuem funções especializadas em limpeza, manipulação e transformação de dados.

Em estudos de poluição do ar, a coleta de dados é realizada principalmente por experimentos laboratoriais ou por meio de instrumentos de medição colocados em vias de grande movimento, túneis, parques, próximos a fábricas e outros locais de interesse. Também é comum a instalação de estações de monitoramento automático que medem diversos parâmetros periodicamente. Essas estações geralmente são controladas por órgãos ambientais, que disponibilizam os dados gratuitamente² pela internet³.

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

²No Brasil. Em outros países, pode ser necessário pagar para a obtenção dos dados.

³Alguns portais, como o do Instituto Nacional de Meteorologia (INMET), requerem uma solicitação informando os dados desejados. Após o pedido ser processado, os dados são enviados por e-mail ou, quando o volume é muito grande, são postados em uma mídia física para o endereço do solicitante.

A obtenção de dados já coletados pela internet, no entanto, nem sempre é uma tarefa simples, em especial quando o volume de informação que precisamos baixar é muito grande. Embora dificilmente haja interesse político em dificultar o acesso desses dados, como muitas vezes acontece com dados públicos governamentais e de tribunais, o acesso a eles nem sempre é construído de maneira ótima para quem vai analisá-los. Além disso, raramente a base se encontra formatada para análise, sendo preciso passar também por uma etapa de estruturação.

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*⁴ 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.

Neste capítulo, discutiremos os conceitos básicos de web scraping e apresentaremos alguns portais para se obter dados meteorológicos e de poluição do ar, tanto no Brasil quanto em outros lugares do mundo.

7.1 Web scraping

Web scraping é a tarefa de se extraír dados da internet de forma automatizada. Hoje em dia é muito comum termos acesso rápido e fácil a qualquer conjunto de informações pela web, mas raramente esses dados estão estruturados e em uma forma de fácil obtenção pelo usuário.

Fazer *web scraping* é necessário quando os dados são disponibilizados publicamente, mas o acesso manual a eles é exaustivo ou impraticável, como, por exemplo, quando queremos baixar uma série de 30 anos de um poluente, mas os dados são disponibilizados mês a mês, com cada arquivo em uma página diferente. Quando os dados não são públicos, a construção de *scrapers* deve levar em conta os termos de uso da página, pois algumas não permitem a extração dos dados ou o acesso via algoritmo. Também é recomendável sempre verificar se o órgão ou a empresa já não possui uma API (*Application Programming Interface*), isto é, um sistema criado para facilitar o acesso de terceiros aos dados.

O fluxo do web scraping, como podemos observar no diagrama a seguir (Figura 7.1), é composto por seis etapas: identificar, navegar, replicar, parsear, validar e iterar.

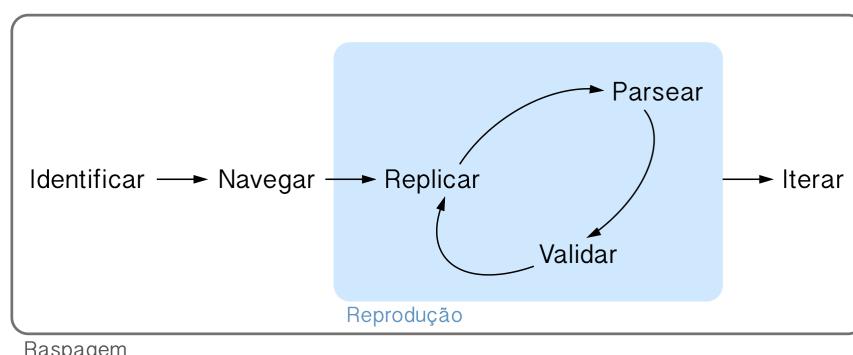


Figura 7.1: O fluxo do web scraping.

⁴Ou raspagem de dados web.

A seguir, descreveremos de forma geral cada uma dessas etapas.

Identificar

No primeiro passo do fluxo, precisamos identificar a informação que vamos coletar, isto é, entender bem qual é a estrutura das páginas que queremos raspar e traçar um plano para extrair tudo que precisamos.

Se, por exemplo, estamos interessados em uma tabela que aparece no corpo de diversas páginas web, precisamos listar todas as páginas que devem ser acessadas (definir o conjunto de *links* que serão acessados) e avaliar se essa tabela sempre aparece com o mesmo formato.

Navegar

O objetivo desta etapa é descobrir qual e que tipo de requisição é feita para o servidor que hospeda o site gerar os dados que queremos extrair.

Esta etapa exige algum conhecimento de programação web, pois consiste em usar ferramentas de desenvolvedor do navegador para encontrar a fonte dos dados a partir das chamadas HTTP ou dos resultados das funções JavaScript.

Replicar

Se tivéssemos que fazer várias requisições HTTP para chegar até a informação que queremos, seria nesta etapa que tentaríamos replicar essas chamadas. Aqui, é necessário compreender absolutamente tudo que a página está fazendo para trazer o conteúdo até você, como a existência de parâmetros, cookies, *tokens* etc.

No R, é possível fazer requisições GET e POST a partir das funções `GET()` e `POST()` do pacote `httr`.

Parsear

O anglicismo *parsear* vem do verbo *to parse*, que quer dizer algo como analisar ou estudar, mas que, no contexto do web scraping, significa extraír os dados desejados de um arquivo HTML. Esta etapa é essencialmente dependente da estrutura de dados que está sendo baixada e de como ela foi disponibilizada na página.

Validar

Se tudo ocorreu bem, validar os resultados será uma tarefa simples. Precisamos apenas reproduzir o procedimento descrito até agora para algumas outras páginas de modo a verificar se estamos de fato extraíndo corretamente tudo o que queremos.

Caso encontremos algo de errado precisamos voltar ao terceiro passo, tentar replicar corretamente o comportamento do site e parsear os dados certos nas páginas.

Iterar

O último passo consiste em colocar o *scraper* em produção. Aqui, ele já deve estar funcionando corretamente para todos os casos desejados e estar pronto para extraír todos os dados que

precisamos.

7.2 Dados no Brasil

Dados de poluição do ar no Brasil geralmente são disponibilizados pelas Órgãos Estaduais de Meio Ambiente, sendo que o acesso a esses dados, em geral, pode ser feito direta ou indiretamente no portal de cada órgão. Dos 27 estados brasileiros, apenas 9 monitoram a qualidade do ar: Bahia, Espírito Santo, Minas Gerais, São Paulo, Rio de Janeiro, Rio Grande do Sul, Paraná, Goiás e Distrito Federal.

Uma solução integrada para acessar esses dados foi desenvolvida pelo Instituto de Energia e Meio Ambiente (IEMA), que compilou os dados de monitoramento dos órgãos ambientais de todo o país em uma plataforma unificada e acessível. A [Plataforma de Qualidade do Ar](#), em sua primeira versão, compilou os valores anuais da concentração de poluentes atmosféricos, criando uma base de dados entre os anos de 2000 e 2014 de todos os estados que disponibilizam tais informações. Com o objetivo de explorar em mais detalhes os poluentes monitorados, a plataforma passou a incorporar dados horários e diários a partir de 2015.

Na Figura 7.2, apresentamos um mapa das estações de monitoramento disponíveis na plataforma do IEMA. Fica claro a falta de dados sobre as regiões Norte e Nordeste do país.



Figura 7.2: Mapa de estações de monitoramento disponíveis na Plataforma de Qualidade do Ar do Instituto de Energia e Meio Ambiente.

Para o estado de São Paulo, a CETESB (Companhia Ambiental do Estado de São Paulo) oferece um sistema de consulta de medidas em tempo real e de relatórios diários, mensais e anuais. O sistema, chamado *Qualar*, também permite a exportação de diversas séries históricas de poluentes

e parâmetros meteorológicos, além de informações sobre as estações de monitoramento.

Para facilitar o acesso aos dados, já que o sistema possui restrições estruturais (como baixar dados de vários poluentes ou estações de uma única vez) e também pode se tornar lento quando precisamos acessar séries muito longas, nós criamos o pacote `koffing`, na linguagem R. O processo se resume ao uso da função `scraper_cetesb()`, que tem, entre outros argumentos, o parâmetro a ser baixado, a estação medidora e o login e senha de acesso ao Qualar. Dada uma lista de parâmetros e estações, a função pode ser utilizada dentro de um *looping* para baixar os dados de diversas estações e poluentes automaticamente. Para instalar o pacote, basta rodar o seguinte comando `devtools::install_github("atmoschem/koffing")`.

Dados de emissão podem ser obtidos no portal do Sistema de Estimativas de Emissões e Remoções de Gases de Efeito Estufa (SEEG), que produz estimativas anuais das emissões de gases de efeito estufa (GEE) no Brasil e documentos analíticos sobre a evolução das emissões. O SEEG avalia os cinco setores que são fontes de emissões — Agropecuária, Energia, Mudanças de Uso da Terra, Processos Industriais e Resíduos. Os dados constituem uma série que cobre o período de 1970 até 2017, exceto para o setor de Mudança de Uso da Terra que tem a série de 1990 a 2017. São considerados todos os gases de efeito estufa contidos no inventário nacional como CO₂, CH₄, N₂O e os HFC.

A maioria dos órgãos de monitoramento ambiental também disponibilizam dados climáticos, como temperatura, radiação solar, umidade, velocidade e direção do vento e precipitação. Bases mais consolidadas podem ser encontradas no [portal do Instituto Nacional de Meteorologia \(INMET\)](#).

A seguir, apresentaremos os principais portais com acesso a dados internacionais de clima e poluição.

7.3 Dados em outros países

Lançado em março de 2000, o programa MOPPIT (*Measurements Of Pollution In The Troposphere*) lançado pela NASA tem como objetivo medir o monóxido de carbono troposférico em escala global. Os dados podem ser baixados diretamente do seguinte site: <https://search.earthdata.nasa.gov/>.

A NASA também possui outros canais de visualização e disponibilização de dados, como o portal [Atmospheric Science Data Center](#), para dados atmosféricos, e o ambiente [Giovanni](#), para parâmetros geofísicos.

O portal suíço [AirVisual](#) disponibiliza visualizações, métricas e previsões para dados de material particulado e meteorológicos em mais de 10000 pontos de monitoramento espalhadas em todo o mundo (Figura 7.3).

Nos EUA, a Agência de Proteção Ambiental (EPA) é o órgão federal que regulamenta e monitora os níveis de poluição da terra, água e ar com o objetivo de proteger a saúde humana e o meio ambiente. Em conjunto com outras agências ambientais, a EPA criou o [Airnow](#), uma plataforma de monitoramento da qualidade do ar com informações horárias de ozônio e material particulado para todos os estados americanos, totalizando mais de 400 cidades.

Na Europa, a Agência Europeia de Meio Ambiente (EEA) é a responsável por implementar as diretivas da União Europeia com respeito ao controle de emissões e à qualidade do ar. A agência mantém um [portal de monitoramento](#) horário de ozônio, material particulado, dióxido de nitrogênio,



Figura 7.3: Exemplo de visualização do portal AirVisual para a estação Parque Dom Pedro II, em São Paulo.

dióxido de enxofre e monóxido de carbono para diversas cidades em toda a Europa. Os dados do portal podem ser baixados na página de [Air Quality e-Reporting](#).

Capítulo 8

Discussão

Durante a última década, novas tecnologias vêm mudando bastante a forma como estatísticos encaram a análise de dados. A utilização de computadores, instrumentos de medição cada vez mais sofisticados e a internet geram um volume de informação cada vez maior, abrindo novas possibilidades em áreas que historicamente sofriam com a escassez de dados. As preocupações que antes giravam em torno de encontrar o modelo mais adequado para os dados e a derivar resultados assintóticos para a construção de intervalos de confiança e testes de hipóteses, foram em certo grau substituídas pelo desenvolvimento de algoritmos que melhor capturam os padrões escondidos nos dados e que possam ser generalizados para além da amostra. Além disso, a necessidade de extrair dados da internet, manipular bases gigantescas e comunicar resultados de forma atraente passaram a frequentar o radar do estatístico.

Essa nova *forma* de se fazer Estatística chamada *Data Science* incorporou diversas ideias interessantes à análise de dados, mas o seu principal requisito é um que sempre fez parte da formação dos estatísticos: a análise crítica. Entender o fenômeno de interesse, o objetivo do estudo, as restrições de cada técnica e ser capaz de avaliar criticamente os resultados é essencial independentemente da abordagem de análise que estamos seguindo.

O aspecto multidisciplinar faz os estudos de poluição do ar um exemplo claro disso. Os principais desafios nessa área envolvem a modelagem de fenômenos naturalmente complexos, e a utilização de modelos muito simples para representá-los pode levar a conclusões no mínimo superficiais e, por consequência, políticas públicas insuficientes.

Neste trabalho, mostramos que os resultados encontrados por Salvo e Geiger (2014) e Salvo *et al.* (2017) sobre a associação entre a concentração de ozônio e a proporção de carros a gasolina rodando na cidade de São Paulo podem levar a algumas conclusões equivocadas. O modelo de regressão linear utilizado não captura a relação não linear que modelos mais flexíveis sugeriram existir entre as variáveis. Embora as análises realizadas apontarem que, de fato, o aumento do uso de etanol estar associado com maiores médias diárias de ozônio, entender a forma dessa relação é importante para uma maior compreensão do mecanismo gerador do ozônio e para a criação de estratégias de redução de emissões.

A ausência de padrão no efeito da proporção estimada de carros a gasolina indicada pelo *XG-Boost*, o modelo que resultou no melhor ajuste, levanta dúvidas sobre a possibilidade de encontrar um modelo simples e interpretável para explicar a associação desse preditor com a concentração de ozônio e também nos faz questionar qual precisão essa medida representa o uso de etanol em cada região da cidade. Apesar de Salvo e Huse (2013) terem realizado um trabalho bem cuidoso para

estimar essa quantidade, algumas das suposições feitas são muito fortes, e podem estar gerando bastante ruído quando utilizamos essas estimativas para explicar a concentração de ozônio.

Primeiro, estamos supondo que a proporção de carros a gasolina é a mesma para toda a cidade, o que pode ser falso principalmente para uma metrópole com as dimensões e as desigualdades sociais de São Paulo. Também assumimos que o efeito dessa variável é o mesmo em todas as regiões, o que pode não ser verdade como mostramos ao ajustar um modelo para cada estação de monitoramento. Por fim, não são consideradas medidas de evaporação do etanol nos instantes de abastecimento, o que seria um preditor importante pois é uma das principais fontes de compostos voláteis orgânicos, um dos elementos responsáveis pela formação do ozônio troposférico.

Além disso, também expandimos a análise realizada por *Salvo et al. (2017)* investigando o efeito da proporção estimada de carros a gasolina e da concentração de ozônio na mortalidade geral em crianças e idosos. Os modelos considerando o uso de gasolina/etanol sugeriram uma associação negativa deste preditor com a mortalidade em idosos. No entanto, mais uma vez os ajustes não identificaram uma relação clara entre as variáveis e, somado ao comportamento sazonal da proporção estimada de carros a gasolina (com maiores valores tendendo a acontecer no verão, ao contrário da mortalidade), é difícil dizer se a associação encontrada não é espúria. Assim como esperado, dada a farta literatura sobre o tema, todos os modelos considerando a concentração de ozônio como variável explicativa sugeriram uma associação positiva entre esse preditor e a mortalidade em idosos. Nenhum dos modelos considerados ajustou bem a mortalidade em crianças, indicando que os preditores considerados não são suficientes para explicar essa variável.

Os resultados nas duas análises mostram, sobretudo, a dificuldade de se analisar dados de poluição do ar, as diversas estratégias existentes para abordar cada problema e o quanto devemos ser cuidados ao interpretar os resultados. Em problemas tão complexos, dificilmente vamos conseguir respostas simples, sendo preferível na maioria das vezes levantar hipóteses em vez de chegar em conclusões.

Esta tese buscou ser uma ponte entre a Estatística e as outras disciplinas que compõem o estudo da poluição do ar. O compartilhamento de conhecimento de forma acessível é essencial para que haja a colaboração entre pesquisadores, direcionando os trabalhos na direção certa. Este esforço continuará, pois há muitos tópicos que não foram abordados aqui, como estudos de previsão, estratégias de imputação de dados ou o uso de modelos de redes neurais. Além disso, novas técnicas e modelos vêm surgindo todos os dias, demandando que a construção dessa ponte seja um processo contínuo, e não parado no tempo.

Ao passo que a ciência cria novas tecnologias, a tecnologia também muda o jeito como fazemos ciência. A Estatística não é diferente.

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. 37

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. 44

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. 34, 45

Bollerslev(1986) Tim Bollerslev. Generalized autoregressive conditional heteroscedasticity. *J. Econ.*, 31: 307–327. Citado na pág. 54

Box e Jenkins(1970) George E. P. Box e Gwilym M. Jenkins. *Time series analysis: forecasting and control*. Holden-Day, San Francisco. Citado na pág. 48

Breiman(2001) Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–231. Citado na pág. 29, 30

Breiman e Friedman(1985) Leo Breiman e Jerome 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. 46

Buhr et al.(1992) M. P. Buhr, M. Trainer, D. D. Parrish, R. E. Sievers e F. C. Fehsenfeld. Assessment of pollutant emission inventories by principal component analysis of ambient air measurements. *Geophysical Research Letters*, 19(10): 1009–1012. doi: 10.1029/92GL01020. Citado na pág. 51

Bussab e Morettin(2013) W. O. Bussab e P. A. Morettin. *Estatística Básica*. Editora Saraiva, São Paulo. Citado na pág. 4

Cameron e Miller(2015) Adrian C. Cameron e Douglas L. Miller. A practitioner's guide to cluster-robust inference. *J. Human Resources*, 50(2): 317–372. Citado na pág. 36

Carlaw et al.(2007) David C. Carslaw, Sean D. Beavers 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. 2

- Casella e Berger(2001)** George Casella e Roger L. Berger. *Statistical Inference*. Duxbury Press; 2nd edition. Citado na pág. 31, 32, 41
- 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 PM_{2.5} 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. 104
- Chavent et al.(2009)** Marie Chavent, Hervé Guegan, Vanessa Kuentz, Brigitte Patouille e Jérôme Saracco. PCA and PMF based methodology for air pollution sources identification and apportionment. 20: 928–942. URL <https://hal.archives-ouvertes.fr/hal-00332015/file/Env-2009-preprint.pdf>. Citado na pág. 51
- Chuang et al.(2011)** Ya-Hsiu Chuang, Sati Mazumdar, Taeyoung Park, Gong Tang, Vincent. C. Arena e Mark J. Nicolich. Generalized linear mixed models in time series studies of air pollution. *Atmospheric Pollution Research*, 2(4): 428 – 435. URL <http://www.sciencedirect.com/science/article/pii/S1309104215304694>. Citado na pág. 54
- Conceição et al.(2001a)** Gleice M. S. Conceição, Simone G. E. K. Miraglia, Humberto S. Kishi, Paulo H. N. Saldiva e Julio M. 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, 104
- Conceição et al.(2001b)** Gleice M. S. Conceição, Paulo H. N. Saldiva e Julio M. Singer. Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbimortalidade: uma introdução baseada em dados da cidade de São Paulo. *Revista Brasileira de Epidemiologia*, 4(3): 206–219. Citado na pág. 2, 42, 44
- Cortina-Januchs et al.(2015)** Maria Guadalupe Cortina-Januchs, Joel Quintanilla-Dominguez, Antonio Vega-Corona e Diego Andina. Development of a model for forecasting of pm10 concentrations in salamanca, mexico. *Atmospheric Pollution Research*, 6(4): 626 – 634. ISSN 1309-1042. doi: <https://doi.org/10.5094/APR.2015.071>. URL <http://www.sciencedirect.com/science/article/pii/S1309104215301951>. Citado na pág. 3
- Coull et al.(2001)** Brent Coull, Joel Schwartz e Matt Wand. Respiratory health and air pollution: additivemixed model analyses. *Biostatistics*, 2(3): 337–349. Citado na pág. 54
- 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. 103
- Demidenko(2013)** Eugene Demidenko. *Mixed models: theory and applications with R*. Wiley, New York. Citado na pág. 36
- Dobson(1990)** Annette J. Dobson. *An introduction to generalized linear models*. Chapman and Hall, New York. Citado na pág. 42
- 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. 56
- 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. 7
- Elangasinghe et al.(2014)** Madhavi Anushka Elangasinghe, Naresh Singhal, Kim N. Dirks e Jennifer A. Salmond. Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmospheric Pollution Research*, 5(4): 696 – 708. ISSN 1309-1042. URL <http://www.sciencedirect.com/science/article/pii/S1309104215302786>. Citado na pág. 3

- Engle(1982)** Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50: 987–1007. Citado na pág. 54
- 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. 77
- Everitt e Hothorn(2011)** Brian Everitt e Torsten Hothorn. *An Introduction to Applied Multivariate Analysis with R*. Use R! Springer-Verlag New York. Citado na pág. 53
- 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. 103
- Feng et al.(2015)** Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin e Jingjie Wang. Artificial neural networks forecasting of pm2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, 107: 118 – 128. doi: <https://doi.org/10.1016/j.atmosenv.2015.02.030>. URL <http://www.sciencedirect.com/science/article/pii/S1352231015001491>. Citado na pág. 3
- Fisher(1918)** Ronald Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 652(2): 399–433. Citado na pág. 53
- Galecki e Burzykowski(2013)** Andrzej Galecki e Tomasz Burzykowski. *Linear Mixed-Effects Models Using R*. Springer. Citado na pág. 54
- Garrett e Casimiro(2011)** Pedro Garrett e Elsa Casimiro. Short-term effect of fine particulate matter (pm2.5) and ozone on daily mortality in Lisbon, Portugal. *Environ Sci Pollut Res*, 18: 1585–1582. Citado na pág. 104
- Goodfellow et al.(2016)** Ian Goodfellow, Yoshua Bengio e Aaron Courville. *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>. Citado na pág. 30
- Gower(1971)** C. John Gower. A general coefficient of similarity and some of its properties. *Biometrics*. Citado na pág. 73
- Hastie e Tibshirani(1990)** Trevor Hastie e Robert Tibshirani. *Generalized additive models*. London:Chapman & Hall. Citado na pág. 3, 45, 46, 67
- Hastie et al.(2008)** Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer. Citado na pág. 2, 5, 29, 30, 31, 32, 39, 45, 46, 47, 48, 67, 69
- 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. 103
- Jacobson(2002)** Mark Z. Jacobson. *Atmospheric Pollution History, Science, and Regulation*. Cambridge. Citado na pág. 103
- 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. 78
- 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. 5, 31, 32, 34, 38, 39, 47, 48, 61, 63, 64, 65, 67

- 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. 66
- 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. 55
- 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. 55
- 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.bmjjournals.org/content/50/Suppl_1/S12. Citado na pág. 2
- Kiefer e Wolfowitz(1952)** J. Kiefer e J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statista*, 23: 462–466. Citado na pág. 70
- 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 PM_{2.5} exposures in the mid-atlantic states. *American Chemical Society*, 46: 11913–11921. Citado na pág. 2, 54
- Kumar e Ridder(2010)** Ujjwal Kumar e Koen De Ridder. GARCH modelling in association with FFT-ARIMA to forecast ozone episodes. *Atmospheric Environment*, 44(34): 4252 – 4265. ISSN 1352-2310. doi: <https://doi.org/10.1016/j.atmosenv.2010.06.055>. URL <http://www.sciencedirect.com/science/article/pii/S1352231010005340>. Citado na pág. 55
- Leiva(2015)** Victor Leiva. *The Birnbaum-Saunders distribution*. Academic Press; 1 edition. Citado na pág. 44
- 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. 44
- Liao *et al.*(1999)** Duanping Liao, John Creason, Card Shy, Ron Williams, Randall Watsfs e Roy Zweidinger. Daily variation of particulate air pollution and poor cardiac autonomic control in the elderly. *Environ Health Perspect*, 107(7): 521–525. Citado na pág. 53
- 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. 66
- Madronich(2014)** Sacha Madronich. Ethanol and ozone. *Nature Geoscience: news & views*, 7: 395–397. Citado na pág. 101

- Magalhães e Lima(2013)** Marcos Magalhães e Antonio Carlos Pedroso Lima. *Noções de Probabilidade e Estatística*. Edusp. Citado na pág. 4
- McCulloch e Searle(2001)** Charles E. McCulloch e Shayle R. Searle. *Generalized, linear, and mixed models*. Wiley, New York. Citado na pág. 36
- 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. 17, 49, 50, 51, 54
- Muggeo(2003)** Vito Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22: 3055–3071. Citado na pág. 38, 39
- 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. 77
- Nelder e Wedderburn(1972)** John A. Nelder e R. William M. Wedderburn. Generalized linear models. *Stat Soc A*, 135: 370–384. Citado na pág. 3, 41, 65
- Nicholson(2001)** W. Keith Nicholson. *Elementary Linear Algebra*. McGraw-Hill Ryerson, 2^a edição. Citado na pág. 52
- 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. 36, 42, 43
- 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. 77
- 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. 103
- Polezer et al.(2018)** Gabriela Polezer, Yara S. Tadano, Hugo V. Siqueira, Ana F.L. Godoi, Carlos I. Yamamoto, Paulo A. de AndrÃ©, Theotonio Pauliquevis, Maria de Fatima Andrade, Andrea Oliveira, Paulo H.N. Saldiva, Philip E. Taylor e Ricardo H.M. Godoi. Assessing the impact of pm2.5 on respiratory disease using artificial neural networks. *Environmental Pollution*, 235: 394 – 403. ISSN 0269-7491. doi: <https://doi.org/10.1016/j.envpol.2017.12.111>. URL <http://www.sciencedirect.com/science/article/pii/S0269749117337107>. Citado na pág. 4
- 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. 5
- Ribeiro et al.(2016)** Marco Tilio 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. 72
- Saldiva et al.(1994)** P. Saldiva, A. Lichtenfels, P. Paiva, I. Barone, M. Martins, E. Massad, J. Pereira, V. Xavier, J. Singer e G. 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, 31, 44, 103, 104

- 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](#), [3](#), [8](#), [18](#), [19](#), [35](#), [40](#), [55](#), [63](#), [78](#), [104](#), [119](#)
- Salvo e Huse(2013)** Alberto Salvo e Cristian Huse. Build it, but will they come? Evidence from consumer choice between gasoline and sugarcane ethanol. *Journal of Environmental Economics and Management*, (66): 251–279. Citado na pág. [119](#)
- 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. [78](#)
- 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. [iii](#), [viii](#), [xiv](#), [xvii](#), [2](#), [3](#), [8](#), [9](#), [31](#), [35](#), [63](#), [78](#), [79](#), [80](#), [84](#), [86](#), [87](#), [88](#), [89](#), [92](#), [94](#), [96](#), [101](#), [104](#), [110](#), [119](#), [120](#)
- 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 Dockery(1992a)** Joel Schwartz e Douglas Dockery. Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology*, 135(1): 12–25. Citado na pág. [104](#)
- Schwartz e Dockery(1992b)** Joel Schwartz e Douglas W. Dockery. Particulate air pollution and daily mortality in Steubenville, Ohio. *Am J Epidemiol.*, 1(135): 12–19. Citado na pág. [2](#), [44](#), [103](#)
- 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, 2^a edição. Citado na pág. [17](#), [18](#), [34](#), [48](#), [49](#), [50](#), [51](#), [55](#)
- Singer et al.(2012)** Julio Singer, JuvÃ^ancio Nobre e Francisco Marcelo Rocha. *Análise de Dados Longitudinais (versão parcial preliminar)*. <http://www.ime.usp.br/jmsinger/Textos>, São Paulo. Citado na pág. [39](#), [54](#)
- Stingone et al.(2017)** Jeanette A. Stingone, Om P.Pandey, Luz Claudio e Gaurav Pandey. Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among u.s. children. *Environmental Pollution*, 230: 230–240. Citado na pág. [3](#)
- Tecer(2009)** Lokman Hakan Tecer. A factor analysis study: Air pollution, meteorology, and hospital admissions for respiratory diseases. *Toxicological & Environmental Chemistry*, 91(7): 1399–1411. doi: 10.1080/02772240902732316. Citado na pág. [51](#)
- Thurston e Spengler(1985)** George D. Thurston e John D. Spengler. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan boston. *Atmospheric Environment (1967)*, 19(1): 9 – 25. ISSN 0004-6981. Citado na pág. [51](#)
- WHO(2004)** World Health Organization WHO. Health aspects of air pollution. Relatório Técnico E83080. Citado na pág. [104](#)
- Wickham(2010)** Hadley Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1): 3–28. Citado na pág. [8](#)

- Wickham e Grolemund(2017)** Hadley Wickham e Garrett Grolemund. *R for Data Science*. O'Reilly, 1^a edição. Citado na pág. 5, 7
- Wilkinson(2005)** Leland Wilkinson. *The Grammar of Graphics*. Statistics and Computing. Springer. 2nd edition. Citado na pág. 8
- Williams(1987)** Alistair D. Williams. Generalized linear model diagnostic using the deviance and single case deletion. *Applied Statistics*, 36: 181–191. Citado na pág. 43
- Wood(2006)** Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, 1^a edição. Citado na pág. 44
- Yoon et al.(2009)** S. Yoon, S. Ha, H. Roh e C. 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. 77
- Zannetti(1990)** P. Zannetti. *Air pollution modelling: theories, computational methods and available software*. Springer Science+Business Media, LLC, New York. Citado na pág. 1, 56