

Victor Teixeira de Melo Mayrink

**Avaliação do Algoritmo Gradient Boosting em Aplicações de Previsão de  
Carga Elétrica a Curto Prazo**

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Orientador: Prof. D.Sc. Henrique Steinherz Hippert

Juiz de Fora

2016

Ficha catalográfica elaborada através do programa de geração automática da Biblioteca Universitária da UFJF, com os dados fornecidos pelo(a) autor(a)

Mayrink, Victor Teixeira de Melo.

Avaliação do algoritmo Gradient Boosting em aplicações de previsão de carga elétrica a curto prazo / Victor Teixeira de Melo Mayrink. -- 2016.

91 f.

Orientador: Henrique Steinherz Hippert

Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2016.

1. Previsão de carga elétrica. 2. Amortecimento exponencial. 3. Árvores de decisão. 4. Gradient Boosting. I. Hippert, Henrique Steinherz, orient. II. Título.

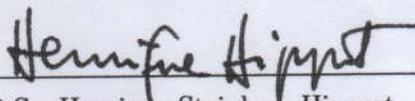
Victor Teixeira de Melo Mayrink

**Avaliação do Algoritmo Gradient Boosting em Aplicações de Previsão de  
Carga Elétrica a Curto Prazo**

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Aprovada em 31 de Agosto de 2016.

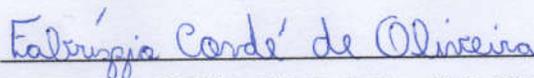
BANCA EXAMINADORA



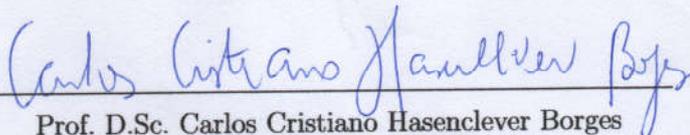
Prof. D.Sc. Henrique Steinherz Hippert - Orientador  
Universidade Federal de Juiz de Fora



Prof. Ph.D. Carlos Eduardo Pedreira  
Universidade Federal do Rio de Janeiro



Prof. D.Sc. Fabrício Condé de Oliveira  
Universidade Salgado de Oliveira



Prof. D.Sc. Carlos Cristiano Hasenclever Borges  
Universidade Federal de Juiz de Fora

*Aos meus pais.*

## AGRADECIMENTOS

Agradeço à Deus. Aos meus pais, pelo grande exemplo de vida. Ao Henrique, pela orientação na elaboração deste trabalho, mesmo durante seu período de férias. Aos amigos da pós-graduação, pelas trocas de experiências e companheirismo. À FAPEMIG pelo suporte financeiro.

*“Essentially, all models are wrong,  
but some are useful”*

George E. P. Box

## RESUMO

O armazenamento de energia elétrica em larga escala ainda não é viável devido a restrições técnicas e econômicas. Portanto, toda energia consumida deve ser produzida instantaneamente; não é possível armazenar o excesso de produção, ou tampouco cobrir eventuais faltas de oferta com estoques de segurança, mesmo que por um curto período de tempo. Conseqüentemente, um dos principais desafios do planejamento energético consiste em realizar previsões acuradas para as demandas futuras. Neste trabalho, apresentamos um modelo de previsão para o consumo de energia elétrica a curto prazo. A metodologia utilizada compreende a construção de um comitê de previsão, por meio da aplicação do algoritmo *Gradient Boosting* em combinação com modelos de árvores de decisão e a técnica de amortecimento exponencial. Esta estratégia compreende um método de aprendizado supervisionado que ajusta o modelo de previsão com base em dados históricos do consumo de energia, das temperaturas registradas e de variáveis de calendário. Os modelos propostos foram testados em duas bases de dados distintas e demonstraram um ótimo desempenho quando comparados com resultados publicados em outros trabalhos recentes.

**Palavras-chave:** Previsão de Carga Elétrica. Amortecimento Exponencial. Árvores de decisão. Gradient Boosting.

## ABSTRACT

The storage of electrical energy is still not feasible on a large scale due to technical and economic issues. Therefore, all energy to be consumed must be produced instantly; it is not possible to store the production leftover, or either to cover any supply shortages with safety stocks, even for a short period of time. Thus, one of the main challenges of energy planning consists in computing accurate forecasts for the future demand. In this paper, we present a model for short-term load forecasting. The methodology consists in composing a prediction committee by applying the Gradient Boosting algorithm in combination with decision tree models and the exponential smoothing technique. This strategy comprises a supervised learning method that adjusts the forecasting model based on historical energy consumption data, the recorded temperatures and calendar variables. The proposed models were tested in two different datasets and showed a good performance when compared with results published in recent papers.

**Keywords:** Short Term Load Forecasting. Exponential Smoothing. Decision Trees. Gradient Boosting.

## SUMÁRIO

1	Introdução .....	14
1.1	Justificativa .....	16
1.2	Classificação dos problemas de previsão de carga elétrica .....	17
1.3	Objetivos .....	19
1.4	Organização do trabalho .....	20
2	Revisão de Literatura .....	21
3	Materiais e Métodos .....	27
3.1	Caracterização dos dados .....	27
3.1.1	<i>Rio-96/97</i> .....	27
3.1.2	<i>GEFCom2012</i> .....	28
3.1.3	<i>Análise exploratória de dados</i> .....	30
3.2	Métodos de Amortecimento Exponencial .....	31
3.2.1	<i>Amortecimento exponencial simples</i> .....	31
3.2.2	<i>Método de Holt-Winters</i> .....	34
3.2.3	<i>Amortecimento exponencial com dupla sazonalidade</i> .....	36
3.3	Árvores de Decisão .....	38
3.3.1	<i>Introdução</i> .....	38
3.3.2	<i>Algoritmo de construção</i> .....	40
3.3.3	<i>Características das árvores de decisão</i> .....	45
3.3.3.1	<i>Interpretabilidade</i> .....	45
3.3.3.2	<i>Manipulação de dados numéricos e categóricos</i> .....	45
3.3.3.3	<i>Manipulação de informações incompletas</i> .....	46
3.3.3.4	<i>Insensibilidade à entradas irrelevantes</i> .....	46
3.3.3.5	<i>Pré-processamento de dados</i> .....	47
3.3.3.6	<i>Poder preditivo</i> .....	47
3.3.3.7	<i>Quadro comparativo</i> .....	48
3.3.4	<i>Algoritmos mais complexos</i> .....	48
3.4	<i>Gradient Boosting</i> .....	49

3.4.1	<i>Bagging vs. Boosting</i> .....	50
3.4.2	<i>Algoritmo Gradient Boosting</i> .....	52
3.4.3	<i>Parâmetros e estratégias de controle</i> .....	54
3.5	Metodologia .....	56
3.5.1	<i>Rio-96/97</i> .....	56
3.5.1.1	<i>Variáveis preditoras</i> .....	59
3.5.2	<i>GEFCom2012</i> .....	60
3.5.2.1	<i>Análise de dados</i> .....	61
3.5.2.2	<i>Previsão das temperaturas</i> .....	63
3.5.2.3	<i>Combinação de temperaturas</i> .....	64
3.5.2.4	<i>Previsão das séries de cargas</i> .....	66
3.5.2.5	<i>Adequação às restrições do problema</i> .....	67
4	Resultados.....	70
4.1	<i>Rio-96/97</i> .....	70
4.2	<i>GEFCom2012</i> .....	74
5	Conclusão.....	77
	Referências .....	80
	APÊNDICES .....	85
A	Análise Exploratória de Dados .....	85
A.1	<i>Rio-96/97</i> .....	85
A.2	<i>GEFCom2012</i> .....	88

## LISTA DE ILUSTRAÇÕES

2.1	Evolução do número de publicações na área de previsão de carga elétrica (Hong e Fan, 2016). . . . .	22
3.1	Série de carga <i>Rio-9697</i> no primeiro mês de dados . . . . .	28
3.2	Série de cargas e temperaturas amortecidas - <i>GEFCom2012</i> (Abril/2008). . .	30
3.3	Relação <i>peso x defasamento</i> para diferentes valores de $\alpha$ . . . . .	34
3.5	Exemplo de árvore de decisão em um problema de regressão. As variáveis $x_1$ e $x_2$ representam os atributos do espaço de entradas; $y_i$ é o valor da variável de resposta correspondente à $i$ -ésima observação da amostra; e $c_j$ é uma constante de aproximação para os valores $y_i$ que pertencem à região $R_j$ . .	39
3.9	Exemplo de árvore de decisão em problemas com variáveis irrelevantes. . . . .	47
3.10	Estratégias para construção de comitês de previsores: <i>Bagging</i> x <i>Boosting</i> (os percentuais são valores parametrizáveis que indicam a proporção da subamostragem de dados em cada subconjunto da amostra total de treinamento). . . . .	51
3.12	Salto no nível médio da série de cargas da zona 10. . . . .	62
3.13	Comportamentos atípicos na série de carga da zona 09. . . . .	63
3.14	Diagrama de dispersão ‘temperatura x carga’ e ajuste polinomial de 3ª ordem	65
A.1	Consumo de energia elétrica ao longo do ano ( <i>Rio-96/97</i> ). . . . .	85
A.2	<i>Boxplot</i> do consumo de energia por dia da semana ( <i>Rio-96/97</i> ). . . . .	86
A.3	<i>Boxplot</i> do consumo de energia por horário do dia ( <i>Rio-96/97</i> , seg-sex). . . . .	87
A.4	Diagrama de dispersão ‘carga x temperatura’ por faixa de horário ( <i>Rio-96/97</i> , seg-sex). . . . .	87
A.5	Participação de cada zona geográfica no consumo total do sistema durante todo o período observado ( <i>GEFCom2012</i> ). . . . .	88
A.7	Consumo de energia elétrica ao longo do ano ( <i>GEFCom2012</i> - carga total do sistema). . . . .	90
A.8	<i>Boxplot</i> do consumo de energia por horário do dia ( <i>GEFCom2012</i> - carga total do sistema, seg-sex). . . . .	90

A.9 Diagrama de dispersão ‘ <i>carga x temperatura</i> ’ por faixa de horário ( <i>GEFCom2012</i> - carga total do sistema, seg-sex). . . . .	91
--	----

## LISTA DE TABELAS

1.1	Classificação dos problemas de previsão de carga elétrica segundo o horizonte de planejamento (Hong, 2010). . . . .	18
3.1	Resumo das características de cada base de dados. . . . .	30
3.2	Notação adotada para representar a amostra de dados. . . . .	40
3.3	Características qualitativas de diferentes métodos de aprendizado. Legenda: ▲ = bom; ◆ = regular; ▼ = ruim. (Hastie <i>et al.</i> , 2009) . . . . .	48
3.4	Parâmetros ótimos determinados para o método de amortecimento exponencial com dupla sazonalidade ( <i>Rio-96/97</i> ) . . . . .	57
3.5	Parametrização do algoritmo <i>Gradient Boosting</i> ( <i>Rio-96/97</i> ) . . . . .	59
3.6	Lista de variáveis preditoras candidatas ( <i>Rio-96/97</i> ) . . . . .	60
3.7	Lista de variáveis preditoras candidatas ( <i>GEFCom2012</i> ) . . . . .	66
3.8	Parametrização do algoritmo <i>Gradient Boosting</i> ( <i>GEFCom2012</i> ) . . . . .	67
4.1	Erros MAPE na amostra de teste ( <i>Rio-96/97</i> ) . . . . .	71
4.2	Distribuição dos pesos de previsão - <i>GEFCom2012</i> . . . . .	74
4.3	Resultados da competição <i>GEFCom2012</i> em toda amostra de teste (Hong <i>et al.</i> , 2014) . . . . .	75

# 1 Introdução

A eletricidade é a principal forma de energia distribuída para fins residenciais e comerciais, além de ser amplamente utilizada na indústria. A demanda por energia elétrica reflete o ritmo das atividades econômicas, o acesso da população à rede elétrica e a sua capacidade de adquirir e usufruir de bens de consumo que demandam eletricidade. Desta maneira, o consumo de energia elétrica compreende um importante indicador de desenvolvimento socioeconômico de uma nação. Particularmente no Brasil, o consumo per capita de energia tem demonstrado uma tendência de crescimento bastante evidente, a uma taxa média de 45 kWh/ano nas últimas quatro décadas, segundo dados do Banco Mundial (Figura 1.1).

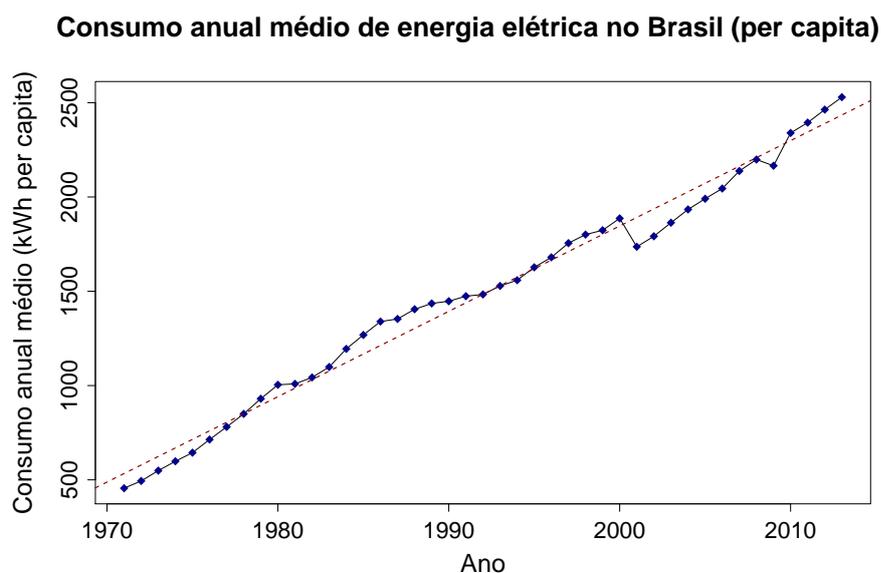


Figura 1.1: Consumo anual médio de energia elétrica no Brasil. Fonte: Banco Mundial<sup>1</sup>.

Para suportar o aumento contínuo na demanda de eletricidade é necessário um planejamento energético antecipado. Os investimentos não devem se restringir à ampliação dos sistemas de geração, transmissão e distribuição de energia. Também é necessário concentrar esforços na otimização dos recursos disponíveis. Eficiência energética tem sido um dos principais temas de pesquisa da área de energia nos últimos anos.

Diferentemente das *commodities* de energia, como óleo e gás, a eletricidade não pode ser armazenada em larga escala de forma economicamente viável, considerando os recursos

<sup>1</sup>disponível em: <http://data.worldbank.org/indicador/EG.USE.ELEC.KH.PC?locations=BR>

tecnológicos existentes. Esta impossibilidade de armazenamento exige um equilíbrio contínuo entre a produção e o consumo de eletricidade. Toda energia consumida deve ser produzida instantaneamente e, quando há desequilíbrios, mesmo que por frações de minuto, todo o sistema corre o risco de desligamentos em cascata, os chamados “apagões” (ABRADEE, 2015).

Neste contexto, a produção de energia elétrica exige um planejamento eficiente que, por sua vez, requer previsões acuradas para as demandas futuras. A longo prazo, é importante garantir que o sistema elétrico será capaz de suprir o crescimento da demanda, ao passo que, as previsões de curto prazo são importantes para orientar o planejamento operacional e garantir o equilíbrio contínuo entre a produção e o consumo.

Do ponto de vista das empresas, a geração excedente de energia é desperdiçada e, portanto, se traduz em prejuízo financeiro. Por outro lado, em mercados competitivos, a produção insuficiente de energia elétrica obriga as empresas a recorrerem ao mercado de energia à vista (*spot market*), onde o valor de compra é, geralmente, muito superior ao custo de produção.

Ainda no âmbito operacional, o desequilíbrio entre a oferta e a demanda de eletricidade pode provocar anormalidades na rede elétrica, comprometendo a qualidade da energia transmitida. Esta perda de qualidade compreende alterações na frequência e tensão da rede e pode causar danos aos equipamentos elétricos dos consumidores finais. Em situações extremas o fornecimento de energia deve ser interrompido.

No cenário nacional, por exemplo, a qualidade da energia elétrica é monitorada pela ANEEL (Agência Nacional de Energia Elétrica), órgão regulador que têm autonomia para aplicar sanções pecuniárias às empresas que não atendem às exigências mínimas de qualidade no fornecimento de energia elétrica. Assim, valoriza-se o compromisso com a qualidade do serviço prestado e, conseqüentemente, a necessidade de um planejamento energético eficiente.

O escopo deste trabalho compreende o problema de previsão de carga elétrica em aplicações de caráter operacional, ou seja, a curto prazo. O objetivo consiste em construir um modelo estatístico para prever demandas futuras para o consumo de energia com base em dados passados e variáveis exógenas. O método proposto combina um poderoso algoritmo de aprendizado de máquina, chamado *Gradient Boosting*, com árvores de decisão e um modelo de previsão baseado na técnica amortecimento exponencial.

## 1.1 Justificativa

Existem diversos riscos associados aos erros de previsão de carga. Hobbs *et al.* (1999) citam as principais consequências relacionadas aos erros de previsão, seja no caso de estimações subdimensionadas ou superdimensionadas.

No caso de previsões superdimensionadas, destacam-se:

- Unidades geradoras podem ser acionadas desnecessariamente, elevando os custos com combustíveis e manutenção;
- Energia extra pode ser comprada indevidamente a um alto custo, ou pode-se perder uma oportunidade lucrativa de vender energia;
- Potencial hidroelétrico pode ser turbinado sem necessidade, quando poderia ser armazenado para atender demandas futuras.
- Preços de energia podem ser cotados por um valor muito elevado, prejudicando as vendas;
- Interrupções no abastecimento e medidas de controle de carga podem ser implementadas indevidamente, frustrando consumidores e reduzindo as vendas.

No caso de previsões subdimensionadas, os principais riscos são:

- Recursos insuficientes podem ser disponibilizados para atender requisitos de segurança, colocando a confiabilidade do sistema em risco;
- Pode ser necessário comprar energia a preços acima do mercado para atender a demanda que não foi prevista;
- Interrupções no fornecimento e medidas de controle de carga podem ser necessárias, quando poderiam ser evitadas se houvesse um planejamento mais eficaz;
- Ofertas para venda de energia podem ser cotadas a um preço abaixo do mercado, ou até mesmo inferior ao custo de produção da energia para a própria empresa, diminuindo a lucratividade da operação.

Em geral, observa-se que os custos dos erros são bastante elevados e podem provocar consequências graves, tanto no âmbito operacional como financeiro, o que destaca a necessidade de se utilizar modelos elaborados para previsão de carga elétrica, com o objetivo de produzir estimativas com o máximo de acurácia possível.

No entanto, as séries de carga elétrica apresentam um comportamento bastante complexo, que inclui a presença de três comportamentos periódicos sobrepostos: *i*) um perfil diário de demanda de energia, que reflete o padrão de consumo da população e, em geral, apresenta picos no início do período noturno e cargas mais baixas durante a madrugada; *ii*) um comportamento periódico semanal, resultante da distinção entre os hábitos de consumo da população nos dias úteis e finais de semana; e *iii*) uma sazonalidade anual como consequência das temperaturas médias registradas em cada estação do ano, provocando um forte impacto no consumo de energia, uma vez que equipamentos elétricos são amplamente utilizados para climatização de ambientes (refrigeração e aquecimento).

Além dos três padrões sazonais, as séries de carga elétrica também estão relacionadas com variáveis exógenas, como as condições do tempo (temperatura, umidade, luminosidade e etc) e fatores econômicos. Estas séries também apresentam comportamentos atípicos durante feriados, pontos facultativos e eventos especiais (finais de campeonatos esportivos, eleições e etc.).

Os altos riscos associados aos erros de previsão de carga elétrica e o comportamento complexo das séries de consumo de energia justificam o interesse crescente em pesquisas envolvendo o problema de previsão de carga, conforme discutiremos na revisão de literatura apresentada no próximo capítulo.

## **1.2 Classificação dos problemas de previsão de carga elétrica**

Em situações práticas, os problemas de previsão de carga elétrica podem ter restrições e objetivos muito distintos, o que requer a construção de modelos de previsão adequados para cada tipo de situação. Além disso, as características gerais das séries de carga elétrica podem variar bastante entre regiões e perfis de consumidores. O número e a proporção entre o consumo de clientes residenciais, comerciais e industriais; além de fatores geográficos, climáticos, econômicos, sociais, culturais, entre outros, são características

determinantes para o comportamento das séries de consumo de energia. Feinberg e Genethliou (2005) argumentam que não existe um modelo de previsão universal, que seja adequado em todas as situações.

Os problemas de previsão de carga podem ser classificados de acordo com vários critérios, entre eles: (i) *Horizonte de previsão*, (ii) *Ciclo de atualização*, (iii) *Resolução temporal*, e (iv) *tipo de previsão (pontual ou probabilística)*.

- ***Horizonte de previsão***: compreende o prazo de antecedência em que as previsões são realizadas. De acordo com Bunn e Farmer (1985), o horizonte de previsão para séries de carga elétrica pode variar desde alguns minutos à frente, até 40 anos ou mais. Hong (2010) estabelece dois tipos de classificação, conforme a Tabela 1.1. No mercado de energia, o planejamento de longo-prazo compreende decisões estratégicas como a construção de novas usinas geradoras e grandes linhas de transmissão. O planejamento de médio prazo envolve decisões como a compra de energia, insumos e a manutenção de unidades geradoras. A curto prazo são tratadas as questões operacionais relativas ao funcionamento dos sistemas de geração, transmissão ou distribuição de eletricidade.

Tabela 1.1: Classificação dos problemas de previsão de carga elétrica segundo o horizonte de planejamento (Hong, 2010).

Classificação <sup>1</sup>	Classificação <sup>2</sup>	Caráter	Ciclo de Atualização	Horizonte
STLF	VSTLF	Operacional	≤ horário	Até um dia
	STLF	Operacional	diário	Até duas semanas
LTLF	MTLF	Tático	mensal	Até três anos
	LTLF	Estratégico	anual	30 anos ou mais

- ***Ciclo de atualização***: corresponde à frequência com que as previsões são recalculadas à medida em que novos dados são observados. Assim, previsões

<sup>1</sup>Previsão de carga a curto prazo - STLF (*Short Term Load Forecasting*) e previsão de carga a longo prazo - LTLF (*Long Term Load Forecasting*)

<sup>2</sup>Previsão de carga a muito curto prazo - VSTLF (*Very Short Term Load Forecasting*); previsão de carga a curto prazo - STLF (*Short Term Load Forecasting*); previsão de carga a médio prazo - MTLF (*Medium Term Load Forecasting*) e previsão de carga a longo prazo - LTLF (*Long Term Load Forecasting*)

estratégicas para um horizonte de 10 anos podem ser recalculadas e atualizadas anualmente. O ciclo de atualização está diretamente relacionado à aplicação das previsões. A Tabela 1.1, extraída de Hong (2010), sugere valores usuais para cada tipo de aplicação.

- **Resolução temporal:** refere-se à frequência de amostragem dos dados, os quais podem ser medidos a cada minuto, hora, dia, mês, ano e etc.
- **Previsão pontual ou probabilística:** a resposta de um modelo de previsão pode ser pontual ou probabilística. Modelos de previsão pontual fornecem um valor numérico como estimativa para as observações futuras. No entanto, é possível construir modelos para fornecer previsões na forma de distribuição de probabilidades, conforme discute Gneiting e Katzfuss (2014).

### 1.3 Objetivos

Os principais objetivos deste trabalho são:

- Revisar as principais publicações sobre o problema de previsão de carga elétrica a curto prazo e pesquisar as tendências atuais e o estado da arte nesta linha de pesquisa.
- Estudar as características gerais das séries de consumo de energia, bem como as particularidades associadas à cada região.
- Propor uma estratégia para combinar séries de temperaturas mensuradas em diferentes pontos da região geográfica das zonas de consumo.
- Combinar o algoritmo *gradient boosting* com modelos de árvores de decisão e uma técnica de amortecimento exponencial para séries com dois padrões sazonais, para produzir um modelo de previsão mais acurado.
- Apresentar uma estratégia do tipo *top-bottom* para fazer a agregação das previsões no caso do problema de previsão hierárquica de carga elétrica e comparar os resultados com a estratégia convencional *bottom-top*.
- Testar e validar as metodologias propostas em bases de dados reais, comparando o desempenho com resultados publicados por outros pesquisadores.

## 1.4 Organização do trabalho

O texto está organizado em cinco capítulos. O próximo capítulo apresenta uma revisão de literatura sobre o tema, incluindo os trabalhos mais recentes e o estado da arte nesta linha de pesquisa. O capítulo 3, materiais e métodos, descreve as características das bases de dados utilizadas e apresenta os métodos utilizados para construir o modelo de previsão proposto neste trabalho. Em seguida, no capítulo 4, os resultados obtidos foram analisados e comparados com outros trabalhos que utilizaram as mesmas bases de dados. Finalmente, o capítulo 5 discute as considerações finais e as conclusões deste trabalho.

## 2 Revisão de Literatura

Previsões para o consumo de energia sempre foram muito importantes para orientar o planejamento e as tomadas de decisões nas companhias de energia elétrica, contudo, o tema alcançou maior relevância nas últimas décadas, devido ao aumento da competitividade no mercado de energia elétrica (Feinberg e Genethliou, 2005). Este novo cenário foi provocado pela adoção de novas políticas energéticas que ocorreram em vários países, inclusive no Brasil, com o objetivo de reduzir a intervenção pública no mercado de energia.

O problema de previsão de carga, apesar de sua notoriedade em publicações recentes, não é um tema novo. Uma revisão bibliográfica publicada em 1980 (Huck *et al.*, 1980), já contabilizava mais de 70 trabalhos na área. Bunn e Farmer (1985) já haviam discutido as implicações econômicas e operacionais das previsões de carga elétrica a curto prazo, muito antes da desregulamentação do mercado de energia na maioria dos países.

Outros fatores também contribuíram para a difusão dos trabalhos sobre previsão de carga, entre eles o desenvolvimento dos computadores digitais, que permitiu a manipulação de grandes volumes de dados e a utilização de métodos de previsão computacionalmente intensivos. O advento das tecnologias *Smart Grid* também colaborou para as pesquisas nesta área, uma vez que viabilizou a aquisição de dados de consumo de energia com alta confiabilidade e maiores resoluções temporais e espaciais.

Com o amadurecimento das pesquisas, novas necessidades e desafios surgiram na área de previsão de carga, entre eles: a previsão de demanda em *microgrids* (Amjady *et al.* (2010); Javed *et al.* (2012); Chitsaz *et al.* (2015)), que se caracteriza por zonas de consumo menores e, portanto, com curvas de carga menos suaves e mais susceptíveis à efeitos de natureza estocástica; o problema de previsão hierárquica (Han *et al.* (2012); Hong *et al.* (2014)), cujo o objetivo é fornecer previsões para a demanda de energia em diferentes níveis de abastecimento, como bairros, cidades, estados e etc.; e o problema de previsão probabilística (Gneiting *et al.* (2007); Hong e Fan (2016)), em que se deve fornecer previsões na forma de distribuição de probabilidades.

Hong e Fan (2016) apresentam a evolução do número de publicações relevantes na área de previsão de carga ao longo das últimas cinco décadas (Figura 2.1). As publicações foram

estratificadas classificando-se os modelos de previsão em quatro categorias: longo prazo (LTLF), curto prazo (STLF), modelos de previsão hierárquicos (S/HLF) e probabilísticos (PLF). Embora o estudo tenha concentrado-se na evolução dos trabalhos sobre modelos probabilísticos, é possível notar que, em geral, o número total de publicações cresceu muito nos últimos anos.

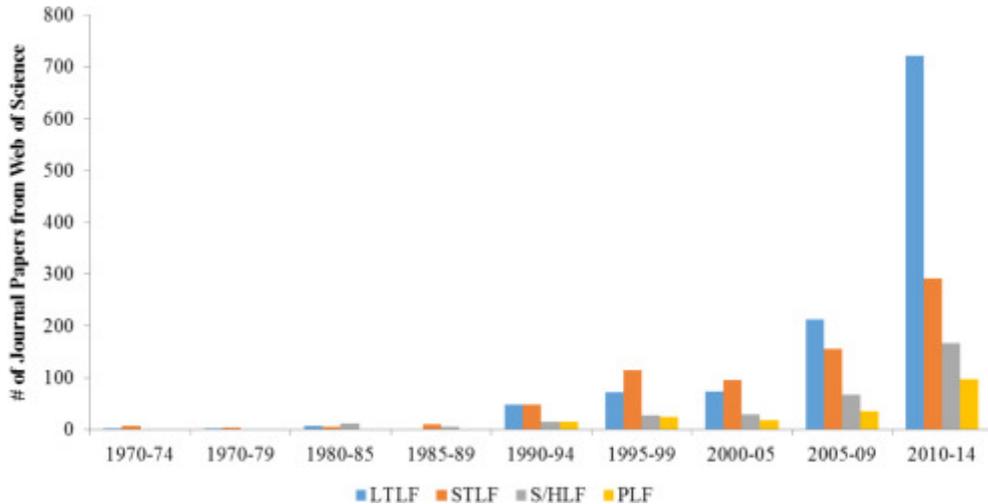


Figura 2.1: Evolução do número de publicações na área de previsão de carga elétrica (Hong e Fan, 2016).

Atualmente, a literatura envolvendo o problema de previsão de carga elétrica é bastante extensa. De acordo com Hong Hong *et al.* (2014), já existe um vasto conteúdo dedicado à previsão de consumo de energia, entretanto, a maior parte dos trabalhos ainda está em nível teórico e apresenta pouca contribuição prática. Segundo os autores, ainda não existe um processo de *benchmarking* formal, ou sequer um conjunto de dados público que seja bastante difundido. As publicações raramente reproduzem resultados de trabalhos passados para efeitos de comparação.

A maior parte dos modelos de previsão utiliza técnicas estatísticas clássicas ou algoritmos de inteligência computacional, como redes neurais artificiais, lógica nebulosa e sistemas especialistas (Feinberg e Genethliou, 2005). Basicamente, os métodos de previsão de séries temporais podem ser classificados como univariados ou multivariados (Wei, 2006). Nos modelos univariados, as previsões são calculadas com base apenas em observações passadas da própria série temporal, enquanto os modelos multivariados permitem incluir variáveis exógenas que podem influenciar indiretamente o comportamento da série de interesse.

Taylor *et al.* (2006) realizaram um estudo comparativo entre modelos univariados para previsão de carga elétrica. Os autores testaram modelos auto-regressivos e de amortecimento exponencial adaptados para séries com dois padrões sazonais. Métodos mais complexos, como redes neurais e regressão linear com análise de componentes principais (PCA) também foram avaliados, porém, os melhores resultados foram obtidos com a técnica de amortecimento exponencial com dupla sazonalidade proposta em seu trabalho anterior (Taylor, 2003).

Carneiro (2014) combinou diversos modelos univariados para previsão de carga elétrica. Os previsores foram combinados utilizando uma média ponderada das respostas de cada modelo, em que os coeficientes de ponderação foram otimizados com base em uma amostra de dados de ajuste. Entre os modelos testados, o destaque também foi para o método de amortecimento exponencial com dupla sazonalidade, que obteve o melhor resultado individualmente, e recebeu pesos maiores no processo de otimização dos coeficientes de ponderação.

As condições do tempo, em especial a temperatura, podem exercer uma forte influência no padrão de consumo de energia elétrica. Assim, modelos multivariados são muito frequentes na literatura. Nesta linha, destacam-se os métodos de regressão linear múltipla (Hong *et al.* (2011); Charlton e Singleton (2014)) e diversas metodologias baseadas em técnicas de inteligência computacional, entre elas, redes neurais artificiais (Hippert *et al.*, 2001), lógica *fuzzy* (Ranaweera *et al.*, 1996) e *Boosting* (Ben Taieb e Hyndman (2014), Lloyd (2014)).

As redes neurais compreendem uma das técnicas de aprendizado mais utilizadas em problemas de previsão de carga elétrica a curto prazo. Hippert *et al.* (2001) revisaram cerca de 40 trabalhos publicados em revistas de grande influência, com o objetivo de investigar a descrença na utilização de redes neurais em problemas de previsão de carga. As duas principais deficiências encontradas foram: *i*) muitos trabalhos avaliados apresentaram modelos superajustados aos dados de treinamento, seja por excesso de parâmetros ou por excesso de treinamento; *ii*) em geral, os resultados apresentados não foram testados e avaliados de forma plenamente convincente do ponto de vista científico.

Ranaweera *et al.* (1996) apresentam um sistema *fuzzy* para prever a carga elétrica total e de pico um dia à frente. Apesar de destacarem a importância do papel do especialista na construção de modelos *fuzzy*, os autores utilizaram uma metodologia capaz de criar o

conjunto de regras com base apenas na amostra de dados de treinamento. Além disso, os resultados foram aprimorados com a implementação de um processo de atualização dinâmica da base de regras, à medida que novas observações da amostra de teste eram apresentadas ao sistema. Os modelos propostos superaram o desempenho de uma rede neural, considerando as mesmas variáveis de entrada.

Hong *et al.* (2014) organizaram uma competição denominada *Global Energy Forecasting Competition 2012 (GEFCom2012)* para previsão de carga elétrica a curto prazo. A competição abordou o problema de previsão hierárquica de carga e, posteriormente, a base de dados utilizada foi disponibilizada para trabalhos futuros.

Ao todo, mais de 100 equipes submeteram seus trabalhos para a competição *GEFCom2012*. Alguns autores reportaram resultados bastante promissores com a utilização de técnicas de *boosting*. Ben Taieb e Hyndman (2014) apresentaram uma lista com várias variáveis preditoras e utilizaram uma técnica denominada *component wise gradient boosting*, utilizando *splines* penalizadas. Este método facilita a manipulação de dados com alta dimensionalidade, uma vez que implementa um procedimento iterativo com seleção automática de atributos. Para a mesma competição, Lloyd (2014) implementou um comitê de previsão, cuja resposta foi definida por uma média ponderada entre três modelos: *Gradient Boosting* com árvores de decisão, processos gaussianos e regressão linear. Os dois trabalhos, ficaram entre os cinco mais bem colocados na competição.

Além dos modelos de previsão apresentados por Ben Taieb e Hyndman (2014) e Lloyd (2014), outros dois trabalhos demonstraram resultados relevantes. Nedellec *et al.* (2014) utilizaram uma abordagem multi-escala, decompondo o comportamento da série em três componentes: *i)* longo prazo, *ii)* médio prazo, e *iii)* curto prazo. Para cada componente, os autores adotaram uma abordagem diferente, a longo prazo, o comportamento da série foi modelado através de um modelo de regressão via *kernel*; os efeitos de médio prazo foram previstos com modelos aditivos generalizados; e o algoritmo *random forest* foi utilizado para modelar os padrões de curto prazo.

Charlton e Singleton (2014) aplicaram uma série refinamentos no modelo de regressão linear proposto por Hong (2010), que foi utilizado como referência (*benchmark*) da competição. Entre os aperfeiçoamentos realizados, destacaram-se: um método mais suavizado para previsão de temperaturas, uma estratégia para combinar previsões de

carga obtidas com diferentes séries de temperaturas, e a segmentação dos dados por horário do dia, estação do ano e tipo de dia (final de semana ou dia útil). Para cada divisão de dados os autores treinaram um modelo de regressão linear múltipla.

Um dos principais desafios da competição *GEFCom2012* foi selecionar quais séries de temperatura deveriam ser utilizadas para prever a demanda de energia em cada uma das zonas de consumo. Ao todo, haviam 20 zonas de consumo e 11 estações meteorológicas distribuídas na mesma região geográfica. Os organizadores da competição, propositalmente, não forneceram nenhuma informação sobre a posição geográfica das zonas de consumo e das estações meteorológicas. Após revisar as metodologias utilizadas pelas equipes que participaram da competição, Hong *et al.* (2015) apresentaram uma proposta para seleção de estações meteorológicas em problemas de previsão de cargas.

A *GEFCom2012* contou com a participação de pesquisadores do meio acadêmico, da indústria de energia elétrica, além de profissionais e entusiastas da comunidade de aprendizado de máquina. Devido à sua ampla aceitação, dois anos após a *GEFCom2012*, os organizadores decidiram lançar uma segunda edição da competição. Desta vez, a chamada *GEFCom2014*, tinha como objetivo abordar o problema de previsão probabilística de carga elétrica. Em outras palavras, os trabalhos submetidos deveriam prover a distribuição de probabilidades para as demandas futuras de energia, discretizada em 99 percentis, e não apenas um valor pontual como na competição anterior.

A necessidade de modelos probabilísticos para previsão de carga elétrica não é uma novidade. Bunn e Farmer (1985) já haviam destacado a importância de conhecer a distribuição estatística dos erros de previsão ao longo dos dias ou períodos do ano. Entretanto, este tipo de abordagem tem conquistado mais atenção em publicações mais recentes, como destaca Hong e Fan (2016) ao revisar os trabalhos sobre modelos probabilísticos para previsão de carga elétrica. Apesar de ser um problema antigo, porém pouco estudado, os modelos probabilísticos ainda enfrentam algumas questões em aberto como, por exemplo, a definição de medidas de erro e critérios para avaliação e comparação de desempenho entre diferentes modelos de previsão.

Previsões probabilísticas estão fora do escopo deste trabalho, porém, destacamos a seção especial publicada no 32º volume do periódico *International Journal of Forecasting*, que publicou diversos trabalhos com metodologias bem sucedidas na competição *GEFCom2014*.

De modo geral, as técnicas de inteligência computacional têm dificuldade em modelar dados auto-correlacionados, o que é uma característica inerente das séries temporais. Por outro lado, os métodos específicos para previsão de séries temporais não conseguem capturar satisfatoriamente as relações entre as variáveis externas e a série de interesse, em especial relações não-lineares. O objetivo deste trabalho é combinar a habilidade específica dos métodos de previsão de séries temporais com a capacidade de aprendizado das técnicas de inteligência computacional. A metodologia proposta consiste em obter previsões iniciais usando o método de amortecimento exponencial com dupla sazonalidade (Taylor, 2003) e, em seguida, modelar os resíduos de previsão através do algoritmo *Gradient Boosting* (Friedman, 2001).

## 3 Materiais e Métodos

Este capítulo descreve as características das bases de dados utilizadas neste trabalho e apresenta: *i)* a técnica de amortecimento exponencial, *ii)* os modelos de árvores de decisão, e *iii)* o algoritmo *Gradient Boosting*; que compõem o conjunto de métodos utilizados para construir o modelo de previsão de carga proposto.

### 3.1 Caracterização dos dados

Este trabalho utiliza duas bases de dados distintas, doravante denominadas de *Rio-96/97* e *GEFCom2012*. Ambas as bases de dados se referem ao problema de previsão de consumo de energia elétrica a curto prazo com resolução temporal horária. Entretanto, o comportamento das séries, suas características gerais e as regras de previsão são muito diferentes para os dois problemas.

As duas bases de dados já foram bastante utilizadas na literatura, o que motivou sua adoção neste trabalho. A base de dados *Rio-96/97*, é a mesma utilizada por Taylor *et al.* (2006), Neto (2014), e Carneiro (2014). Já os dados da base *GEFCom2012* foram utilizados em uma competição de previsão de carga elétrica promovida pelo *IEEE Working Group on Energy Forecasting* (Hong *et al.*, 2014), conforme discutimos na revisão de literatura apresentada no capítulo anterior.

#### 3.1.1 *Rio-96/97*

A base de dados *Rio-96/97* contém a série de consumo de energia elétrica durante o período de 01/01/1996 a 28/12/1997 em uma área do estado do Rio de Janeiro. A série de temperaturas da mesma região está disponível para o período de 01/04/1996 a 28/12/1997. Os dados apresentam resolução horária e foram fornecidos pela concessionária de energia que atende a região.

A Figura 3.1 ilustra os valores observados para o consumo de energia no primeiro mês de dados. É fácil notar que a série possui um padrão sazonal diário e outro semanal, o que é um comportamento típico das séries de carga elétrica. Na prática, o consumo de energia também apresenta um comportamento sazonal anual, mas que pode ser desprezado em

aplicações de previsão a curto prazo.

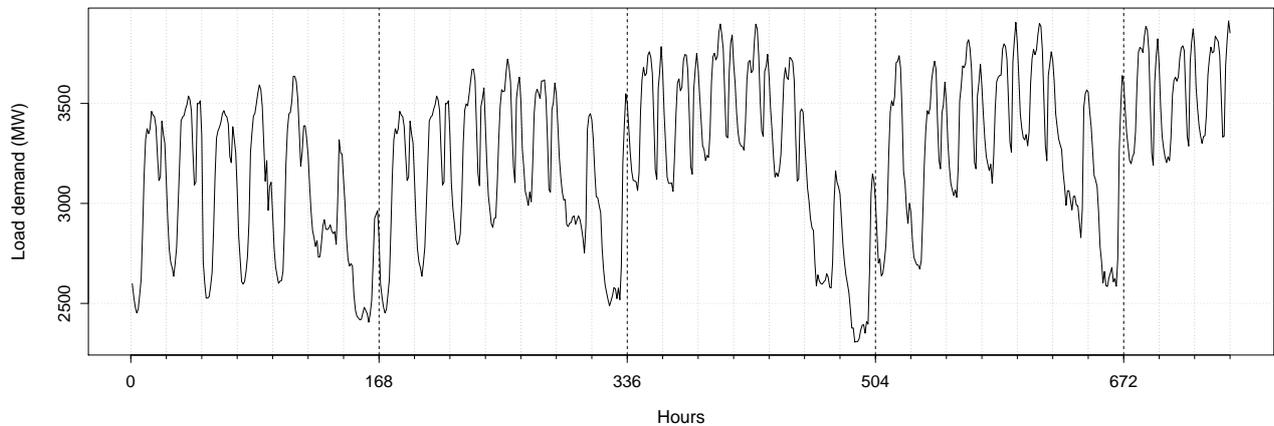


Figura 3.1: Série de carga *Rio-9697* no primeiro mês de dados

Nos dados *Rio-96/97*, o objetivo é prever o perfil diário de cargas. Em outras palavras, permite-se utilizar os dados até a última hora do dia  $d$  para prever a carga nas próximas 24 horas do dia  $d + 1$ .

Feriados e dias especiais possuem uma curva de demanda muito distinta dos dias comuns. Em geral, a previsão de demanda nessas datas é realizada por um especialista, que deve considerar uma série de fatores externos. Esta tarefa está fora do escopo de trabalho para esta base de dados. Sendo assim, as cargas elétricas nos feriados foram substituídas pelo valor médio observado no mesmo dia e horário das semanas anterior e posterior ao feriado.

### 3.1.2 *GEFCom2012*

A base de dados *GEFCom2012* contém as séries de carga elétrica em 20 zonas de consumo localizadas no estado da Carolina de Norte (EUA). Também estão disponíveis 11 séries de temperatura do ar mensuradas em diferentes estações meteorológicas, distribuídas na mesma região geográfica das zonas de consumo. As séries apresentam resolução temporal horária, e compreendem o período de 01/01/2004 a 07/07/2008.

A utilização de múltiplas zonas de consumo compreende o problema de previsão hierárquica de consumo de energia, cujo objetivo é fornecer previsões de carga elétrica para diferentes níveis hierárquicos da rede de distribuição. No caso da competição *GEFCom2012*, o objetivo é prever a carga individual de cada zona e carga total do

sistema (somatório da carga nas 20 zonas de consumo).

Para facilitar a comparação entre os resultados, seguiremos neste trabalho o mesmo regulamento utilizado na competição *GEFCom2012*:

- Horizonte de planejamento de uma semana à frente com resolução horária;
- O desempenho das soluções foi avaliado em 8 semanas intermediárias (06/03/2005 a 12/03/2005, 20/06/2005 a 26/06/2005, 10/09/2005 a 16/09/2005, 25/12/2005 a 31/12/2005, 13/02/2006 a 19/02/2006, 25/05/2006 a 31/05/2006, 02/08/2006 a 08/08/2006, 22/11/2006 a 28/11/2006) e na semana final (01/07/2008 a 07/07/2008);
- Para as semanas intermediárias foram disponibilizadas as séries reais de temperaturas observadas nas 11 estações meteorológicas, enquanto que para a semana final todas as séries de temperaturas foram omitidas;
- Para as semanas intermediárias permite-se a utilização de todos os dados disponíveis, isto é, o modelo pode ser ajustado com informações passadas e futuras em relação à semana a ser prevista (*backcasting*). Para a última semana, evidentemente, estão disponíveis apenas dados passados (*forecasting*);
- Não foi fornecida nenhuma informação a respeito da localização geográfica das zonas de consumo e estações meteorológicas. Portanto, fica a critério do projetista do modelo decidir quantas e quais séries de temperaturas serão utilizadas na construção e ajuste dos modelos de previsão;
- A solução deve conter as previsões para cada uma das 20 zonas de consumo, bem como a previsão para a carga total do sistema. A fim de manter a coerência, exige-se que as previsões para a carga total do sistema sejam iguais ao somatório das previsões para cada zona de consumo;
- Foi fornecida uma lista com os feriados locais. Algumas semanas do período de avaliação foram intencionalmente selecionadas de forma a contemplar feriados;
- Para participar da competição os modelos de previsão devem utilizar apenas os dados fornecidos pelos organizadores.

A figura 3.2 ilustra a série de cargas e temperaturas para a base de dados *GEFCom2012*. Ambas as séries foram suavizadas por um processo de médias móveis centradas com janela de tamanho 24, para neutralizar os efeitos de sazonalidade diária. Verifica-se que o nível médio da série oscila bastante quando a temperatura está instável e as cargas são mais altas em temperaturas muito baixas.

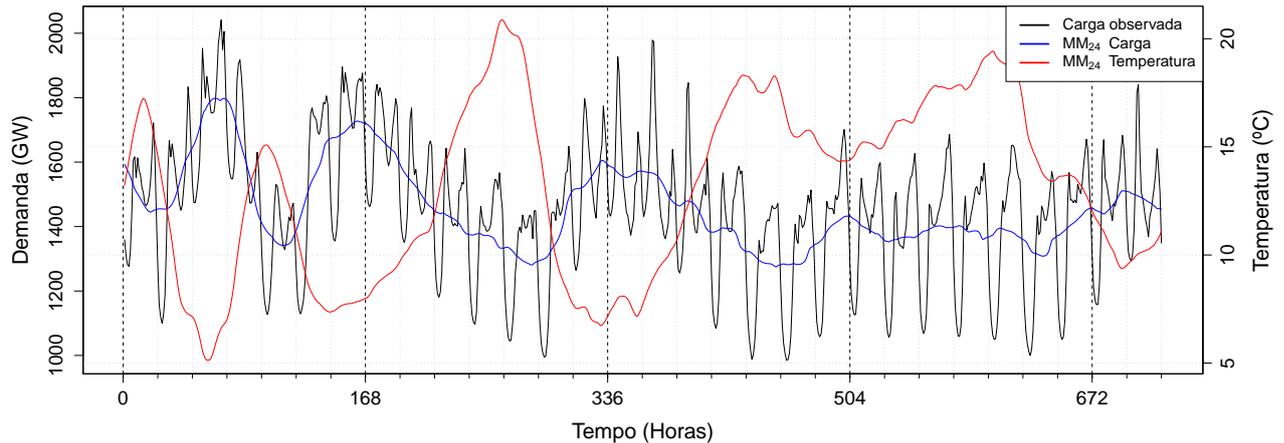


Figura 3.2: Série de cargas e temperaturas amortecidas - *GEFCom2012* (Abril/2008).

### 3.1.3 Análise exploratória de dados

O apêndice A apresenta uma análise estatística detalhada sobre as duas bases de dados descritas nesta seção. A Tabela 3.1 resume as principais características das bases *Rio-96/97* e *GEFCom2012*.

Tabela 3.1: Resumo das características de cada base de dados.

Característica	Rio-96/97	GEFCom2012
Região	Rio de Janeiro	Carolina do Norte (EUA)
Período	01/01/96 a 28/12/97	01/01/04 a 07/07/08
Carga média <sup>1</sup>	3.060 MW	1.647 MW <sup>2</sup>
Horizonte de planejamento	24 horas	168 horas
Resolução temporal	horária	horária
Ciclo de atualização	não há	não há
Tipo de previsão	pontual	pontual
Zonas de consumo	1	20
Séries de temperatura	1	11

## 3.2 Métodos de Amortecimento Exponencial

Os métodos de amortecimento exponencial foram propostos no final da década de 1950 (Brown, 1959; Holt, 1957; e Winters, 1960; *apud* Hyndman e Athanasopoulos (2014)), e têm motivado alguns dos modelos de previsão mais bem sucedidos. As previsões obtidas com as técnicas de amortecimento exponencial compreendem uma média ponderada das observações passadas, onde os pesos correspondentes a cada observação são reduzidos exponencialmente, à medida que os valores tornam-se mais antigos (Hyndman e Athanasopoulos, 2014).

As técnicas clássicas de amortecimento exponencial são amplamente utilizadas para fazer análises e previsões de séries temporais genéricas. Contudo, é possível aprimorar o desempenho dos métodos usuais introduzindo adequações específicas para cada tipo de aplicação. Taylor (2003) apresentou uma extensão para o método clássico de Holt-Winters com adaptações para o problema de previsão de carga elétrica.

Nesta seção introduzimos o conceito básico da técnica de amortecimento exponencial e revisamos o método clássico de Holt-Winters, que deu origem às adaptações propostas por Taylor para aplicações em problemas de previsão de carga elétrica.

### 3.2.1 Amortecimento exponencial simples

O método de amortecimento exponencial simples, como o próprio nome diz, é o exemplo mais elementar de aplicação do conceito de amortecimento exponencial. Este método é indicado para séries de dados que não apresentam tendência ou sazonalidade. Portanto, são séries que podem ser modeladas simplesmente como flutuações aleatórias em torno de um determinado nível médio, conforme a Equação 3.1.

$$Z_t = L_t + \epsilon_t \quad (3.1)$$

Onde  $Z_t$  é o valor observado da série no instante  $t$ ;  $L_t$  é o nível médio da série; e  $\epsilon_t$  é um erro aleatório independente e com média nula. Este é um caso bastante particular de série temporal, que ocorre quando os dados não demonstram tendência de crescimento ou decrescimento, nem variações sazonais periódicas. Na prática, isto não significa que o

---

<sup>1</sup>Referente à todo período de amostragem.

<sup>2</sup>Referente à carga total do sistema, ou seja, a soma das 20 zonas de consumo.

nível médio da série seja absolutamente constante; na verdade, o nível  $L_t$  pode apresentar variações suaves no decorrer do tempo, mas os dados da amostra não evidenciam que estas variações seguem algum tipo de comportamento regular.

Neste caso, como  $\epsilon_t$  é uma variável aleatória com média zero, a melhor previsão para um valor futuro da série que pode ser obtida no instante  $t$  (representada por  $\hat{Z}_{t+h|t}$ ) deve ser uma estimativa do nível da série no último instante disponível  $\hat{L}_t$ . Portanto temos:

$$\hat{Z}_{t+h|t} = \hat{L}_t \quad (3.2)$$

O nível médio da série pode ser estimado a partir da média das observações passadas, uma vez que a média dos erros  $\epsilon_t$  é nula.

$$\begin{aligned} \bar{Z}_t &= \frac{Z_t + Z_{t-1} + Z_{t-2} + \dots + Z_1}{t} \\ &= \frac{L_t + \epsilon_t + L_{t-1} + \epsilon_{t-1} + L_{t-2} + \epsilon_{t-2} + \dots}{t} \\ &= \frac{L_t + L_{t-1} + L_{t-2} + \dots}{t} + \frac{\epsilon_t + \epsilon_{t-1} + \epsilon_{t-2} + \dots}{t} \\ &= \text{média}(L_t, L_{t-1}, L_{t-2}, \dots) + \text{média}(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \dots) \\ &\approx \text{média}(L_t, L_{t-1}, L_{t-2}, \dots) \end{aligned} \quad (3.3)$$

Entretanto, o nível médio pode variar ao longo do tempo; logo, como desejamos obter uma estimativa do nível médio no instante corrente, é interessante atribuir um peso maior às observações mais recentes. Portanto, uma boa estimativa para o nível médio da série  $\hat{L}_t$  pode ser obtida através de uma média ponderada das observações passadas, conforme Equação 3.4.

$$\hat{Z}_{t+h|t} = \hat{L}_t = \frac{\alpha_0 Z_t + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \alpha_3 Z_{t-3} + \dots}{\sum \alpha} \quad (3.4)$$

Onde  $\alpha_0 \geq \alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \dots$ , para garantir que os pesos das observações mais recentes sejam sempre maiores que o das observações mais antigas. Na prática, contudo, é impossível definir um número crescente de pesos, portanto, a solução é utilizar um procedimento recursivo que requer apenas um parâmetro  $\alpha$ , conforme Equação 3.5.

$$\hat{Z}_{t+h|t} = \hat{L}_t = \alpha Z_t + (1 - \alpha) \hat{L}_{t-1} \quad (3.5)$$

A Equação 3.5 resume o método de amortecimento exponencial simples, onde a *constante de amortecimento*  $0 < \alpha < 1$  é o único parâmetro a ser definido.

Expandindo a Equação 3.5 temos:

$$\begin{aligned}
 \hat{L}_t &= \alpha Z_t + (1 - \alpha) \hat{L}_{t-1} \\
 &= \alpha Z_t + (1 - \alpha) [\alpha Z_{t-1} + (1 - \alpha) \hat{L}_{t-2}] \\
 &= \alpha Z_t + \alpha(1 - \alpha) Z_{t-1} + (1 - \alpha)^2 \hat{L}_{t-2} \\
 &= \alpha Z_t + \alpha(1 - \alpha) Z_{t-1} + (1 - \alpha)^2 [\alpha Z_{t-2} + (1 - \alpha) \hat{L}_{t-3}] \\
 &= \alpha Z_t + \alpha(1 - \alpha) Z_{t-1} + \alpha(1 - \alpha)^2 Z_{t-2} + (1 - \alpha)^3 \hat{L}_{t-3} \\
 &= \alpha Z_t + \alpha(1 - \alpha) Z_{t-1} + \alpha(1 - \alpha)^2 Z_{t-2} + \dots + \alpha(1 - \alpha)^d Z_{t-d} + \dots
 \end{aligned} \tag{3.6}$$

Portanto, a estimação do nível médio corresponde à uma média ponderada das últimas observações, onde os pesos decaem conforme uma progressão geométrica de razão  $(1 - \alpha)$ . Na prática, se o valor de  $\alpha$  é baixo (próximo de 0), o decaimento dos pesos é lento e a diferença de pesos entre observações recentes e antigas não é muito significativa. Em contrapartida, se o valor de  $\alpha$  é alto (próximo de 1), as observações recentes tornam-se muito mais relevantes do que os valores passados. A Figura 3.3 ilustra a relação *peso x defasamento* para diferentes valores de  $\alpha$ .

Em geral, utiliza-se algum algoritmo de otimização para determinar o valor ótimo da constante de amortecimento  $\alpha$  (com base no ajuste obtido em uma amostra de dados). Neste caso específico, como se trata de um problema de otimização unidimensional em um intervalo limitado, uma simples busca em grade já é suficiente.

Na prática, o valor ótimo de  $\alpha$  está relacionado com a volatilidade da série. Assim, séries que possuem um comportamento muito suave, com mudanças de nível lentas, tendem a apresentar uma constante de amortecimento com valor relativamente baixo. Por outro lado, séries instáveis, com mudanças abruptas de nível, tendem a apresentar uma constante de amortecimento mais elevada.

De modo geral, os métodos de amortecimento exponencial compreendem um processo recursivo, portanto, é necessário definir um procedimento de inicialização. No caso do método de amortecimento exponencial simples a inicialização é trivial. Em geral, admite-se que o nível médio no instante inicial é o próprio valor observado ( $\hat{L}_{t=1} = Z_{t=1}$ ), e executamos o processo recursivo definido pela Equação 3.5 a partir do instante  $t = 2$ .

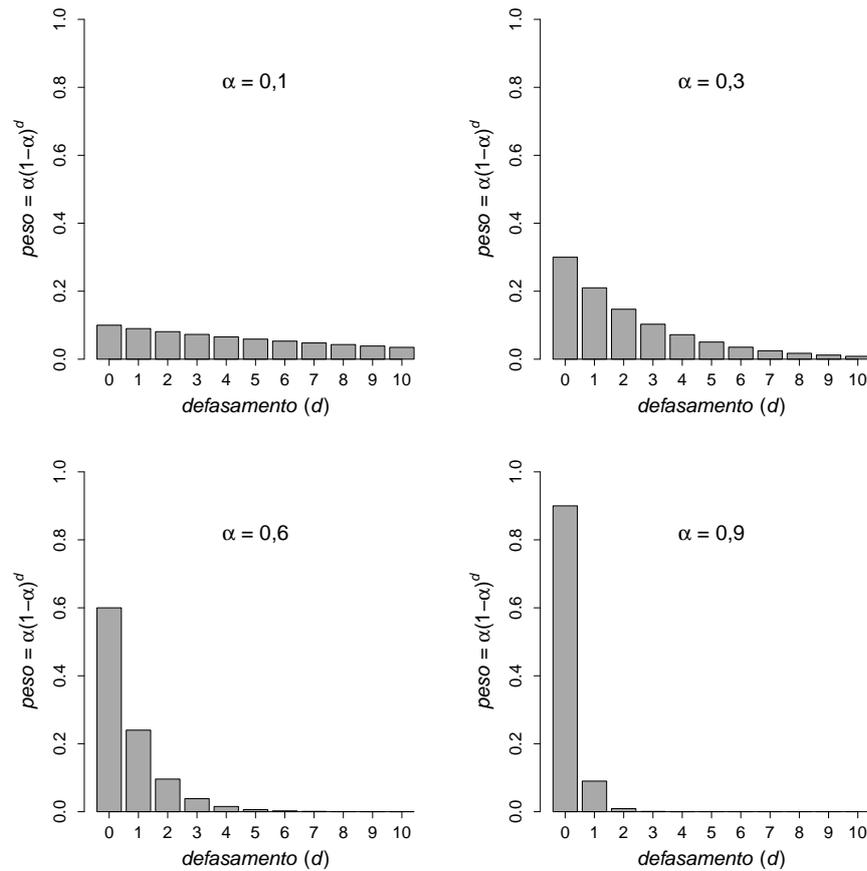


Figura 3.3: Relação *peso* x *defasamento* para diferentes valores de  $\alpha$ .

### 3.2.2 Método de Holt-Winters

O método de Holt-Winters utiliza o mesmo conceito de amortecimento exponencial apresentado na seção anterior. No entanto, este método é indicado para séries temporais mais complexas, que apresentam tendência e padrão sazonal. Evidentemente, estes fenômenos estão sobrepostos quando as observações são amostradas; porém, é possível decompor as séries nas componentes de *nível*, *tendência*, *sazonalidade* e *erro*. Existem basicamente dois modelos de decomposição para este tipo de série: o aditivo e o multiplicativo.

$$\text{Modelo Aditivo : } Z_t = L_t + B_t + S_t + \epsilon_t \quad (3.7)$$

$$\text{Modelo Multiplicativo : } Z_t = (L_t + B_t)S_t + \epsilon_t \quad (3.8)$$

Onde  $L_t$  é a componente relativa ao nível médio da série;  $B_t$  representa a tendência; e  $S_t$  refere-se ao comportamento sazonal. O erro  $\epsilon_t$ , assim como no caso anterior, é uma variável

aleatória com média zero, independente e identicamente distribuída. Conforme expressam as Equações 3.7 e 3.8, o modelo aditivo é indicado para séries em que a amplitude das variações sazonais são aproximadamente constantes, independentemente do nível médio no instante corrente; enquanto que o modelo multiplicativo é preferível quando a amplitude das variações sazonais é proporcional ao nível médio da série. De modo geral, quando as séries possuem variações de nível muito lentas, como é o caso dos dados de carga elétrica, os dois modelos tendem a apresentar resultados muito semelhantes. Neste trabalho vamos nos concentrar apenas no modelo aditivo.

No método de Holt-Winters o comportamento periódico é modelado por fatores independentes (um para cada período sazonal), enquanto a tendência segue um modelo linear. Desta forma, as previsões são calculadas de acordo com a Equação 3.9.

$$\hat{Z}_{t+h} = \hat{L}_t + h\hat{B}_t + \hat{S}_{t+h-s} \quad (3.9)$$

Cada componente da série é atualizada separadamente, conforme às Equações 3.10<sup>1</sup>.

$$\begin{cases} \hat{L}_t = \alpha(Z_t - \hat{S}_{t-m}) + (1 - \alpha)(\hat{L}_{t-1} + \hat{B}_{t-1}) & (3.10a) \\ \hat{B}_t = \beta(\hat{L}_t - \hat{L}_{t-1}) + (1 - \beta)\hat{B}_{t-1} & (3.10b) \\ \hat{S}_t = \gamma(Z_t - \hat{L}_{t-1} - \hat{B}_{t-1}) + (1 - \gamma)\hat{S}_{t-s} & (3.10c) \end{cases}$$

Onde  $\alpha$ ,  $\beta$  e  $\gamma$  são as constantes de amortecimento das componentes de nível, tendência e sazonalidade, respectivamente. O parâmetro  $s$  representa o período sazonal da série e está diretamente relacionado à frequência de amostragem dos dados; por exemplo, em séries com sazonalidade anual e amostragem mensal, deve-se adotar  $s = 12$ . Nota-se que as Equações 3.10 isolam o efeito de cada componente e utilizam o mesmo conceito de amortecimento exponencial introduzido na seção anterior.

Existem diversas formas de inicialização para o método de Holt-Winters. Makridakis *et al.* (2008) sugerem um método bastante simples para iniciar as previsões a partir do instante  $s + 1$ : o nível médio é obtido pela média das observações no primeiro período sazonal (Equação 3.11); o coeficiente angular que representa a tendência é a média das inclinações entre as observações correspondentes dos dois primeiros períodos

---

<sup>1</sup>É possível encontrar na literatura pequenas variações do método de Holt-Winters. Neste trabalho, adotamos como referência o procedimento descrito por Hyndman e Athanasopoulos (2014)

sazonais(Equação 3.12); e os fatores sazonais são dados pela diferença entre o valor observado da série e o nível médio estimado (Equação 3.13).

$$\hat{L}_s = \frac{1}{s}(Z_1 + \dots + Z_s) \quad (3.11)$$

$$\begin{aligned} \hat{B}_s &= \frac{1}{s} \left[ \frac{Z_{s+1} - Z_1}{s} + \frac{Z_{s+2} - Z_2}{s} + \dots + \frac{Z_{s+s} - Z_s}{s} \right] \\ &= \frac{1}{s^2} [(Z_{s+1} + Z_{s+2} + \dots + Z_{s+s}) - (Z_1 + Z_2 + \dots + Z_s)] \end{aligned} \quad (3.12)$$

$$\hat{S}_1 = Z_1 - \hat{L}_s \quad ; \quad \hat{S}_2 = Z_2 - \hat{L}_s \quad ; \quad \dots \quad ; \quad \hat{S}_s = Z_s - \hat{L}_s \quad (3.13)$$

### 3.2.3 Amortecimento exponencial com dupla sazonalidade

Taylor (2003) apresentou uma extensão para o método clássico de Holt-Winters, introduzindo uma nova componente sazonal e uma equação de atualização associada. Estas modificações permitiram modelar séries com dois padrões sazonais, e tornou o método mais adequado para aplicações de previsão de carga a curto prazo. Este método foi denominado de amortecimento exponencial de Holt-Winters com dupla sazonalidade (DSHW - *Double Seasonal Holt-Winters*). Em seu trabalho, Taylor utilizou a abordagem multiplicativa, no entanto, vamos adotar o modelo aditivo<sup>2</sup>, conforme Equação 3.14.

$$Z_t = L_t + B_t + D_t + W_t + \epsilon_t \quad (3.14)$$

Analogamente ao método de Holt-Winters convencional, os termos  $L_t$  e  $B_t$  representam as componentes de nível e tendência, respectivamente; enquanto,  $D_t$  e  $W_t$  representam as duas componentes sazonais, que no caso do problema de previsão de carga elétrica são as componentes de sazonalidade diária e semanal.

O método de Taylor, utiliza um modelo de tendência linear e aplica uma correção autorregressiva, conforme sugere Gardner (1985) *apud* Taylor (2003). A formulação adaptada para o caso aditivo segue o modelo representado na Equação 3.15

---

<sup>2</sup>Em geral, as variações de nível e o efeito de tendência não são relevantes para previsão de carga elétrica a curto prazo. Conforme veremos na seção , este é o caso da série *Rio-96/97*, o justifica nossa decisão de utilizar o modelo aditivo.

$$\hat{Z}_{t+h|t} = \hat{L}_t + h\hat{B}_t + \hat{D}_{t+h-s_1} + \hat{W}_{t+h-s_2} + \lambda^h e_t \quad (3.15)$$

Onde o termo  $\lambda^h e_t$  é a correção auto-regressiva efetuada com o último erro observado, utilizada para reduzir a auto-correlação dos resíduos. As equações de atualização estão expressas abaixo:

$$\begin{cases} \hat{L}_t = \alpha(Z_t - \hat{D}_{t-s_1} - \hat{W}_{t-s_2}) + (1 - \alpha)(\hat{L}_{t-1} + \hat{B}_{t-1}) & (3.16a) \\ \hat{B}_t = \beta(\hat{L}_t - \hat{L}_{t-1}) + (1 - \beta)\hat{B}_t & (3.16b) \\ \hat{D}_t = \delta(Z_t - \hat{L}_t - \hat{W}_{t-s_2}) + (1 - \delta)\hat{D}_{t-s_1} & (3.16c) \\ \hat{W}_t = \omega(Z_t - \hat{L}_t - \hat{D}_{t-s_1}) + (1 - \omega)\hat{W}_{t-s_2} & (3.16d) \end{cases}$$

As constantes  $\alpha$ ,  $\beta$ ,  $\delta$  e  $\gamma$  representam os coeficientes de amortecimento para as componentes de nível, tendência, sazonalidade diária e sazonalidade semanal, respectivamente. Os parâmetros  $s_1$  e  $s_2$  referem-se aos períodos sazonais diário e semanal, e estão diretamente ligados à frequência de amostragem dos dados.

De acordo com os resultados discutidos por Taylor *et al.* (2006), em problemas de previsão de carga a curto-prazo, o efeito da tendência normalmente é negligenciável, uma vez que em períodos curtos de tempo as variações de carga são dominadas pelos comportamentos sazonais. Portanto, é possível remover a componente de tendência do método de amortecimento exponencial DSHW, expresso pelas Equações 3.15 e 3.16. Para isso, ignoramos a Equação 3.16b e os demais termos relativos à tendência linear (isto é, o termo  $h\hat{B}_t$  na Equação 3.15 e o termo  $\hat{B}_{t-1}$ , na Equação 3.16a). Esta simplificação facilita a parametrização do modelo e, em geral, produz melhores resultados quando o efeito de tendência pode ser negligenciado. O modelo simplificado é expresso pelas Equações 3.17 e 3.18.

$$\hat{Z}_{t+h|t} = \hat{L}_t + \hat{D}_{t+h-s_1} + \hat{W}_{t+h-s_2} + \lambda^h e_t \quad (3.17)$$

$$\begin{cases} \hat{L}_t = \alpha(Z_t - \hat{D}_{t-s_1} - \hat{W}_{t-s_2}) + (1 - \alpha)\hat{L}_{t-1} & (3.18a) \\ \hat{D}_t = \delta(Z_t - \hat{L}_t - \hat{W}_{t-s_2}) + (1 - \delta)\hat{D}_{t-s_1} & (3.18b) \\ \hat{W}_t = \omega(Z_t - \hat{L}_t - \hat{D}_{t-s_1}) + (1 - \omega)\hat{W}_{t-s_2} & (3.18c) \end{cases}$$

### 3.3 Árvores de Decisão

Nesta seção introduzimos o conceito de árvores de decisão. Conforme iremos discutir, estes modelos apresentam diversas características que facilitam sua aplicação em problemas de regressão e classificação de padrões. Neste contexto, os problemas de previsão de séries temporais podem ser tratados como um caso especial de regressão, onde a variável de resposta apresenta uma autocorrelação temporal e os modelos são ajustados com base em informações passadas.

As árvores de decisão compreendem os modelos elementares que irão formar o comitê de previsão a partir do algoritmo *Gradient Boosting*, que será apresentado em detalhes na próxima seção.

#### 3.3.1 Introdução

Árvores de decisão são modelos representados em uma estrutura hierárquica que se assemelha à maneira como os seres humanos raciocinam e tomam decisões em determinadas situações. A Figura 3.4, por exemplo, representa uma árvore de decisão para fazer o diagnóstico inicial sobre o defeito de um veículo.

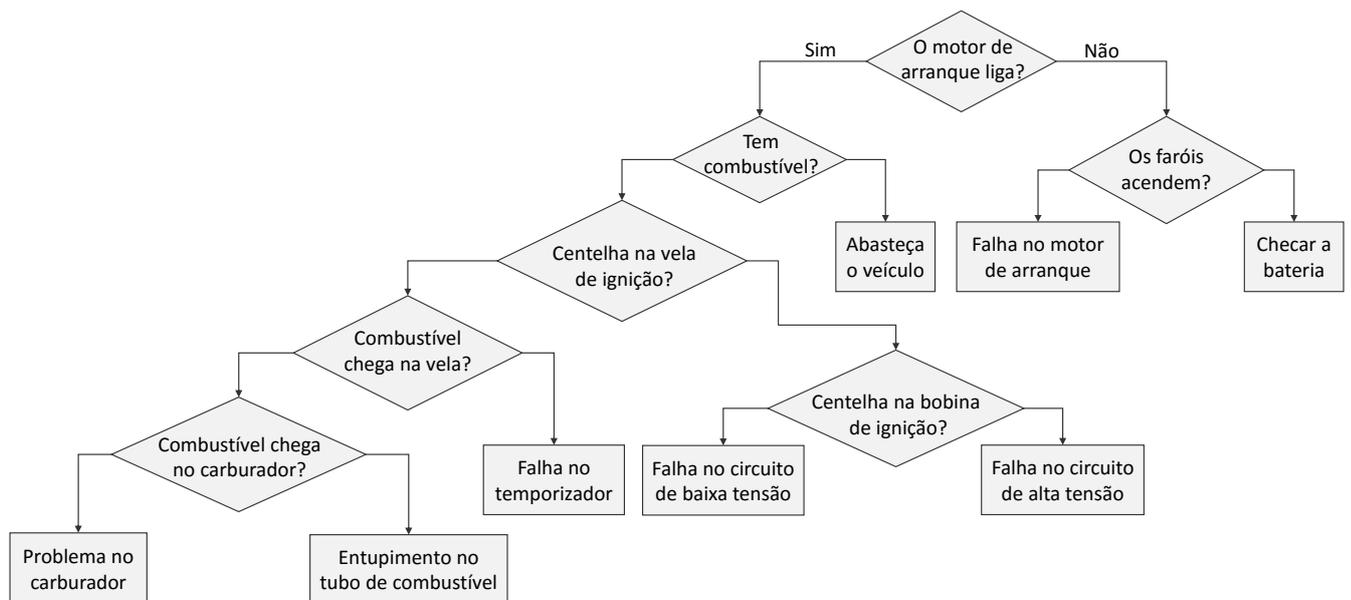


Figura 3.4: Exemplo de árvore de decisão para diagnóstico de defeito em veículo<sup>1</sup>

<sup>1</sup>exemplo retirado de <https://www.cs.auckland.ac.nz/courses/compsci367s2c/tutorials/tutorial01.pdf>

Observando a Figura 3.4, nota-se que as árvores de decisão são formadas por dois tipos de nós: nós de decisão e nós terminais. Cada nó de decisão executa um teste, cujo resultado define a direção para o próximo nó. Os testes são executados até que se atinja um nó terminal, onde é atribuída uma resposta para o problema.

Estes modelos podem ser construídos com base no conhecimento de um especialista no assunto; contudo, neste trabalho, estamos interessados em algoritmos de aprendizado supervisionado que possibilitem inferir automaticamente as regras de decisão da árvore, com base em uma amostra de dados sobre o problema. Desta forma, é possível utilizar os modelos de árvores de decisão em aplicações de aprendizado de máquina, como problemas de regressão (Figura 3.5) ou classificação de padrões (Figura 3.6).

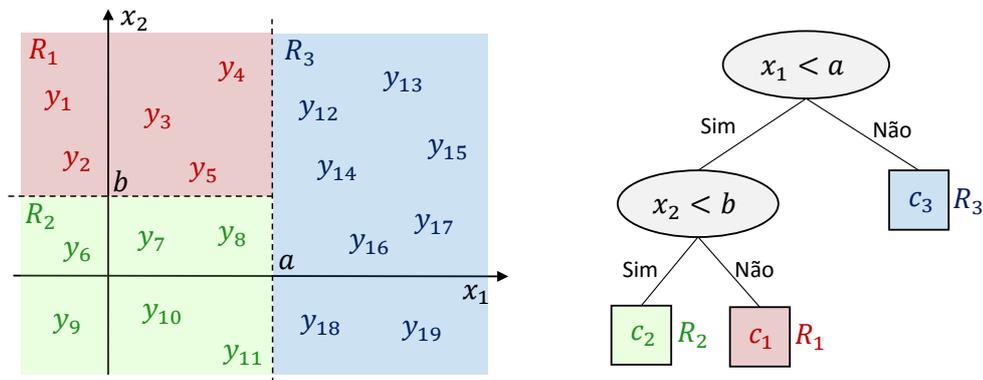


Figura 3.5: Exemplo de árvore de decisão em um problema de regressão. As variáveis  $x_1$  e  $x_2$  representam os atributos do espaço de entradas;  $y_i$  é o valor da variável de resposta correspondente à  $i$ -ésima observação da amostra; e  $c_j$  é uma constante de aproximação para os valores  $y_i$  que pertencem à região  $R_j$ .

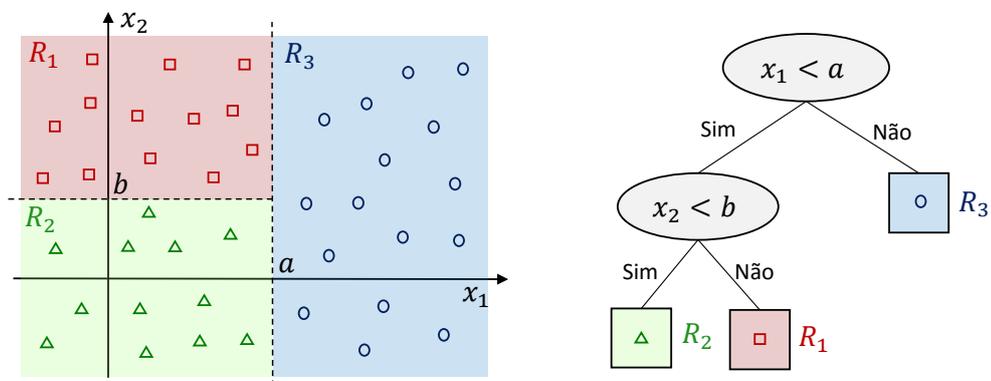


Figura 3.6: Exemplo de árvore de decisão em um problema de classificação. As variáveis  $x_1$  e  $x_2$  representam os atributos do espaço de entradas; e os símbolos  $\square$ ,  $\circ$  e  $\triangle$  indicam os rótulos de classe observados na amostra.

### 3.3.2 Algoritmo de construção

A indução de árvores de decisão é uma das formas mais simples, e ainda assim, mais bem-sucedidas, de algoritmos de aprendizagem (Russell e Norvig, 2003). De acordo com Ripley (1996), o primeiro algoritmo de indução de árvores de decisão foi o AID (Morgan e Sonquist, 1963). Posteriormente, os algoritmos CART (Breiman *et al.*, 1984) e C4.5 (Quinlan, 1993), tornaram-se muito difundidos em aplicações práticas. Atualmente, diversos métodos de aprendizado supervisionado para a construção de árvores de decisão já foram propostos na literatura; Kotsiantis (2013) e Loh (2011) apresentam revisões recentes sobre as principais características das árvores de decisão e algoritmos de construção já propostos, tanto em problemas de classificação, como de regressão.

Neste trabalho iremos restringir nossa atenção ao algoritmo CART para aplicações em problemas de regressão. Consideremos uma amostra composta por  $N$  pares  $(\mathbf{x}_i, y_i)$ , onde  $\mathbf{x}_i = [x_1^{[i]}, x_2^{[i]}, \dots, x_P^{[i]}]$  é um vetor formado por  $P$  atributos (numéricos ou categóricos), e  $y_i$  é a variável de resposta que se pretende aproximar.

Tabela 3.2: Notação adotada para representar a amostra de dados.

$n \setminus p$	$x_1$	$x_2$	$x_3$	...	$x_P$	$y$
1	$\mathbf{x}_1 = [x_1^{[1]}, x_2^{[1]}, x_3^{[1]}, \dots, x_P^{[1]}]$					$y_1$
2	$\mathbf{x}_2 = [x_1^{[2]}, x_2^{[2]}, x_3^{[2]}, \dots, x_P^{[2]}]$					$y_2$
3	$\mathbf{x}_3 = [x_1^{[3]}, x_2^{[3]}, x_3^{[3]}, \dots, x_P^{[3]}]$					$y_3$
$\vdots$	$\vdots$					$\vdots$
$N$	$\mathbf{x}_N = [x_1^{[N]}, x_2^{[N]}, x_3^{[N]}, \dots, x_P^{[N]}]$					$y_N$

As árvores de decisão particionam recursivamente o espaço de entradas, formando regiões disjuntas associadas à cada um dos nós terminais ( $R_1$ ,  $R_2$  e  $R_3$ , nos exemplos das Figuras 3.5 e 3.6). A aproximação  $c_j$  atribuída no nó terminal da árvore pode ser definida como uma medida de tendência central das observações que se encontram na região correspondente (em geral, a média em problemas de regressão; ou a moda, no caso de classificadores).

A função de aproximação  $\hat{f}(\mathbf{x})$  que representa o mapeamento efetuado por uma árvore de decisão contendo  $J$  nós terminais, pode ser expressa matematicamente conforme a

Equação 3.19.

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^J c_j I(\mathbf{x} \in R_j) \quad (3.19)$$

Onde a função  $I(\mathbf{x} \in R_j)$  é definida como:

$$I(\mathbf{x} \in R_j) = \begin{cases} 1 & \text{se } \mathbf{x} \in R_j \\ 0 & \text{caso contrário} \end{cases} \quad (3.20)$$

A construção de uma árvore de decisão consiste, então, em encontrar o melhor particionamento e as constantes de aproximação correspondentes  $\{R_j, c_j\}_{j=1}^J$ , considerando uma determinada medida de erro, por exemplo a soma dos erros quadráticos, isto é:

$$\{R_j, c_j\}_{j=1}^J = \left\{ \arg \min_{R_j, c_j} \left[ \sum_{i=1}^N (y_i - \hat{f}(\mathbf{x}_i))^2 \right] \right\}_{j=1}^J \quad (3.21)$$

Entretanto, no âmbito global, encontrar o melhor particionamento em termos de minimização dos erros quadráticos é, em geral, computacionalmente impraticável. Portanto, a solução, neste caso, é utilizar uma estratégia recursiva gulosa (*greedy*) (Hastie *et al.*, 2009). Isto é, o processo inicia com todos os dados disponíveis e efetua um particionamento que divide a amostra em duas partes  $R'$  e  $R''$ , conforme ilustra a Figura 3.7.

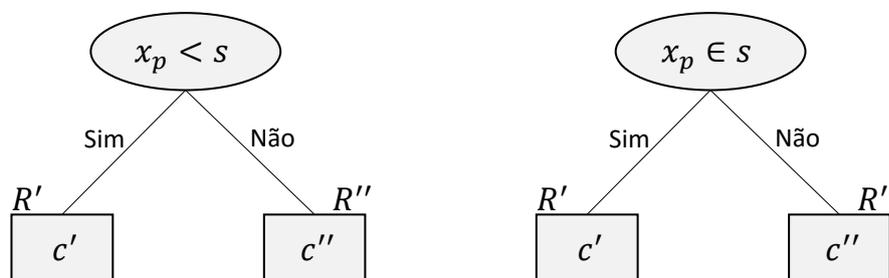


Figura 3.7: Particionamento binário durante o processo de construção da árvore, considerando  $x_p$  um atributo numérico (esq.) ou categórico (dir.)

O problema agora se resume em determinar apenas quatro parâmetros: *i*) o atributo de entrada  $x_p$  que será usado no nó de decisão; *ii*) o ponto de corte (no caso de variáveis numéricas), ou subconjunto de categorias (no caso de variáveis categóricas),  $s$  que irá segmentar a amostra em duas partes; e *iii*) as constantes de aproximação  $c'$  e  $c''$ . Ou seja, precisamos resolver o problema de minimização expresso na Equação 3.22.

$$\{x_p, s, c', c''\} = \arg \min_{x_p, s} \left[ \arg \min_{c'} \sum_{x_i \in R'} (y_i - c')^2 + \arg \min_{c''} \sum_{x_i \in R''} (y_i - c'')^2 \right] \quad (3.22)$$

Como estamos adotando a minimização do erro quadrático médio, é fácil demonstrar que as constantes de aproximação ótimas  $c'$  e  $c''$  correspondem às médias da variável de resposta das observações localizadas nas regiões  $R'$  e  $R''$ , ou seja:

$$c' = \text{média}(y_i | \mathbf{x}_i \in R') \quad c'' = \text{média}(y_i | \mathbf{x}_i \in R'') \quad (3.23)$$

Para cada variável de entrada  $x_p$ , a determinação do ponto de corte, ou subconjunto de categorias,  $s$  pode ser efetuada muito rapidamente<sup>3</sup>. Além disso, uma vez determinado o par  $(x_p, s)$ , as constantes de aproximação podem ser obtidas diretamente, a partir das Equações 3.23. Portanto, é computacionalmente viável realizar uma busca exaustiva em todos os atributos de entrada e selecionar o melhor par  $(x_p, s, c', c'')$ .

Uma vez encontrado o particionamento ótimo, o processo deve ser repetido recursivamente em cada uma das regiões obtidas, até que um determinado critério de parada seja satisfeito (algoritmo 1).

---

**Algoritmo 1** Pseudocódigo para construção de uma árvore de decisão por busca exaustiva (Loh, 2011)

---

1. Inicie pelo nó raiz utilizando todos os dados disponíveis
  2. Para cada atributo  $x_p$ , com  $p = 1 \dots P$ , determine o ponto de corte ou subconjunto  $s$  e as constantes de aproximação  $c'$  e  $c''$  que minimizam a soma dos erros quadráticos
  3. Se o critério de parada foi alcançado encerre a execução. Caso contrário, execute o passo 2 em cada um dos nós obtidos
- 

<sup>3</sup>Desde que  $x_p$  não seja uma variável categórica com um número elevado de classes distintas, conforme discutiremos na seção 5.3.2

Existem diversas formas de implementar um critério de parada durante o processo de construção da árvore. Os métodos mais simples consistem em fixar um limite máximo de profundidade ou um número mínimo de observações em cada nó terminal.

Também é possível estabelecer um limiar mínimo de ganho na função erro; neste caso, por exemplo, o particionamento só é executado se a redução na soma dos erros quadráticos for maior que o limiar estabelecido. No entanto, um particionamento aparentemente ruim pode, posteriormente, vir a produzir um particionamento com maior redução na função objetivo; o que não seria possível com a utilização desta estratégia.

Portanto, o ideal é determinar o tamanho ótimo da árvore utilizando uma estratégia adaptativa durante o próprio processo de indução (Hastie *et al.*, 2009). A solução adotada por Chen e Guestrin (2016), por exemplo, introduz um termo de regularização na função objetivo, penalizando árvores com estruturas muito complexas. No entanto, o procedimento mais comum, conforme aponta Kotsiantis (2013), consiste em dividir a construção da árvore em duas etapas: inicialmente, na fase de *crescimento*, gera-se uma árvore de decisão superdimensionada (adotando um limite máximo de profundidade e um número mínimo de observações em cada nó terminal como critério de parada); e, em seguida, na fase de *poda*, aplica-se um algoritmo para simplificar a estrutura da árvore.

O algoritmo CART implementa um mecanismo de poda que ficou conhecido como *cost-complexity pruning*. O procedimento inicia a partir da árvore de decisão superdimensionada  $T_0$ , obtida na etapa de crescimento. Em seguida, define-se  $T \subset T_0$  como uma subárvore derivada de  $T_0$ , resultante de uma operação de junção entre dois nós terminais descendentes do mesmo nó pai (Figura 3.8).

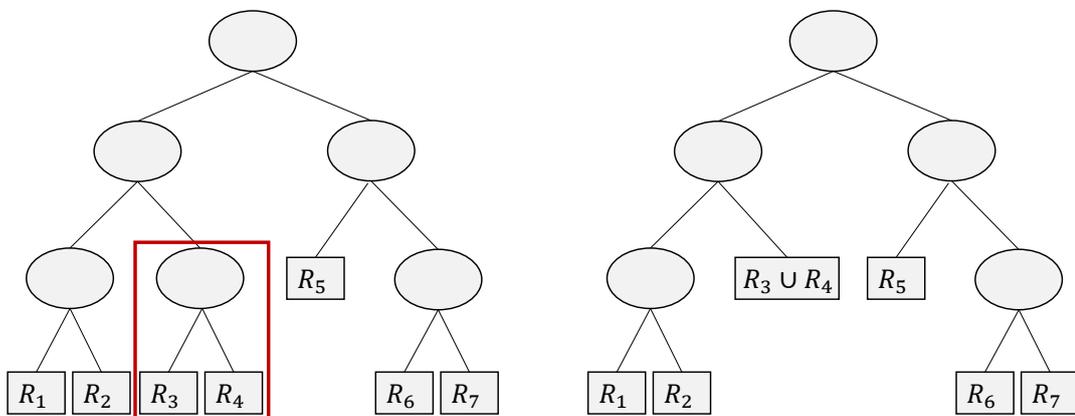


Figura 3.8: Exemplo de uma operação básica de poda.

A medida de erro de cada subárvore  $T \subset T_0$  pode ser definida como o somatório das somas dos erros quadráticos em cada região da subárvore  $T$ :

$$R(T) = \sum_{R_j \in T} \sum_{x_i \in R_j} (y_i - c_j)^2 \quad (3.24)$$

O critério de custo-complexidade é definido conforme a Equação 3.25

$$R_\xi(T) = R(T) + \xi|T| \quad (3.25)$$

Onde  $|T|$  representa o tamanho da árvore  $T$ , que pode ser definido, por exemplo, como o número de nós terminais. O parâmetro  $\xi \geq 0$  é um termo de regularização que penaliza o tamanho da árvore. O objetivo, portanto, consiste em determinar a árvore  $T_\xi \subset T_0$  que minimiza  $R_\xi(T)$ :

$$T_\xi = \arg \min_T [R(T) + \xi|T|] \quad (3.26)$$

Valores elevados de  $\xi$  resultam em árvores pequenas. Por outro lado, conforme a notação sugere, se  $\xi = 0$ , a solução é a própria árvore original  $T_0$ . O parâmetro  $\xi$  pode ser visto como um multiplicador de Lagrange, portanto, encontrar a árvore ótima para todo  $\xi$  é equivalente à encontrar a árvore que minimiza  $R(T)$  para cada tamanho possível ( $|T| = 1, 2, \dots, |T_0|$ ) (Ripley, 1996). Breiman *et al.* (1984) demonstrou que existe uma sequência de subárvores  $T_0, T_1^*, T_2^*, T_J$ , tal que cada uma é ótima dentro de um intervalo de valores para  $\xi$ , ou seja, existe uma sequência:

$$0 = \xi_0 < \xi_1 < \xi_2 < \dots < \infty \quad (3.27)$$

tal que  $T_k^*$  é a árvore ótima para  $\xi \in [\xi_k, \xi_{k+1})$ .

Breiman *et al.* (1984) descreve um procedimento para gerar uma sequência de subárvores  $T_0, T_1, \dots, T_J$  que inicia a partir da árvore  $T_0$  e executa, a cada passo, a operação de junção que provoca o menor impacto na função  $R(T)$ . Posteriormente, Ripley (1996) demonstrou que essa sequência contém a árvore ótima  $T_\xi$ .

Uma estimativa para o valor de  $\xi$  ótimo, pode ser obtida a partir de um processo de validação cruzada do tipo *k-fold*<sup>4</sup>. Em geral, utilizam-se 5 ou 10 subamostras para fazer a validação e seleciona-se o valor  $\hat{\xi}$  que minimiza a soma dos erros quadráticos entre as amostras de validação. A árvore final será  $T_{\hat{\xi}}$ .

<sup>4</sup>A seção 6.1 apresenta mais detalhes sobre o processo de validação cruzada *k-fold*

### 3.3.3 Características das árvores de decisão

Nesta seção discutimos as principais características das árvores de decisão, apresentando tanto os aspectos positivos como os negativos, além de alguns detalhes de implementação. Por fim, apresentamos uma comparação qualitativa entre os modelos de árvores de decisão e outras técnicas de inteligência computacional.

#### 3.3.3.1 Interpretabilidade

O fluxo de execução em uma árvore de decisão assemelha-se à forma como os seres humanos lidam com diversas situações práticas. Portanto, a possibilidade de interpretar as regras inferidas durante processo de construção é uma das principais vantagens das árvores de decisão, quando comparadas à outras técnicas de aprendizado de máquina.

As regras lógicas seguidas por uma árvore de decisão são muito mais fáceis de interpretar do que os pesos sinápticos de uma rede neural, por exemplo. Em aplicações reais, as pessoas sentem-se mais confortáveis utilizando modelos que elas possam compreender (Kotsiantis, 2013).

#### 3.3.3.2 Manipulação de dados numéricos e categóricos

Os nós de decisão implementam testes lógicos que podem ser executados naturalmente tanto em atributos numéricos como categóricos. Portanto, não é necessário codificar as variáveis categóricas em representações numéricas, o que é inevitável em outros algoritmos de aprendizado, como redes neurais ou regressão linear. Em determinadas situações, a forma como esta codificação é implementada pode afetar o desempenho dos modelos, conforme apontam os resultados apresentados por Crone *et al.* (2006) e Fitkov-Norris *et al.* (2012).

No entanto, a busca exaustiva descrita na seção anterior pode se tornar impraticável na presença de variáveis categóricas com muitos níveis. É fácil notar que uma variável ordinal que possui  $q$  valores distintos na amostra de dados pode ser dividida em  $q - 1$  partições binárias da forma  $x_p < s$ , onde  $s$  é um ponto de corte numérico ou ordinal. No entanto, as variáveis categóricas não ordenáveis apresentam  $(2^q - 1)$  divisões distintas da forma  $x_p \in s$ , onde  $s$  é um subconjunto não vazio dos  $q$  valores distintos que o atributo apresenta.

Além disso, conforme aponta Loh (2011), se uma variável categórica possui muitos valores distintos, sua chance de ser selecionada torna-se muito alta; pois há muitas divisões possíveis a serem consideradas. Com isso, o processo de indução da árvore pode ficar enviesado, e este tipo de variável pode ganhar uma importância exagerada no modelo final.

### 3.3.3.3 Manipulação de informações incompletas

Outra vantagem muito importante das árvores de decisão reside na sua capacidade de manipular dados faltantes. Hastie *et al.* (2009) citam duas maneiras de adaptar as árvores para trabalhar com dados faltantes.

- *Definir uma nova categoria:* este método pode ser aplicado em dados categóricos e consiste simplesmente em criar uma nova categoria para indicar que a informação não está disponível. Esta abordagem é adequada nos casos em que os dados faltantes são de natureza não-aleatória, isto é, quando as observações contendo dados incompletos comportam-se de forma diferente daquelas em que a informação está disponível.
- *Variáveis suplentes:* consiste em definir uma sucessão de atributos suplentes, que podem ser usados caso o atributo primário não esteja disponível. Durante o processo de treinamento, seleciona-se o melhor atributo e o correspondente ponto de corte (ou subconjunto de categorias) para fazer o particionamento em cada um dos nós da árvore, utilizando apenas as observações em que os atributos estão disponíveis. O primeiro atributo suplente será aquele que melhor reproduzir o particionamento obtido com o atributo primário. O próximo suplente será aquele que alcançar a segunda melhor reconstituição do particionamento obtido pelo atributo primário, e assim sucessivamente. Este método explora a correlação entre as variáveis de entrada para garantir um bom desempenho em bases de dados com informações incompletas.

### 3.3.3.4 Insensibilidade à entradas irrelevantes

O processo de treinamento das árvores de decisão envolve um procedimento automático de seleção de atributos. Na construção de cada nó da árvore, o atributo que proporcionar

o melhor ganho na função objetivo será selecionado para fazer o particionamento. Em situações normais, eventuais variáveis irrelevantes não serão selecionadas e, portanto, não são incluídas na árvore. Consequentemente, este tipo de variável não introduz ruído no modelo, fato que garante às árvores de decisão alta robustez à variáveis irrelevantes.

A Figura 3.9 representa um problema de classificação em um espaço de entradas bidimensional. A variável  $x_2$  é completamente irrelevante para separar as três classes do problema. Observe que esta variável não aparece no modelo final e, portanto não introduz ruído.

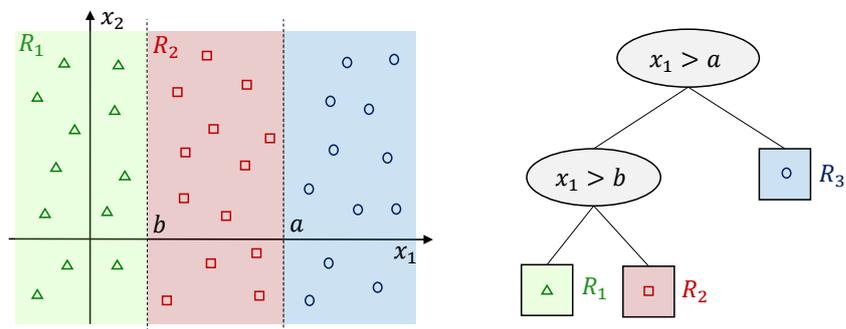


Figura 3.9: Exemplo de árvore de decisão em problemas com variáveis irrelevantes.

### 3.3.3.5 Pré-processamento de dados

O pré-processamento dos dados é uma das etapas mais importantes na construção de qualquer modelo aprendido, inclusive das árvores de decisão. No entanto, conforme já discutimos anteriormente, esses modelos apresentam algumas características que tornam o processo de aprendizado menos sensível ao pré-processamento dos dados, entre elas: *i*) insensibilidade à escala das variáveis de entrada; *ii*) manipulação natural de dados categóricos; *iii*) insensibilidade à atributos irrelevantes; e *iv*) manipulação de dados faltantes.

### 3.3.3.6 Poder preditivo

Nos algoritmos mais simples como o CART, que descrevemos nesta seção, as aproximações nos nós terminais são sempre constantes e os particionamentos consideram um único atributo por vez. Portanto, as fronteiras entre as regiões são sempre perpendiculares aos eixos e a função de aproximação é fortemente descontínua. De acordo com Hastie *et al.* (2009), esta característica não é muito danosa em problemas de classificação, onde a variável de resposta é naturalmente descontínua; mas pode prejudicar substancialmente

sua aplicação em problemas de regressão e previsão (nesse trabalho esse problema será aliviado com o uso de comitês de previsores, conforme descrito na seção 5.4).

### 3.3.3.7 Quadro comparativo

A tabela abaixo, extraída de Hastie *et al.* (2009), resume as principais características qualitativas entre algumas técnicas de inteligência computacional: redes neurais artificiais (ANN - *Artificial Neural Networks*), máquinas de vetor suporte (SVM - *Support Vector Machines*), árvores de decisão CART (CART - *Classification and Regression Trees*), *splines* de regressão adaptativa multivariada (MART - *Multivariate Adaptive Regression Splines*), e *k*-vizinhos mais próximos (*k*-NN - *k-nearest neighbors*).

Tabela 3.3: Características qualitativas de diferentes métodos de aprendizado.  
Legenda: ▲ = bom; ◆ = regular; ▼ = ruim. (Hastie *et al.*, 2009)

Característica	ANN	SVM	CART	MARS	<i>k</i> -NN
Manipulação de dados de diferentes tipos	▼	▼	▲	▲	▼
Manipulação de dados faltantes	▼	▼	▲	▲	▲
Robustez à <i>outliers</i>	▼	▼	▲	▼	▲
Insensibilidade à transformações monótonas das entradas	▼	▼	▲	▼	▼
Manipulação de grande volume de dados ( <i>N</i> elevado)	▼	▼	▲	▲	▼
Robustez à variáveis irrelevantes	▼	▼	▲	▲	▼
Habilidade para extrair combinações lineares dos atributos	▲	▲	▼	▼	◆
Interpretabilidade	▼	▼	◆	▲	▼
Poder preditivo	▲	▲	▼	◆	▲

### 3.3.4 Algoritmos mais complexos

O algoritmo CART é um dos métodos de aprendizado supervisionado mais simples que existem para a construção de árvores de decisão. É possível propor alterações a partir do algoritmo original no intuito de aprimorar o poder preditivo das árvores de decisão. O

algoritmo C4.5 (Quinlan, 1993), por exemplo, permite efetuar particionamentos múltiplos em um mesmo nó de decisão, ao invés da estratégia binária do algoritmo CART.

Loh (2011) cita três outros algoritmos (CRUISE, GUIDE e QUEST) que implementam estratégias para evitar que a seleção dos atributos seja enviesada na presença de variáveis categóricas, conforme discutimos na seção 5.3.2.

Hastie *et al.* (2009) discutem ainda a possibilidade de implementar nós de decisão para efetuar particionamentos baseados em uma combinação linear dos atributos, do tipo  $\sum a_p x_p < s$ , onde os coeficientes  $a_p$  e o ponto de corte  $s$  seriam otimizados considerando uma determinada função objetivo. Modelos lineares também poderiam ser utilizados para fazer as aproximações nos nós terminais, ao invés de se utilizar uma função constante. Contudo, como os próprios autores alertam, esse tipo de modificação reduz a interpretabilidade dos modelos e aumenta o custo computacional; além de tornar os modelos mais sensíveis à atributos irrelevantes, e dificultar a utilização de variáveis categóricas e manipulação de dados faltantes.

No entanto, o poder preditivo das árvores de decisão pode ser aprimorado com a formação de *comitês de previsão*. Na próxima seção, apresentamos o conceito de *boosting*, cuja ideia central consiste em combinar diversos modelos simples, chamados *weak learners*, para formar um poderoso comitê de previsão. Conforme iremos discutir, a construção de comitês de árvores de decisão permite aprimorar significativamente o poder preditivo individual de uma árvore, ao mesmo tempo em que consegue manter a maioria das vantagens que descrevemos nesta seção.

### 3.4 *Gradient Boosting*

Conforme discutimos na seção anterior, as árvores de decisão apresentam diversas características vantajosas para a aplicação em problemas de aprendizado de máquina. No entanto, o poder preditivo individual de uma árvore de decisão é, normalmente, muito inferior aos resultados que podem ser alcançados com o uso de outras técnicas de inteligência computacional, como redes neurais ou SVM, por exemplo. Entretanto, é possível aprimorar o poder preditivo de um modelo de aprendizado através da formação de um *comitê de previsão*, que é a ideia central das técnicas de *boosting*.

*Boosting* é uma estratégia genérica para aprimorar o desempenho de qualquer

algoritmo de aprendizado (Freund *et al.*, 1999). Originalmente, o método foi proposto para tratar problemas de classificação de padrões, com a introdução do algoritmo *AdaBoost* por Freund e Schapire (1997). Posteriormente, diversas generalizações surgiram a partir da estratégia original, entre elas, o algoritmo *Gradient Boosting* (Friedman, 2001), que pode ser aplicado tanto em problemas de classificação como de regressão, utilizando qualquer função objetivo diferenciável (Hastie *et al.*, 2009).

Neste seção apresentamos o algoritmo *Gradient Boosting*, que compreende uma estratégia sistemática para a construção de comitês de previsão. A utilização deste algoritmo em problemas de previsão de carga elétrica ainda é bastante recente, porém alguns trabalhos já demonstraram resultados muito promissores (Ben Taieb e Hyndman (2014); Lloyd (2014)), conforme a revisão apresentada no capítulo 2.

### 3.4.1 *Bagging vs. Boosting*

De acordo com Zhou (2012), existem, basicamente, dois paradigmas entre os procedimentos de construção de comitês<sup>5</sup>: os métodos de combinação paralela, em que cada modelo é treinado de forma totalmente independente; e os métodos de combinação sequencial, em que o treinamento de um novo modelo depende do resultado obtido pelo modelo anterior.

O *Bagging* (ou *Bootstrap Aggregating*), é o exemplo mais comum dos métodos de combinação paralela. A cada rodada, o algoritmo sorteia aleatoriamente um subconjunto dos dados de treinamento e utiliza essa subamostra para treinar um novo modelo de aprendizado. Em geral, o sorteio é feito com reposição<sup>6</sup>, permitindo a ocorrência de observações replicadas nos subconjuntos de treinamento de cada modelo.

O processo de *validação cruzada* também compreende uma estratégia paralela para construção de *comitês* (Dietterich, 2000); porém, neste método, a técnica mais comum, chamada *k-fold*, consiste em dividir os dados de treinamento em  $k$  subconjuntos disjuntos e de mesma cardinalidade; e, em seguida, treinar  $k$  modelos, utilizando em cada rodada um dos subconjuntos para validação e o restante dos dados para treinamento. Nestes

---

<sup>5</sup>Dietterich (2000) apresenta uma classificação mais específica, dividindo os processos de construção de comitês em cinco categorias: *i*) métodos bayesianos; *ii*) métodos de manipulação da amostra de treinamento; *iii*) métodos de manipulação no espaço de atributos; *iv*) métodos de manipulação na variável de resposta; e *v*) métodos de adição de ruído.

<sup>6</sup>No caso de sorteio sem reposição, o processo é equivalente ao método de validação cruzada de Monte Carlo (Xu e Liang, 2001)

métodos, a previsão final pode ser dada por uma medida de tendência central entre as respostas produzidas por cada um dos modelos (em geral a média, em problemas de regressão, ou a moda, no caso de classificadores).

Os métodos de *Boosting*, por outro lado, seguem o paradigma sequencial, utilizando uma estratégia que busca atuar sobre os erros obtidos na etapa anterior, com o objetivo de reduzir gradativamente os resíduos de previsão. Portanto, o processo de treinamento de cada novo modelo que irá compor o comitê precisa ser alimentado com informações sobre os erros obtidos na etapa anterior. A Figura 3.10 compara os dois paradigmas citados.

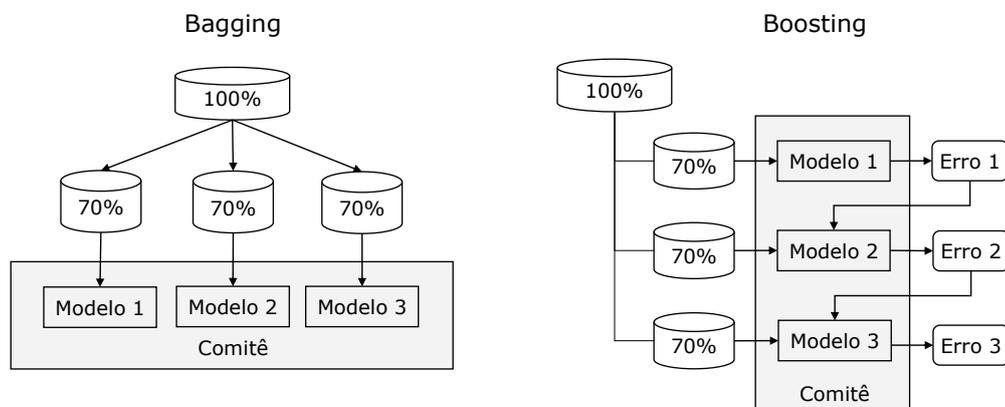


Figura 3.10: Estratégias para construção de comitês de previsores: *Bagging* x *Boosting* (os percentuais são valores parametrizáveis que indicam a proporção da subamostragem de dados em cada subconjunto da amostra total de treinamento).

Teoricamente, é possível utilizar qualquer algoritmo de aprendizado como membro básico de um comitê; entretanto, conforme aponta Alpaydin (2010), estes métodos funcionam particularmente bem para algoritmos de aprendizado *instáveis*; isto é, modelos cujas respostas podem sofrer grandes variações em decorrência de pequenas alterações no processo de treinamento (como, por exemplo, na inicialização ou na subamostragem de dados). Árvores de decisão e redes neurais são exemplos de algoritmos instáveis, enquanto modelos como regressão linear e vizinhos mais próximos são, em geral, muito estáveis (Dietterich, 2000).

A formação de comitês de árvores de decisão motivou a concepção de algoritmos de aprendizado de máquina muito bem sucedidos. Seguindo o paradigma dos métodos paralelos, o algoritmo *Random Forest* (Breiman, 2001) consiste basicamente em um processo de construção de um comitê de árvores de decisão a partir da estratégia de

*bagging*. Já na linha de métodos sequenciais, a implementação mais comum do algoritmo *Gradient Boosting* também utiliza árvores de decisão como membros básicos do comitê.

Do ponto de vista global, a utilização de um comitê de árvores de decisão prejudica a interpretabilidade do modelo. Na prática, o tamanho do comitê pode chegar a centenas ou milhares de árvores, o que inviabiliza uma análise compreensiva sobre as regras inferidas durante o processo de treinamento. Entretanto, as demais vantagens individuais das árvores de decisão podem ser mantidas, como a capacidade de manipular naturalmente dados categóricos; a habilidade para lidar com informações incompletas; a robustez à *outliers*; e a insensibilidade à atributos irrelevantes. Portanto, a construção de comitês de árvores de decisão ainda se mostra uma opção bastante atraente entre as diversas técnicas de inteligência computacional existentes.

### 3.4.2 Algoritmo Gradient Boosting

O algoritmo *Gradient Boosting* consiste em um processo iterativo aditivo. O método inicia com uma previsão constante, cujo valor corresponde à média da variável de resposta na amostra de treinamento ( $f_0(\mathbf{x}) = \bar{y}$ ). A cada iteração, um novo termo é adicionado ao modelo corrente, com o objetivo de reduzir gradualmente o erro de previsão. Assim, as atualizações são calculadas seguindo o sentido inverso do gradiente da função objetivo  $\Psi(y_i, f(\mathbf{x}_i))$ , em relação às aproximações correntes,  $f(\mathbf{x}_i)$ . O processo repete-se até que uma determinada condição de parada seja satisfeita, por exemplo, um número máximo de iterações,  $M$ .

As Equações 3.28 descrevem a estratégia aditiva do algoritmo *Gradient Boosting* para construir o comitê de previsores.

$$\begin{aligned}
 \hat{y}_i^{(0)} &= f_0(\mathbf{x}_i) = \bar{y} \\
 \hat{y}_i^{(1)} &= f_0(\mathbf{x}_i) + \eta f_1(\mathbf{x}_i) = \hat{y}_i^{(0)} + \eta f_1(\mathbf{x}_i) \\
 \hat{y}_i^{(2)} &= f_0(\mathbf{x}_i) + \eta f_1(\mathbf{x}_i) + \eta f_2(\mathbf{x}_i) = \hat{y}_i^{(1)} + \eta f_2(\mathbf{x}_i) \\
 &\dots \\
 \hat{y}_i^{(M)} &= \sum_{m=0}^M f_m(\mathbf{x}) = \hat{y}_i^{(M-1)} + \eta f_M(\mathbf{x})
 \end{aligned} \tag{3.28}$$

Onde  $\hat{y}_i^{(m)}$  representa o valor estimado da variável de resposta para a  $i$ -ésima

observação da amostra de treinamento, após a  $m$ -ésima iteração do algoritmo. O parâmetro  $\eta$ , chamado *passo de aprendizado* (*shrinkage*), controla a velocidade de convergência do algoritmo, suavizando as atualizações efetuadas a cada iteração. O comitê de previsão é formado pelas funções  $f_m(\mathbf{x})$ , que pertencem à uma classe de funções parametrizáveis  $f_m(\mathbf{x}|\boldsymbol{\theta}_m)$ , onde os parâmetros  $\boldsymbol{\theta}_m$  são determinados pelo algoritmo de treinamento dos modelos de aprendizado que compõem o comitê.

No nosso caso, as funções  $f_m(\mathbf{x}|\boldsymbol{\theta}_m)$  são árvores de decisão. Os parâmetros  $\boldsymbol{\theta}_m$  definem os particionamentos e as constantes de aproximação  $\{R_j, c_j\}_{j=1}^J$ , discutidos na seção anterior.

A cada iteração, uma nova árvore de decisão  $f_m(\mathbf{x}|\boldsymbol{\theta}_m)$  é treinada para ajustar os gradientes da função objetivo em relação às previsões do modelo corrente, considerando cada observação da amostra de treinamento ( $r_i^{(m)}$ ).

$$r_i^{(m)} = -\frac{\partial}{\partial f(\mathbf{x}_i)} \Psi(y_i, f(\mathbf{x}_i)) \Bigg|_{f(\mathbf{x}_i)=\hat{f}_m(\mathbf{x}_i)} \quad (3.29)$$

Nos problemas de regressão, em geral, a função objetivo utilizada é o erro quadrático médio (MSE) entre os valores reais e estimados para a amostra de treinamento.

$$\Psi(y_i, f(\mathbf{x}_i)) = \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (y_i - f(\mathbf{x}_i))^2 \quad (3.30)$$

Conseqüentemente, no caso particular da minimização do erro quadrático médio, os gradientes da função objetivo correspondem aos resíduos de previsão da aproximação corrente.

$$r_i^{(m)} = y_i - f_{m-1}(\mathbf{x}_i) \quad (3.31)$$

Portanto, a cada iteração uma nova árvore de decisão é treinada para ajustar os erros obtidos pelo estado corrente do comitê.

$$\boldsymbol{\theta}_m = \{R_j^{(m)}, c_j^{(m)}\}_{j=1}^J = \arg \min_{R_j, c_j} \Psi\left(r_i^{(m)}, f_m(\mathbf{x}_i|\{R_j, c_j\}_{j=1}^J)\right) \quad (3.32)$$

Finalmente, o comitê é atualizado com a adição da nova árvore de decisão recém treinada.

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \eta f_m(\mathbf{x}|\boldsymbol{\theta}_m) \quad (3.33)$$

### 3.4.3 Parâmetros e estratégias de controle

O procedimento descrito pelas Equações 3.28 a 3.33 utiliza uma estratégia iterativa para redução de resíduos. Este processo, se for executado de forma prolongada, tende a ajustar-se demasiadamente aos dados de treinamento; modelando, inclusive os ruídos aleatórios da amostra. Conseqüentemente, o modelo resultante irá apresentar um baixo desempenho quando for testado com dados diferentes daqueles utilizados em sua construção. Este problema é conhecido como *overfitting*, mas é possível adotar algumas estratégias para evitá-lo.

Difícilmente é possível estabelecer de antemão uma boa estimativa para o número de iterações  $M$ . Portanto, a estratégia mais comum, consiste em reservar uma parte dos dados de treinamento como amostra de validação. Assim, é possível monitorar o desempenho do modelo nas duas amostras e interromper o processo de treinamento quando o erro de validação atingir seu valor mínimo. Também é comum adotar algum método de validação cruzada, como, por exemplo, o *k-fold*, discutido na seção 6.1. Apesar de parecer um procedimento redundante (pois neste caso temos um comitê de comitês), esta estratégia, em geral, proporciona bons resultados, pois é possível aproveitar todos os dados de treinamento disponíveis.

A complexidade das árvores de decisão também é um importante ponto de atenção. O procedimento usual de indução consiste em construir uma árvore de decisão superdimensionada e, em seguida, aplicar um algoritmo de poda. Esta estratégia, contudo, tende a produzir árvores muito complexas no início do treinamento, o que prejudica o desempenho global do modelo e aumenta o custo computacional (Hastie *et al.*, 2009). A estratégia mais simples, neste caso, é restringir que todas as árvores do comitê tenham a mesma profundidade  $J$ . Também é comum definir um número mínimo de observações em cada nó terminal da árvore, para evitar particionamentos que contenham poucos dados.

Em geral, árvores muito profundas tendem a cometer *overfitting*; por outro lado, se a estrutura for muito simples, o modelo não consegue capturar as interações entre os

atributos de entrada. Portanto, este parâmetro deve ser definido de acordo com o grau de interação entre os atributos de entrada do problema. Esta característica, em geral, é difícil de estimar, mesmo que de forma subjetiva. No entanto, com base em experiências práticas, Hastie *et al.* (2009) sugerem testar valores dentro do intervalo  $4 \leq J \leq 8$ .

Friedman (2002) demonstrou que ao subamostrar os dados de treinamento em cada iteração, é possível obter melhores resultados em termos de generalização e ainda reduzir o custo computacional do algoritmo.

Conforme apontam os resultados apresentados por Ridgeway (2007), a redução do passo de aprendizado, em geral, produz um melhor desempenho em dados fora da amostra de treinamento. Esta medida, contudo, exige um número maior de iterações, pois as atualizações do modelo são suavizadas.

O Algoritmo 2 resume o procedimento completo de execução do algoritmo *Gradient Boosting*.

---

**Algoritmo 2** *Gradient Boosting* com árvores de decisão e minimização do erro MSE

---

1. Definir os hiper-parâmetros:

(1.a) O número máximo de iterações (tamanho do comitê),  $M$

(1.b) O passo de aprendizado (*shrinkage*),  $\eta$

(1.c) A proporção da subamostragem (*bagging*),  $\rho$

(1.d) A profundidade máxima das árvores,  $J$

(1.e) O número mínimo de observações em cada nó terminal,  $q$

2. Inicialize  $f_0(\mathbf{x}) = \bar{y}$

3. Para  $m = 1$  to  $M$

(3.a) Sub-amostrar os dados de treinamento em uma fração  $\rho$

(3.b) Para cada observação  $i$  da subamostra, calcule:

$$r_i^{(m)} = y_i - f_{m-1}(\mathbf{x}_i)$$

(3.c) Treinar a  $m$ -ésima árvore de decisão para ajustar os resíduos  $r_i^{(m)}$

$$\boldsymbol{\theta}_m = \{R_j^{(m)}, b_j^{(m)}\}_{j=1}^J = \arg \min_{R_j, b_j} \Psi\left(r_i^{(m)}, f_m(\mathbf{x}_i | \{R_j, b_j\}_{j=1}^J)\right)$$

(3.d) Atualize  $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \eta f_m(\mathbf{x} | \boldsymbol{\theta}_m)$

3. Retorne  $\hat{f}(\mathbf{x}) = f_M(\mathbf{x})$

---

## 3.5 Metodologia

Nas três seções anteriores descrevemos a técnica de amortecimento exponencial, o modelo de árvore de decisão e o algoritmo *Gradient Boosting*. Estes três métodos foram combinados para construir os modelos de previsão de carga elétrica propostos neste trabalho.

Esta seção resume a metodologia adotada de um ponto de vista global, descrevendo a maneira como os métodos foram combinados, os parâmetros utilizados, o processo de validação cruzada e outros detalhes de implementação. As metodologias utilizadas nos dados das bases *Rio-96/97* e *GEFCom2012* tiveram de ser adaptadas para atender às características de cada problema e restrições de custo computacional.

### 3.5.1 *Rio-96/97*

As previsões para a base de dados *Rio-96/97* foram calculadas em duas etapas: primeiramente utilizamos o método de amortecimento exponencial com dupla sazonalidade para obter previsões iniciais; em seguida, aplicamos o algoritmo *Gradient Boosting* para modelar os resíduos de previsão do método amortecimento exponencial e efetuar correções nas previsões obtidas na primeira etapa.

O único pré-processamento de dados realizado foi a substituição das cargas observadas nos feriados pelo valor médio registrado no mesmo dia e horário das semanas anterior e posterior, conforme já foi discutido na seção 3.1.1. Este é o mesmo tratamento adotado nos trabalhos anteriores que também adotaram essa base de dados (Hippert (2001); Neto (2014); Carneiro (2014)).

Utilizamos a versão sem tendência linear do método de amortecimento exponencial DSHW (Equações 3.17 e 3.18). Como a série de dados possui amostragem horária e os períodos sazonais são diário e semanal, definimos  $s_1 = 24$  e  $s_2 = 168$ .

As constantes de amortecimento para cada componente foram determinadas de forma a minimizar o erro MAPE (que será discutido no próximo capítulo) na amostra de treinamento. Utilizamos a implementação do algoritmo L-BFGS-B (*Limited Broyden-Fletcher-Goldfarb-Shanno for Bounded Constrained Optimization*) disponível na função *optim* do pacote *base* do R<sup>7</sup>. Os valores ótimos encontrados para cada parâmetro estão

---

<sup>7</sup>O algoritmo L-BFGS-G compreende um método numérico *quasi-Newton* para otimização de funções não-lineares que permite delimitar intervalos para as variáveis de decisão. Maiores detalhes consultar

relacionados na Tabela 3.4.

Tabela 3.4: Parâmetros ótimos determinados para o método de amortecimento exponencial com dupla sazonalidade (*Rio-96/97*)

Parâmetro	$\alpha$	$\delta$	$\omega$	$\lambda$
Valor ótimo	0,000	0,205	0,182	0,942

O valor ótimo encontrado para o parâmetro  $\alpha$  foi praticamente nulo, o que indica que a série de cargas não apresentou variações significativas no nível médio durante o período analisado, o que confirma a suposição de que a tendência pode ser desconsiderada.

Em seguida, calculamos os resíduos  $e_t^{[\text{DSHW}]}$  do método de amortecimento exponencial segundo a Equação 3.34.

$$e_t^{[\text{DSHW}]} = Z_t - \hat{Z}_t^{[\text{DSHW}]} \quad (3.34)$$

Onde  $\hat{Z}_t^{[\text{DSHW}]}$  é previsão inicial para o valor de  $Z_t$ , obtida pelo método de amortecimento exponencial DSHW. Agora nossa tarefa consiste em executar o algoritmo *Gradient Boosting* para ajustar os resíduos  $e_t^{[\text{DSHW}]}$ , incluindo os dados de temperatura que não foram utilizados no método de amortecimento exponencial, isto é:

$$GB(\mathbf{x}_t) = \hat{e}_t^{[\text{DSHW}]} = e_t^{[\text{DSHW}]} + e_t^{[\text{GB}]} \quad (3.35)$$

Onde  $GB(\mathbf{x}_t)$  é a aproximação obtida pelo algoritmo *Gradient Boosting*, a partir de um vetor de variáveis preditoras  $\mathbf{x}_t$ . O termo  $e_t^{[\text{GB}]}$  representa o erro obtido na estimação dos resíduos. Substituindo o termo  $e_t^{[\text{DSHW}]}$  na Equação 3.35 pelo resultado da Equação 3.34, temos:

$$GB(\mathbf{x}_t) = Z_t - \hat{Z}_t^{[\text{DSHW}]} + e_t^{[\text{GB}]} \quad (3.36)$$

Rearranjando os termos da Equação 3.36, temos:

$$\hat{Z}_t^{[\text{DSHW}]} + GB(\mathbf{x}_t) = Z_t + e_t^{[\text{GB}]} \quad (3.37)$$

Analisando a Equação 3.37, observamos que a parte direita da expressão consiste em uma aproximação para valor real observado  $Z_t$  sujeita aos erros  $e_t^{[\text{GB}]}$ . De fato, a parte

esquerda da Equação 3.37 compreende a previsão final do modelo proposto  $\hat{Z}_t^{[GB]}$ , que consiste em somar a aproximação dos resíduos (obtida pelo algoritmo *Gradient Boosting*) às previsões iniciais do método de amortecimento exponencial DSHW, isto é:

$$\hat{Z}_t^{[GB]} = \hat{Z}_t^{[DSHW]} + GB(\mathbf{x}_t) \quad (3.38)$$

A justificativa para este procedimento está no fato de que o algoritmo *Gradient Boosting* utiliza a série de temperaturas, que não é considerada no método de amortecimento exponencial. Portanto, a expectativa é que a média dos erros  $\bar{e}_t^{[GB]}$ , após a correção efetuada pelo *Gradient Boosting*, seja inferior ao erro médio antes das correções  $\bar{e}_t^{[DSHW]}$ . Caso contrário, o ajuste efetuado na segunda etapa foi, na verdade, prejudicial, ao desempenho do método de amortecimento exponencial.

Além disso, a aplicação do método de amortecimento exponencial DSHW reduz significativamente a autocorrelação da série de dados a ser ajustada pelo algoritmo *Gradient Boosting*, conforme demonstra a Figura 3.11.

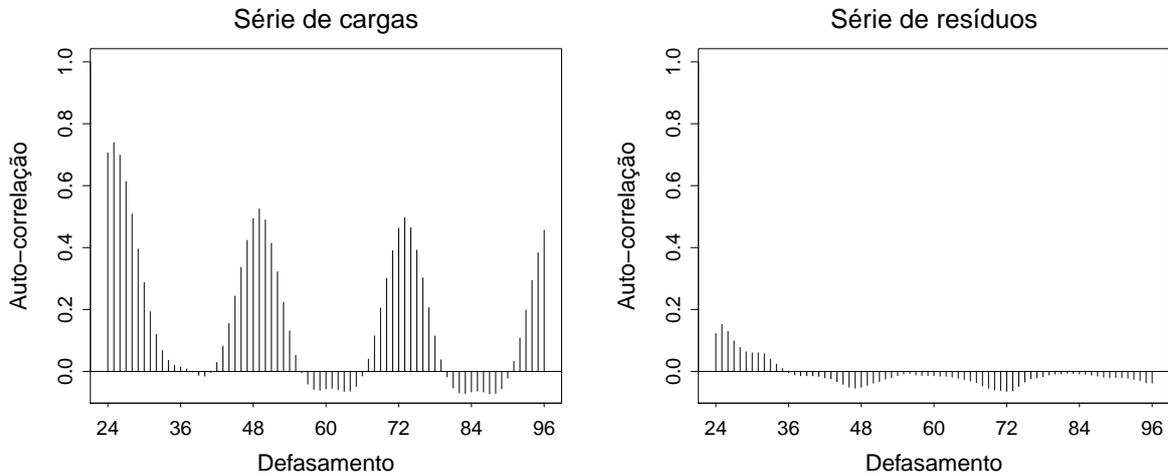


Figura 3.11: Função de auto-correlação da série de cargas (esq.) e de resíduos (dir.)<sup>2</sup>.

A Tabela 3.5 resume os parâmetros utilizados no algoritmo *Gradient Boosting*. Conforme pode-se notar, adotamos o método de validação cruzada *k-fold* com 5 subamostras<sup>3</sup>. Utilizamos uma implementação do algoritmo *Gradient Boosting* chamada

<sup>2</sup>Defasamentos de ordem menor que 24 foram omitidos, pois são inferiores ao horizonte de previsão.

<sup>3</sup>A aplicabilidade dos métodos convencionais de validação cruzada em problemas de previsão de séries temporais já foi questionada em diversas publicações, conforme discussão apresentada por Bergmeir *et al.* (2015). Entretanto, devido à baixa auto-correlação da série de resíduos e às evidências teóricas e experimentais apresentadas por Bergmeir *et al.* (2015), optamos por adotar a validação *k-fold* nos dados

GBM (*Gradient Boosting Machines*), disponível publicamente no repositório oficial da linguagem R (CRAN).

Tabela 3.5: Parametrização do algoritmo *Gradient Boosting* (*Rio-96/97*)

Parâmetro	Valor
Número máximo de iterações, $M$	7.000
Taxa de aprendizado, $\lambda$	0,005
Fração de <i>bagging</i> em cada iteração, $\rho$	50%
Profundidade máxima das árvores de decisão	8
Número mínimo de observações por nó terminal	10
Número de subamostras para validação cruzada	5

### 3.5.1.1 Variáveis preditoras

Nesta seção descrevemos o conjunto de variáveis preditoras  $\mathbf{x}_t$  utilizado pelo algoritmo *Gradient Boosting* para ajustar os resíduos do método DSHW.

- ***Demanda prevista:*** a própria demanda prevista pelo método de amortecimento exponencial pode ser útil para prever o erro de previsão. Desta forma, o algoritmo *Gradient Boosting* pode avaliar as previsões realizadas pelo método de Holt-Winters com dupla sazonalidade e então calcular as correções, considerando outras variáveis exógenas.
- ***Período do ano, semana e dia:*** uma vez que as séries de carga possuem múltiplos padrões sazonais, é importante informar ao modelo a posição em que um dado instante de tempo encontra-se dentro de cada período sazonal. Como essas variáveis são periódicas, aplicamos as funções seno e cosseno sobre o número de instantes de tempo decorridos desde o início de cada período sazonal, escalados dentro do intervalo  $[0, 2\pi]$ . Também foi incluído um número inteiro indicando a hora do dia, que corresponde ao número de instantes de tempo de antecedência em que a previsão inicial foi calculada pelo método de amortecimento exponencial.
- ***Indicadores de final de semana:*** a curva de demanda nos finais de semana possui um comportamento muito diferente quando comparada aos dias úteis. Para incorporar esta informação no modelo foram introduzidas duas variáveis lógicas indicando se uma determinada observação da amostra é referente à um sábado ou domingo.

---

da série *Rio-96/97*.

- **Temperaturas instantânea e amortecida:** a temperatura é um fator que possui grande influência no padrão de consumo de energia elétrica Hong *et al.* (2015); contudo, esta variável não é considerada no método de amortecimento exponencial.

Devido à inércia térmica dos prédios, é importante considerar não apenas as temperaturas correntes, mas também as medições observadas em instantes passados (Ben Taieb e Hyndman, 2014). É possível fazer isto utilizando métodos de suavização com médias móveis ou amortecimento exponencial, conforme proposto por Baker (1985). A média móvel da temperatura ( $TO_t$ ) e a temperatura efetiva ( $TE_t$ ) em um dado instante  $t$  são definidas de acordo com as Equações 3.39 e 3.40:

$$TO_t = \frac{1}{M} \sum_{m=0}^M T_{t-m} \quad (3.39)$$

$$TE_t = \mu TO_t + (1 - \mu) TE_{t-24} \quad (3.40)$$

Onde  $T_t$  é a temperatura no instante  $t$ ;  $M = 4$  e  $\mu = 0,5$  são parâmetros empíricos utilizados por Baker (1985).

Neste ponto é importante lembrar que as árvores de decisão executam um processo automático de seleção de atributos, portanto, as variáveis listadas aqui são apenas candidatas a serem incluídas no modelo.

A Tabela 3.6 resume as variáveis de entrada candidatas.

Tabela 3.6: Lista de variáveis preditoras candidatas (*Rio-96/97*)

ID	Variável	ID	Variável
01	Previsão DSHW ( $Z_t$ )	08	Sábado
02	Período do ano ( $sen$ )	09	Domingo
03	Período do ano ( $cos$ )	10	Hora do dia
04	Período do semana ( $sen$ )	11	$T_t$
05	Período do semana ( $cos$ )	12	$TO_t$
06	Período do dia ( $sen$ )	13	$TE_t$
07	Período do dia ( $cos$ )		

### 3.5.2 GEFCom2012

A metodologia utilizada na base de dados da competição *GEFCom2012* precisou de ser adaptada em função das características do problema e de restrições de custo

computacional.

A aplicação do método de amortecimento exponencial com dupla sazonalidade, seguida do processo de correção de resíduos com o algoritmo *Gradient Boosting*, como fizemos na base *Rio-96/97*, seria computacionalmente inviável no caso dos dados da competição *GEFCom2012*, devido às restrições de tempo. Ao todo, são 21 séries de carga elétrica (20 zonas de consumo mais a carga total do sistema), sendo que cada uma das séries possui 9 semanas de avaliação distribuídas dentro período de amostragem. Portanto, devido à natureza serial dos métodos de amortecimento exponencial, seria necessário estimar os parâmetros para  $9 \times 21 = 189$  modelos de amortecimento exponencial; e, em seguida, executar o algoritmo *Gradient Boosting* para cada modelo. Diante da restrição de custo computacional, a solução alternativa foi utilizar o algoritmo *Gradient Boosting* para prever diretamente as séries de cargas.

Os dados da competição *GEFCom2012* também exigiram que algumas medidas preliminares fossem executadas antes da construção do modelo de previsão propriamente dito. A metodologia, neste caso, foi realizada em cinco etapas:

1. Análise de dados
2. Previsão de temperaturas
3. Combinação de temperaturas
4. Previsão das séries de carga
5. Adequação às restrições do problema

### 3.5.2.1 Análise de dados

A plataforma *Kaggle*, onde a competição *GEFCom2012* foi hospedada, possui um fórum de discussão destinado à troca de informações entre as equipes participantes. Algumas particularidades encontradas nos dados da competição foram reportadas publicamente neste fórum, onde os organizadores da competição (Hong *et al.*, 2014) tiveram a oportunidade de esclarecer diversos questionamentos sobre as características da base de dados. Posteriormente, os trabalhos publicados pelas equipes mais bem colocadas (Ben Taieb e Hyndman (2014); Nedellec *et al.* (2014); Charlton e Singleton (2014); Lloyd (2014)) relataram a utilização destas características para direcionar a construção de seus modelos de previsão.

As séries de cargas das zonas 3 e 7 são absolutamente idênticas; o que indica que uma das zonas, na verdade, foi replicada. Além disso, a série de cargas da zona 2 corresponde exatamente à 92,68% dos valores observados nas zonas 3 e 7. Portanto, não há necessidade de calcular previsões para as zonas 7 e 2, uma vez que elas podem ser obtidas a partir da previsão efetuada para a zona 3.

Uma quebra estrutural pode ser identificada na série de cargas da zona 10 a partir do ano de 2008. No fórum de discussão os organizadores da competição esclareceram que este tipo de comportamento, em geral, ocorre quando há uma redefinição nas fronteiras de uma determinada zona de abastecimento. Devido à essa mudança significativa de comportamento, optamos por separar a série de cargas da zona 10 em duas partes, utilizando o dia 02/01/08 para fazer a divisão entre as duas séries.

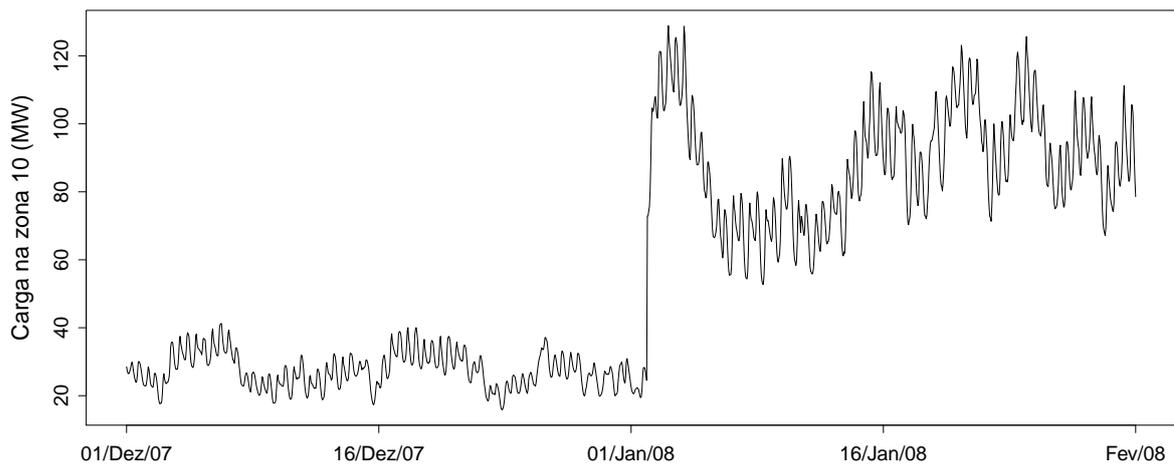


Figura 3.12: Salto no nível médio da série de cargas da zona 10.

A série de cargas da zona 9, apresentou um comportamento muito diferente das demais, com observações atípicas muito frequentes e sem qualquer relação com as temperaturas registradas nas 11 estações meteorológicas. A discussão no fórum sugeriu que essa zona situa-se em uma região industrial (posteriormente este fato foi confirmado pelos próprios organizadores da competição (Hong *et al.*, 2014)). Portanto optamos por não utilizar os dados de temperatura ao calcular as previsões de carga da zona 9, no intuito de evitar a injeção de ruído no modelo de previsão.

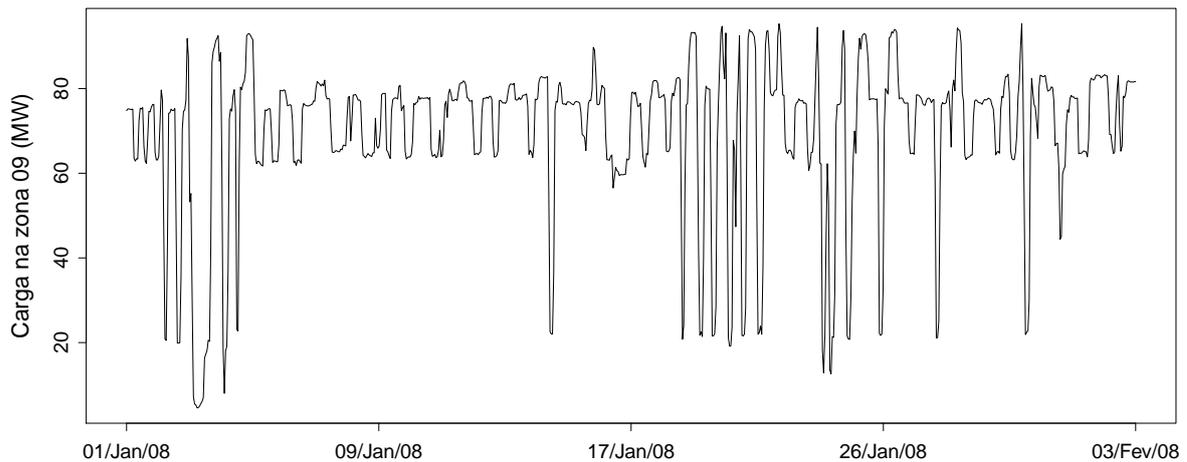


Figura 3.13: Comportamentos atípicos na série de carga da zona 09.

O apêndice A apresenta uma análise estatística mais detalhada sobre as duas bases de dados.

### 3.5.2.2 Previsão das temperaturas

Na competição *GEFCom2012*, as temperaturas registradas na última semana de dados não foram disponibilizadas. Conseqüentemente, ficou a cargo das equipes participantes efetuarem suas próprias previsões com base na série histórica de temperaturas de cada estação meteorológica.

Neste trabalho, adotamos o mesmo procedimento utilizado por Charlton e Singleton (2014). Para cada estação meteorológica  $i = 1, \dots, 11$ , cada dia  $D$  da semana de previsão, e cada hora  $h = 1, \dots, 24$ ; calculamos a temperatura média  $M_{i,D,h}$  correspondente ao mesmo dia e horário nos quatro anos anteriores, além dos 25 dias antecessores e sucessores. Assim, a previsão  $M_{i,D,h}$  é dada pela média de  $4 \times (25 + 1 + 25) = 204$  valores passados.

Em situações práticas, contudo, as previsões de temperatura são obtidas com base em observações meteorológicas realizadas por institutos de pesquisa especializados. Evidentemente, ao analisar a movimentação das massas de ar, é possível efetuar previsões de temperatura muito mais acuradas do que as estimativas que podem ser produzidas por modelos que consideram somente os dados históricos.

No entanto, nenhuma das publicações sobre previsão de carga elétrica revisadas neste trabalho reportou a utilização de modelos ajustados com base em séries de temperaturas previstas. De fato, a maior parte dos trabalhos faz o ajuste dos modelos de previsão

utilizando a série real de temperaturas observadas. Nedellec *et al.* (2014) afirmam que mesmo em suas aplicações práticas, na maior companhia de distribuição de energia elétrica da França, os modelos de previsão são ajustados com base em dados meteorológicos observados, e não nas previsões.

### 3.5.2.3 Combinação de temperaturas

Os dados da competição *GEFCom2012* contém as séries de cargas elétricas de 20 zonas de consumo, além de 11 séries de temperaturas registradas em diferentes estações meteorológicas localizadas na mesma região. Contudo, não foi disponibilizada nenhuma informação a respeito da localização geográfica das zonas de consumo e das estações meteorológicas. Portanto, um dos desafios da competição, consiste em selecionar as séries de temperaturas que serão utilizadas para prever as cargas de cada zona.

O modelo proposto por Hong (2010), que foi adotado como *benchmark* da competição, utiliza um modelo de regressão linear múltipla para prever as cargas em função da série de temperaturas e de algumas variáveis de calendário. O método proposto consiste em ajustar um modelo utilizando cada uma das séries de temperaturas separadamente, e então selecionar aquele que melhor se ajustou aos dados de treinamento.

No entanto, executar o algoritmo *Gradient Boosting* para cada série de temperaturas demanda um custo computacional elevado. Além disso, conforme aponta Charlton e Singleton (2014), realizar as previsões de carga com base em apenas uma série de temperaturas não é muito seguro, pois as medições das estações meteorológicas estão sujeitas à falhas e ruídos; e uma zona de consumo pode estar geograficamente equidistante de duas ou mais estações meteorológicas.

Neste contexto, Hong *et al.* (2015) discute duas abordagens possíveis: *i*) combinar as previsões obtidas por modelos ajustados com base em diferentes séries de temperaturas; ou *ii*) combinar as medições de temperatura e treinar um modelo utilizando a combinação obtida.

Executar o algoritmo *Gradient Boosting* para cada uma das séries de temperaturas requer um custo computacional muito maior do que combinar as temperaturas e treinar um único modelo. Portanto, estratégia adotada neste trabalho consiste em definir uma estação meteorológica virtual, cuja série de temperaturas  $T_t^{[v]}$  é uma média ponderada das medições registradas nas 11 estações meteorológicas no instante  $t$ .

$$T_t^{[v]} = \frac{a_1 T_t^{[1]} + a_2 T_t^{[2]} + \dots + a_{11} T_t^{[11]}}{\sum a_i} \quad (3.41)$$

Onde  $T_t^{[i]}$  denota a temperatura observada na estação meteorológica  $i$  no instante  $t$ . O problema agora consiste em determinar os coeficientes de ponderação  $a_1, \dots, a_{11}$ . Para isso, vamos considerar um polinômio de terceiro grau para ajustar os valores de carga em função das temperaturas observadas, conforme proposto por Hagan e Behr (1987).

$$P_t^{[k,i]} = b_0 + b_1 T_t^{[i]} + b_2 (T_t^{[i]})^2 + b_3 (T_t^{[i]})^3 \quad (3.42)$$

Onde  $P_t^{[k,i]}$  é o ajuste polinomial para a carga na zona  $k = 1, \dots, 21$  em função da temperatura na estação meteorológica  $i = 1, \dots, 11$ , em dado instante  $t$ . Os coeficientes  $b_0, \dots, b_3$  podem ser facilmente estimados pelo critério de mínimos quadrados. A Figura 3.14 ilustra o diagrama de dispersão entre as cargas da zona de consumo  $k = 18$  e as temperaturas das estações meteorológicas  $i = 09$  e  $i = 01$ . A figura também demonstra os ajustes polinomiais  $P_t^{[18,9]}$  e  $P_t^{[18,1]}$ .

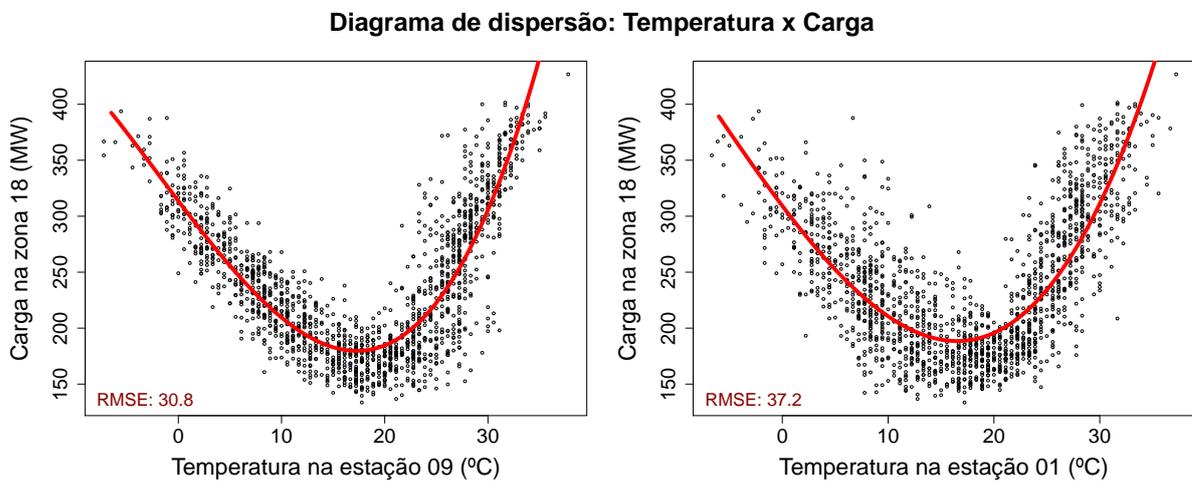


Figura 3.14: Diagrama de dispersão ‘temperatura x carga’ e ajuste polinomial de 3<sup>a</sup> ordem

A Figura 3.14 demonstra uma correlação não linear muito evidente entre os dados de temperatura e carga. Visualmente, podemos notar que as temperaturas registradas na estação meteorológica 09 (esq.) possuem uma dispersão menor no entorno da aproximação polinomial do que as medições registradas na estação 01 (dir.). Assim, podemos utilizar a raiz do erro quadrático médio (RMSE) do ajuste polinomial para classificar as melhores

estações meteorológicas para cada zona de consumo.

$$RMSE(Z_t^{[k]}, P_t^{[k,i]}) = \sqrt{\text{média}([Z_t^{[k]} - P_t^{[k,i]}]^2)} \quad (3.43)$$

Onde  $Z_t^{[k]}$  é a carga observada na zona  $k$  no instante  $t$ .

No entanto, para cada série de carga  $k = 1, \dots, 21$ , estamos interessados em determinar os coeficientes de ponderação  $a_1, \dots, a_{11}$  que irão formar uma série virtual de temperaturas médias ponderadas  $T_t^{[v]}$  (Equação 3.41) de forma a minimizar o erro  $RMSE(Z_t^{[k]}, P_t^{[k,i=v]})$ , isto é:

$$a_i = \arg \min_{a_i} \left[ RMSE \left( Z_t^{[k]}, P_t^{[k,i=v]} \mid T_v = \frac{a_1 T_t^{[1]} + a_2 T_t^{[2]} + \dots + a_{11} T_t^{[11]}}{\sum a_i} \right) \right] \quad (3.44)$$

Os coeficientes ótimos  $a_1, \dots, a_{11}$  foram determinados pelo mesmo algoritmo de otimização *L-BFGS-B*, utilizado para encontrar as constantes do modelo de amortecimento exponencial com dupla sazonalidade (seção 7.1).

### 3.5.2.4 Previsão das séries de cargas

A previsão de cada série de cargas foi realizada separadamente, isto é, executamos o algoritmo *Gradient Boosting* para cada uma das zonas  $z = 1, \dots, 21$  individualmente.

As variáveis preditoras candidatas foram semelhantes às utilizadas na base *Rio-96/97*, conforme demonstra a Tabela 3.7.

Tabela 3.7: Lista de variáveis preditoras candidatas (*GEFCom2012*)

ID	Variável	ID	Variável
01	Instante de tempo	10	Hora do dia
02	Período do ano ( <i>sen</i> )	11	Sábado
03	Período do ano ( <i>cos</i> )	12	Domingo
04	Período do semana ( <i>sen</i> )	13	Fim de semana
05	Período do semana ( <i>cos</i> )	14	Feriado
06	Período do dia ( <i>sen</i> )	15	$T_t^{[v]}$
07	Período do dia ( <i>cos</i> )	16	$(T_t^{[v]})^2$
08	Estação do ano	17	$TO_t^{[v]}$
09	Mês do ano	18	$TE_t^{[v]}$

Adicionamos o instante de tempo (sequência cronológica de números inteiros iniciando no instante  $t = 1$ ), o mês do ano e a temperatura ao quadrado, conforme o modelo proposto por Hong (2010). Charlton e Singleton (2014) alcançaram bons resultados ao dividir a amostra de dados por estação do ano, o que nos motivou a incluir esta variável em nosso modelo. Como os dados da competição *GEFCom2012* possuem feriados no período de teste, incluímos duas variáveis *dummies* de calendário; uma para indicar se a observação é referente a um final de semana (feriados foram tratados como final de semana, conforme sugere Hong (2010)) e outra para indicar especificamente os feriados.

A série de temperaturas utilizada foi a média ponderada resultante da combinação ótima  $T_t^{[v]}$ , conforme descrito na seção anterior. Esta série também foi usada para calcular os valores das temperaturas amortecidas  $TO_t^{[v]}$  e  $TE_t^{[v]}$ , de acordo com as Equações 3.39 e 3.40, respectivamente.

A Tabela 3.8 apresenta a parametrização definida para o algoritmo *Gradient Boosting*, após alguns testes experimentais com os dados da competição *GEFCom2012*. Apesar de cada zona de consumo ter suas próprias características, adotamos a mesma parametrização em todos os modelos, devido a restrições de tempo.

Tabela 3.8: Parametrização do algoritmo *Gradient Boosting* (*GEFCom2012*)

Parâmetro	Valor
Número máximo de iterações, $M$	20.000
Taxa de aprendizado, $\lambda$	0,0025
Fração de <i>bagging</i> em cada iteração, $\rho$	50%
Profundidade máxima das árvores de decisão	8
Número mínimo de observações por nó terminal	10
Percentual da amostra de validação	20%

### 3.5.2.5 Adequação às restrições do problema

Conforme discutimos na seção 3.1.2, as regras estabelecidas para a base de dados *GEFCom2012* impõem que a previsão para a carga total do sistema elétrico seja correspondente à soma das cargas de todas as zonas de consumo. Existem basicamente duas maneiras de satisfazer este requisito:

- **Previsão bottom-top:** nesta abordagem, a carga de cada zona de consumo é prevista de forma independente, e a carga total do sistema é determinada pelo

somatório das previsões de cada zona de consumo. Neste caso, a restrição do problema é automaticamente satisfeita.

- **Previsão top-bottom:** este método realiza, primeiramente, uma previsão para a carga total do sistema e, em seguida, distribui uma fração da carga prevista para cada uma das zonas de consumo.

O método *bottom-top* é relativamente mais simples e foi utilizado em todos os trabalhos revisados (Hong *et al.* (2014); Nedellec *et al.* (2014); Charlton e Singleton (2014); Ben Taieb e Hyndman (2014); Lloyd (2014)). No entanto, neste trabalho, avaliamos as duas abordagens.

Para a estratégia *top-bottom*, a previsão para a carga total do sistema foi obtida com mesma metodologia utilizada em cada uma das zonas de consumo, porém aplicada na série formada pelas somas das cargas em cada zona. Em seguida, as previsões de cada uma das zonas de consumo foram ajustadas com base na razão obtida entre a previsão para a carga total do sistema e o somatório das previsões de cada uma das zonas, de tal forma a garantir que o somatório das cargas de cada zona fosse correspondente à previsão realizada para o consumo total do sistema.

Matematicamente temos:

$$\hat{Z}_t^{[k]'} = \frac{\hat{Z}_t^{[21]}}{\sum_{j=1}^{20} \hat{Z}_t^{[j]}} \hat{Z}_t^{[k]} \quad (3.45)$$

Onde  $\hat{Z}_t^{[k]'}$  é a previsão final ajustada para a zona  $k$ ,  $\hat{Z}_t^{[21]}$  é a previsão realizada para a carga total do sistema e  $\hat{Z}_t^{[k]}$  é a previsão inicial para a zona de consumo  $k$ , todos em um determinado instante  $t$ .

Somando as previsões finais ajustadas, garantimos que:

$$\begin{aligned} \sum_{i=1}^{20} \hat{Z}_t^{[k]'} &= \frac{\hat{Z}_t^{[21]}}{\sum_{j=1}^{20} \hat{Z}_t^{[j]}} \hat{Z}_t^{[1]} + \frac{\hat{Z}_t^{[21]}}{\sum_{j=1}^{20} \hat{Z}_t^{[j]}} \hat{Z}_t^{[2]} + \dots + \frac{\hat{Z}_t^{[21]}}{\sum_{j=1}^{20} \hat{Z}_t^{[j]}} \hat{Z}_t^{[20]} \\ \sum_{i=1}^{20} \hat{Z}_t^{[k]'} &= \frac{\hat{Z}_t^{[21]} \left[ \sum_{j=1}^{20} \hat{Z}_t^{[j]} \right]}{\sum_{j=1}^{20} \hat{Z}_t^{[j]}} \\ \sum_{i=1}^{20} \hat{Z}_t^{[k]'} &= \hat{Z}_t^{[21]} \end{aligned} \quad (3.46)$$

A carga total do sistema é menos susceptível a fatores exógenos atípicos e erros de amostragem. Além disso, os efeitos causados por *outliers* nas séries de cargas de cada zona são suavizados na série agregada e os pesos atribuídos na função de avaliação da competição *GEFCom2012* valorizam bastante as previsões para a carga total do sistema, conforme veremos na seção 4.2. Portanto, podemos esperar que a previsão *top-bottom* apresente um resultado superior, desde que as séries componentes não apresentem muitas quebras estruturais (o que foi verificado apenas na série da zona 10).

## 4 Resultados

De acordo com Hong *et al.* (2014), a validação insuficiente dos modelos propostos é uma das principais deficiências nos trabalhos sobre previsão de carga elétrica. Em geral, os dados utilizados e a implementação dos modelos não são divulgados publicamente, o que dificulta a reprodutibilidade dos experimentos. A maior parte dos trabalhos apresenta comparações apenas entre modelos de previsão implementados pelos próprios autores. Normalmente, o desempenho dos modelos propostos não é comparado com o estado da arte nesta linha de pesquisa.

Neste cenário, a principal motivação para a escolha das bases *Rio-96/97* e *GEFCom2012* foi a disponibilidade dos dados e possibilidade de comparar os resultados com trabalhos já publicados anteriormente. Portanto, seguimos a mesma divisão de dados (treinamento e teste), requisitos de projeto e métricas de desempenho adotados nos trabalhos anteriores.

### 4.1 *Rio-96/97*

O critério de desempenho adotado para a base *Rio-96/97* foi o erro absoluto percentual médio (MAPE), expresso pela Equação 4.1.

$$MAPE = \frac{1}{N} \sum_{t=1}^N APE_t \quad (4.1)$$

Onde  $N$  é o número de observações da amostra; e  $APE(t)$  é o erro absoluto percentual de previsão correspondente ao instante  $t$ , calculado de acordo com a Equação 4.2

$$APE_t = \frac{|Z_t - \hat{Z}_t|}{Z_t} \quad (4.2)$$

O erro MAPE é uma das métricas de desempenho mais utilizadas em previsão de séries temporais, pois fornece uma noção facilmente interpretável da magnitude dos erros de previsão. Contudo, esta não é uma medida de erro adequada para séries de dados que apresentam valores iguais ou relativamente próximos de zero, uma vez que o valor observado da série aparece no denominador da Equação 4.2. Como, este não é o caso da

série de cargas elétricas da base *Rio-96/97*, podemos adotar o erro MAPE como critério de avaliação, assim como foi feito nos trabalhos de Hippert (2001) e Neto (2014) .

A série de cargas da base *Rio-96/97* possui exatamente 104 semanas e foi dividida em dois períodos: treinamento e teste. Os dados das 84 semanas iniciais (01/01/96 a 10/08/97) foram utilizados para ajustar o modelo de previsão, isto é, estimar os parâmetros do método de amortecimento exponencial e executar o algoritmo *Gradient Boosting*. Já os dados das últimas 20 semanas (11/08/97 a 28/12/97) foram utilizados somente para avaliar o desempenho das previsões. Esta é a mesma divisão utilizada no trabalho de Neto (2014).

A parte esquerda da Tabela 4.1 apresenta os resultados obtidos com o método de amortecimento exponencial DSHW e com a inclusão do processo de correção de resíduos através do algoritmo *Gradient Boosting*. Os resultados também são comparados com o algoritmo *Gradient Boosting* utilizado para prever diretamente a carga (e não os resíduos do método DSHW), e com referência aos métodos de previsão ingênuos (*naive*). Para o método *naive24*, as previsões correspondem ao valor observado no mesmo horário do dia anterior ( $\hat{Z}_t = Z_{t-24}$ ); enquanto que o método *naive168* considera o mesmo dia e horário da semana anterior ( $\hat{Z}_t = Z_{t-168}$ ). A parte direita da tabela inclui os resultados publicados por Neto (2014) que utilizou o método DSHW em combinação com um modelo de regressão linear múltipla (RLM) e redes neurais artificiais (RNA).

Tabela 4.1: Erros MAPE na amostra de teste (*Rio-96/97*)

Modelo	Resultado	Modelo (Neto, 2014)	Resultado
Naive24	7,17%	DSHW + RLM	2,29%
Naive168	6,12%	min	2,35%
Gradient Boosting	4,08%	DSHW + RNA <sup>1</sup>	méd 2,37%
DSHW	2,47%	max	2,40%
DSHW + <i>Boosting</i>	<b>2,02%</b>		

O método de previsão *naive168*, apesar de definir as previsões com base em observações mais antigas que o método *naive24*, apresentou um resultado médio melhor. Este fato pode ser explicado pelo padrão periódico semanal da série de cargas, que não é considerado pelo método *naive24*; cujas previsões para as cargas de segunda-feira, por exemplo, são definidas pelo valor observado no mesmo horário do dia anterior, no caso um domingo, o

<sup>1</sup>Devido a característica estocástica do processo de inicialização das redes neurais, Neto (2014) treinou um total de 100 redes e publicou os resultados mínimo, médio e máximo.

que conduz a erros muito altos, uma vez que as demandas nestes dias são muito diferentes.

A utilização do algoritmo *Gradient Boosting* para prever diretamente a carga obteve um desempenho melhor que os previsores ingênuos, porém bem inferior ao método DSHW sozinho. Apesar de não considerar variáveis exógenas, o método DSHW é um modelo proposto especificamente para problemas de previsão de séries de carga elétrica, o que lhe proporcionou um resultado superior ao algoritmo *Gradient Boosting* que é um método genérico para problemas de regressão e classificação

No entanto, a inclusão do processo de correção de resíduos através do algoritmo *Gradient Boosting* ainda foi capaz de proporcionar um ganho de desempenho significativo no método DSHW puro, reduzindo o erro MAPE de 2,47% para 2,02%. Esta redução superou os resultados obtidos por Neto (2014) com o uso dos métodos de regressão linear e redes neurais, o que destaca a capacidade preditiva do algoritmo *Gradient Boosting*.

Uma avaliação mais criteriosa dos modelos de previsão deve considerar também a distribuição dos erros obtidos, e não apenas o valor médio. O *boxplot* da Figura 4.1 comprova a efetividade do procedimento de correção de resíduos.

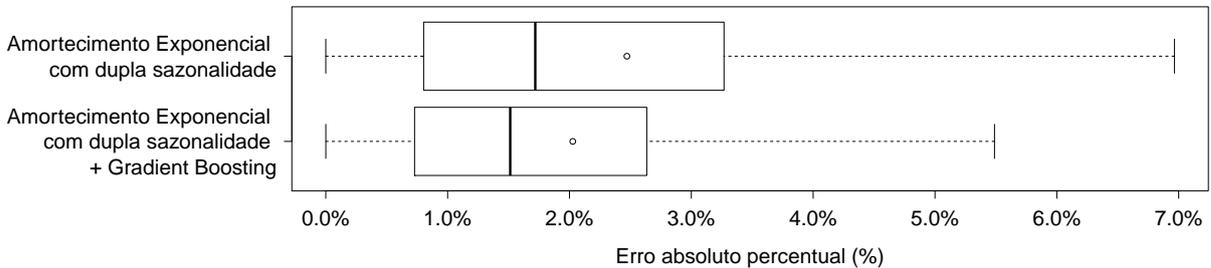


Figura 4.1: *Boxplot* dos erros absolutos percentuais (o símbolo “o” representa a média aritmética; *outliers* foram omitidos para facilitar a visualização).

A Figura 4.2, contém os diagramas de dispersão entre as cargas observadas e previstas, utilizando os modelos de previsão ingênuos, a técnica de amortecimento exponencial DSHW pura, e a implementação do processo de correção de resíduos com o algoritmo *Gradient Boosting*. A figura também apresenta o erro quadrático médio (MSE - *Mean Square Error*) entre os valores reais e previstos.

$$MSE = \frac{1}{N} \sum_{t=1}^N (Z_t - \hat{Z}_t)^2 \quad (4.3)$$

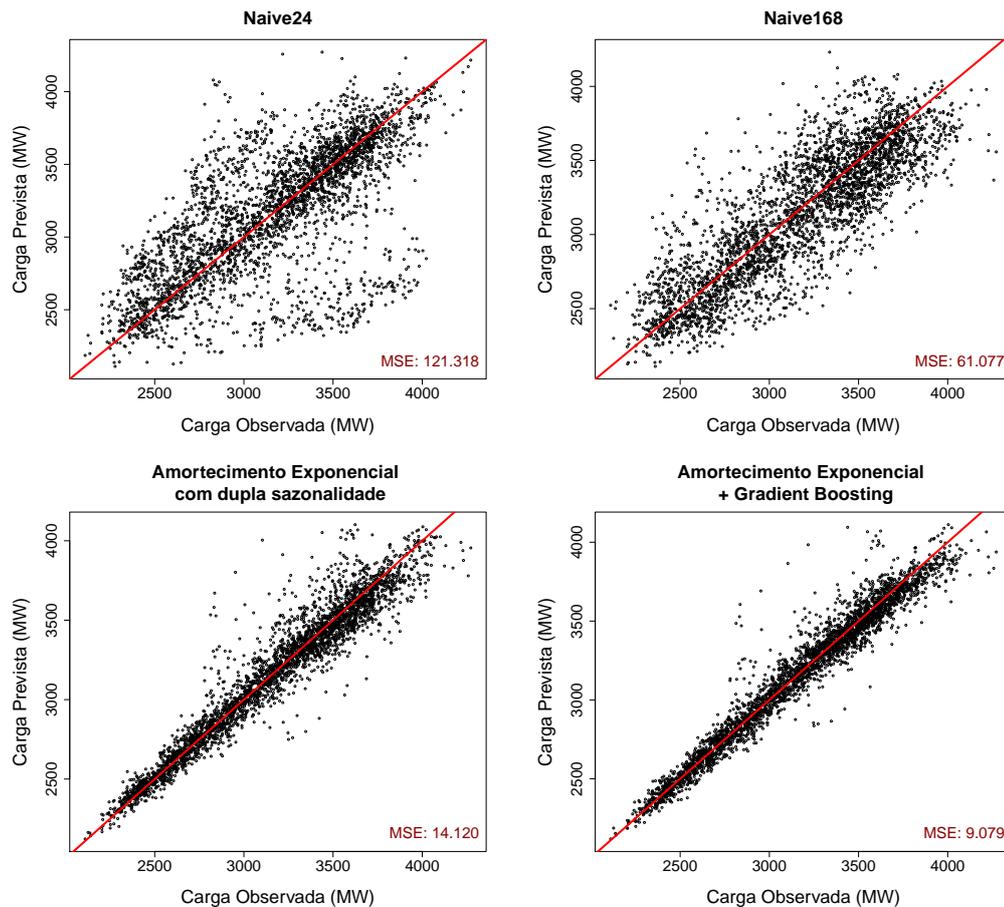


Figura 4.2: Diagrama de dispersão entre a carga prevista e observada para os modelos utilizados neste trabalho.

Analisando os gráficos da figura 4.2, verifica-se que o método *naive24* comete erros muito elevados por desconsiderar o padrão periódico semanal (os pontos na região inferior direita são cargas de segunda-feira subestimadas com base na demanda no domingo). O erro quadrático médio foi reduzido de 14.120 para 9.079 com a introdução do processo de correção de resíduos no método de Holt-Winters com dupla sazonalidade.

O processo de otimização das constantes de amortecimento do método DSHW (Tabela 3.4) foi concluído em cerca de 143s, enquanto a execução do algoritmo *Gradient Boosting* levou aproximadamente 302s. Ambos executados em linguagem R, em um computador portátil pessoal equipado com processador Intel Core i7 com 8GB de memória RAM sob o sistema operacional Windows 10.

## 4.2 *GEFCom2012*

O critério de avaliação da competição *GEFCom2012* foi a raiz do erro quadrático médio ponderado WRMSE (*Weighted Root Mean Square Error*), expresso pela Equação 4.4.

$$WRMSE = \sqrt{\frac{\sum_{k,t} w_{k,t} (Z_t^{[k]} - \hat{Z}_t^{[k]})^2}{\sum_{k,t} w_{k,t}}} \quad (4.4)$$

Onde  $w_{k,t}$  é o peso atribuído à previsão de carga da zona  $k$  no instante  $t$ . O somatório da Equação 4.4 compreende as previsões de todas as semanas da amostra de teste e zonas de consumo, inclusive a zona 21, que representa a carga total do sistema. Os pesos  $w_{k,t}$  foram atribuídos pelos organizadores da competição conforme a Tabela 4.2.

Tabela 4.2: Distribuição dos pesos de previsão - *GEFCom2012*.

Previsão	Peso ( $w_{k,t}$ )
Consumo por zona nas semanas intermediárias	1
Consumo por zona na semana final	8
Consumo total nas semanas intermediárias	20
Consumo total na semana final	160

Esta distribuição de pesos valoriza as previsões referentes à carga total do sistema e à semana final. Neste ponto, vale lembrar que os dados de teste são compostos por 8 semanas intermediárias (*backcasting*) e uma semana final (*forecasting*); e a carga total do sistema (zona 21) é a soma das cargas das 20 zonas de consumo. Portanto, o peso da previsão para a carga total do sistema é equivalente às previsões de todas as zonas juntas; e o peso das previsões para a semana final é equivalente ao peso de todas as semanas intermediárias juntas.

No entanto, como estamos adotando o erro WRMSE, a magnitude dos termos quadráticos  $(Z_t^{[k]} - \hat{Z}_t^{[k]})^2$  na média da Equação 4.4 depende fortemente do nível médio da carga. Portanto, a relevância de cada zona no critério final de avaliação está diretamente ligada à sua participação no consumo total do sistema (ver Figura A.5, no apêndice A).

A Tabela 4.3 foi extraída de Hong *et al.* (2014) e apresenta os resultados alcançados

pelas equipes mais bem sucedidas na competição *GECom2012*. Incluímos também os resultados obtidos neste trabalho, utilizando as estratégias *top-bottom* e *bottom-top* para obter as previsões para a carga total do sistema. Conforme havíamos esperado (subseção 3.5.2.5), a abordagem *top-bottom* alcançou o melhor resultado e teve um desempenho bastante competitivo com os trabalhos mais bem sucedidos na competição, com uma diferença percentual de apenas 1,72% em relação ao melhor resultado publicado (Nedellec *et al.*, 2014).

Tabela 4.3: Resultados da competição *GEFCom2012* em toda amostra de teste (Hong *et al.*, 2014)

Kaggle ID (Publicação)	Resultado	Diferença percentual em relação ao melhor resultado
Tololo (EDF) (Nedellec <i>et al.</i> , 2014)	67.223	-
CountingLab (Charlton e Singleton, 2014)	68.160	1,39%
<b>Este trabalho (<i>top-bottom</i>)</b>	<b>68.377</b>	<b>1,72%</b>
TinTin (Ben Taieb e Hyndman, 2014)	71.033	5,67%
James Lloyd (Lloyd, 2014)	71.387	6,19%
<b>Este trabalho (<i>bottom-top</i>)</b>	<b>73.710</b>	<b>9,65%</b>
Quadrivio	76.816	14,27%
Chaotic Experiments	84.209	25,27%
TheJellyTeam	87.826	30,65%
Andrew L	89.456	33,07%
NHH	90.385	34,46%
<i>Benchmark</i> (Hong (2010); Hong <i>et al.</i> (2014))	100.385	49,33%

O modelo proposto apresentou um resultado superior aos métodos apresentados por Hyndman e Athanasopoulos (2014) e Lloyd (2014), que também utilizaram o algoritmo *Gradient Boosting*. No entanto, a estratégia de agregação *top-bottom*, o método de combinação das estações meteorológicas e o conjunto de variáveis preditoras utilizadas explicam o desempenho obtido, apesar de um modelo de previsão relativamente mais simples.

Com exceção das previsões obtidas pelo modelo *benchmark* da competição, os demais trabalhos não disponibilizaram códigos documentados para efeitos de reprodutibilidade, o que dificultou a utilização de recursos gráficos para comparar o desempenho dos modelos. A Figura 4.3 ilustra os diagramas de dispersão entre as cargas observadas e previstas, utilizando a metodologia proposta neste trabalho (através da abordagem *top-bottom*) e o

modelo de previsão adotado como *benchmark* para a competição *GECOM2012*. Os dados referem-se às previsões para a carga total do sistema.

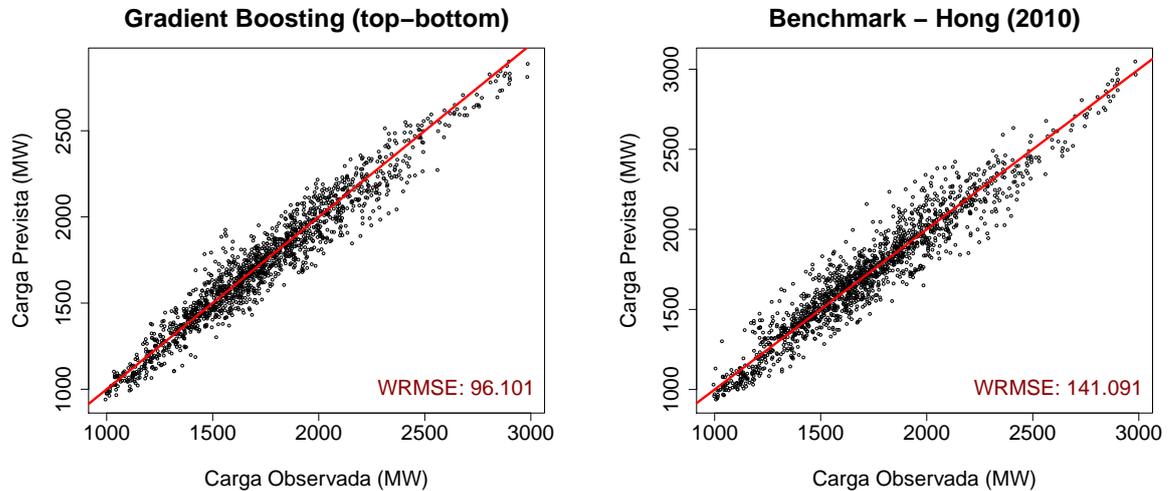


Figura 4.3: Diagrama de dispersão entre a carga prevista e observada para os modelos utilizados neste trabalho.

A previsão das temperaturas para as 11 estações meteorológicas, de acordo com o procedimento descrito por Charlton e Singleton (2014) (seção 7.2.2), é bastante simples e foi executada em menos de 4 segundos. A determinação da ponderação ótima para a temperatura virtual  $T_v(t)$  (seção 7.2.3) para todas as 21 séries de carga foi executada em aproximadamente 15 minutos. O processamento do algoritmo *Gradient Boosting* para todas as zonas de consumo foi concluído em aproximadamente 8 horas. Todas essas etapas foram executadas sequencialmente na mesma máquina descrita na seção anterior.

## 5 Conclusão

Neste trabalho avaliamos a aplicabilidade do algoritmo *Gradient Boosting* com árvores de decisão em problemas de previsão de carga elétrica a curto prazo. Foram consideradas duas bases de dados reais: nos dados da base *Rio-96/97* o algoritmo *Gradient Boosting* foi utilizado para efetuar correções nas previsões obtidas inicialmente através do método de amortecimento exponencial com dupla sazonalidade (DSHW); já nos dados da competição *GEFCom2012* o algoritmo *Gradient Boosting* foi utilizado para prever diretamente os valores de carga elétrica.

O algoritmo *Gradient Boosting* compreende um processo iterativo para formar comitês de previsores. A cada iteração, um novo modelo é treinado e adicionado ao comitê com o objetivo de reduzir os erros de aproximação obtidos na iteração anterior. Portanto, o processo de correção de resíduos, proposto na seção 3.5.1 para a base *Rio-96/97*, pode ser considerado uma adaptação do algoritmo *Gradient Boosting* original para problemas de previsão de séries temporais, em que o primeiro membro do comitê consiste em um modelo de amortecimento exponencial e o restante é formado por árvores de decisão.

A estratégia de correção de resíduos com o método de *boosting* proporcionou um ganho considerável no desempenho de previsão obtido através da técnica de amortecimento exponencial DSHW. Além disso, o modelo proposto também superou os resultados alcançados por Neto (2014), que combinou o mesmo método de amortecimento exponencial com modelos de regressão linear e redes neurais.

A competição *GEFCom2012* abordou o problema de previsão hierárquica de carga elétrica e evidenciou algumas questões que normalmente são negligenciadas em trabalhos específicos sobre previsão de carga, como a seleção das estações meteorológicas, e a agregação das previsões para os diferentes níveis hierárquicos (obedecendo à restrição de que a previsão para um nível superior deve corresponder a soma das previsões dos níveis inferiores). Neste sentido, este trabalho apresentou novas contribuições, com a proposta de combinação das temperaturas a partir do ajuste ótimo à um polinômio de 3º grau (seção 3.5.2.3); e com o método de agregação *top-bottom* discutido na seção 3.5.2.5. Estas duas propostas são essencialmente diferentes das soluções adotadas nos demais trabalhos que foram revisados. A aplicação do algoritmo *Gradient Boosting* em combinação com as

demais estratégias descritas na seção 3.5.2 demonstrou resultados bastante competitivos com os trabalhos mais bem sucedidos na competição *GEFCom2012*.

Uma das dificuldades de aplicação do algoritmo *Gradient Boosting* é a quantidade de parâmetros que devem ser ajustados; contudo, este algoritmo tem recebido grande atenção em publicações recentes, em função de sua flexibilidade e excelente poder preditivo. Os resultados obtidos neste trabalho demonstram que a estratégia utilizada é uma alternativa bastante promissora para aplicação em problemas de previsão de carga elétrica.

No entanto, as metodologias propostas para as duas bases de dados ainda podem ser aperfeiçoadas em trabalhos futuros. Na base de dados *Rio-96/97*, por exemplo, o modelo de previsão foi ajustado com base nos dados do período de treinamento e permaneceu inalterado durante todo o período de teste. Neste sentido, pode-se estudar uma estratégia adaptativa para atualizar o modelo de previsão à medida que novas informações são disponibilizadas. Esta estratégia se justifica a longo prazo, pois o padrão de consumo de energia da população pode ser variante no tempo. Por exemplo, se o número de residências equipadas com equipamentos de refrigeração de ar tende a crescer ao longo do tempo, possivelmente haverá uma mudança na relação entre a carga e a temperatura. Além disso, as tarifas de energia podem variar bastante ao longo do tempo, com impacto direto nos hábitos de consumo da população. Em alguns países, por exemplo, as tarifas de energia variam ao longo do dia, para evitar picos de carga muito elevados. Estas informações, se disponíveis, poderiam ser incluídas como variáveis preditoras no algoritmo *Gradient Boosting*.

No caso da competição *GEFCom2012*, o processo de ajuste dos modelos de previsão poderia ser facilmente paralelizado, uma vez que a carga de cada zona de consumo é prevista de forma independente. Esta estratégia poderia viabilizar a combinação do método de amortecimento exponencial com o processo de correção de resíduos, da mesma forma como fizemos para a base de dados *Rio-96/97*. Além disso, os parâmetros do algoritmo *Gradient Boosting* foram os mesmos em todas as séries de carga, mas poderiam ser definidos separadamente para cada zona de consumo.

O método de previsão de temperaturas também pode ser aprimorado. Utilizamos uma abordagem bastante simplista, descrita por Charlton e Singleton (2014). As previsões calculadas inicialmente foram mantidas durante todo o processo de validação dos modelos de previsão de carga. Contudo, é possível implementar uma metodologia mais

robusta para executar o processo de validação cruzada experimentando, simultaneamente, diferentes modelos de previsão para temperaturas e cargas.

---

Versões parciais deste trabalho foram submetidas e aceitas para apresentação oral e publicação nos anais do IEEE LA-CCI 2016 (*Latin American Conference on Computational Intelligence*), e do XXIII SIMPEP (Simpósio de Engenharia de produção).

# Referências Bibliográficas

- ABRADEE, 2015. Entenda a indústria de energia elétrica. Associação Brasileira de Distribuidores de Energia Elétrica. Disponível em: <http://www.abradee.com.br/escolha-abradee-para-voce/cartilha>. Acessado em: 09/06/2016.
- Alpaydin, E., 2010. Introduction to Machine Learning, segunda Edição. The MIT Press.
- Amjady, N., Keynia, F., Zareipour, H., 2010. Short-term load forecast of microgrids by a new bilevel prediction strategy. Smart Grid, IEEE Transactions on 1 (3), 286–294.
- Baker, A., 1985. Load Forecasting for Scheduling Generation on a Large Interconnected System. In: Bunn, D., Farmer, E. (Eds.), Comparative Models for Electrical Load Forecasting. John Wiley & Sons Ltd, 1985.
- Ben Taieb, S., Hyndman, R. J., 2014. A gradient boosting approach to the Kaggle load forecasting competition. International journal of forecasting 30 (2), 382–394.
- Bergmeir, C., Hyndman, R. J., Koo, B., *et al.*, 2015. A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction. Monash University Department of Econometrics and Business Statistics Working Paper 10, 15.
- Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.
- Breiman, L., Friedman, J., Stone, C. J., Olshen, R., 1984. Classification and Regression Trees. Chapman and Hall.
- Bunn, D., Farmer, E., 1985. Economic and Operational Context of Electric Load Forecasting. In: Bunn, D., Farmer, E. (Eds.), Comparative Models for Electrical Load Forecasting. John Wiley & Sons Ltd. 1985.
- Byrd, R. H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing 16 (5), 1190–1208.

- Carneiro, A. C. M. S., 2014. Previsão do consumo de energia elétrica a curto prazo, usando combinações de métodos univariados. 2014. Dissertação (Mestrado em Modelagem Computacional) - Universidade Federal de Juiz de Fora. Juiz de Fora. 2014.
- Charlton, N., Singleton, C., 2014. A refined parametric model for short term load forecasting. *International Journal of Forecasting* 30 (2), 364–368.
- Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. arXiv preprint arXiv:1603.02754.
- Chitsaz, H., Shaker, H., Zareipour, H., Wood, D., Amjady, N., 2015. Short-term electricity load forecasting of buildings in microgrids. *Energy and Buildings* 99, 50–60.
- Crone, S. F., Lessmann, S., Stahlbock, R., 2006. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research* 173 (3), 781–800.
- Dietterich, T. G., Ensemble methods in machine learning. In: *International workshop on multiple classifier systems*, 2000. Springer, pp. 1–15.
- Feinberg, E. A., Genethliou, D., 2005. Load Forecasting. In: Chow, J. H., Wu, F. F., Momoh, J. A. (Eds.), *Applied Mathematics for Restructured Electric Power Systems*. Springer, Cap. 12, pp. 269–285.
- Fitkov-Norris, E., Vahid, S., Hand, C., Evaluating the impact of categorical data encoding and scaling on neural network classification performance: the case of repeat consumption of identical cultural goods. In: *International Conference on Engineering Applications of Neural Networks*, 2012. Springer, pp. 343–352.
- Freund, Y., Schapire, R., Abe, N., 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence* 14 (771-780), 1612.
- Freund, Y., Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55 (1), 119–139.
- Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

- Friedman, J. H., 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38 (4), 367–378.
- Gneiting, T., Balabdaoui, F., Raftery, A. E., 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (2), 243–268.
- Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application* 1, 125–151.
- Hagan, M. T., Behr, S. M., 1987. The time series approach to short term load forecasting. *IEEE Transactions on Power Systems* 2 (3), 785–791.
- Han, X., Han, L., Gooi, H., Pan, Z., 2012. Ultra-short-term multi-node load forecasting-a composite approach. *IET generation, transmission & distribution* 6 (5), 436–444.
- Hastie, T. J., Tibshirani, R. J., Friedman, J. H., 2009. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York”.
- Hippert, H. S., 2001. *Previsão de cargas a curto prazo: uma avaliação da viabilidade do uso de redes neurais artificiais*. 2001. Tese (Doutorado Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro. 2010.
- Hippert, H. S., Pedreira, C. E., Souza, R. C., 2001. Neural networks for short-term load forecasting: A review and evaluation. *Power Systems, IEEE Transactions on* 16 (1), 44–55.
- Hobbs, B. F., Jitprapaikularn, S., Konda, S., Chankong, V., Loparo, K. A., Maratukulam, D. J., 1999. Analysis of the value for unit commitment of improved load forecasts. *Power Systems, IEEE Transactions on* 14 (4), 1342–1348.
- Hong, T., 2010. *Short term electric load forecasting*. 2010. Tese (Ph.D. Operations Research and Electrical Engineering) - North Carolina State University. Raleigh. 2014.
- Hong, T., Fan, S., 2016. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*.

- Hong, T., Pinson, P., Fan, S., 2014. Global energy forecasting competition 2012. *International Journal of Forecasting* 30 (2), 357–363.
- Hong, T., Wang, P., White, L., 2015. Weather station selection for electric load forecasting. *International Journal of Forecasting* 31 (2), 286–295.
- Hong, T., Wang, P., Willis, H. L., A Naïve multiple linear regression benchmark for short term load forecasting. In: *Power and Energy Society General Meeting, 2011 IEEE*, 2011. IEEE, pp. 1–6.
- Huck, G., Mahmoud, A., Comerford, R., J., A., E., D., 1980. Load Forecast Bibliography. *IEEE Transactions on Power Apparatus and Systems*.
- Hyndman, R. J., Athanasopoulos, G., 2014. *Forecasting: principles and practice*. OTexts.
- Javed, F., Arshad, N., Wallin, F., Vassileva, I., Dahlquist, E., 2012. Forecasting for demand response in smart grids: An analysis on use of anthropologic and structural data and short term multiple loads forecasting. *Applied Energy* 96, 150–160.
- Kotsiantis, S. B., 2013. Decision trees: a recent overview. *Artificial Intelligence Review* 39 (4), 261–283.
- Lloyd, J. R., 2014. GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *International Journal of Forecasting* 30 (2), 369–374.
- Loh, W.-Y., 2011. *Classification and regression trees*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (1), 14–23.
- Makridakis, S., Wheelwright, S. C., Hyndman, R. J., 2008. *Forecasting methods and applications*. John Wiley & Sons.
- Morgan, J. N., Sonquist, J. A., 1963. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association* 58 (302), 415–434.
- Nedellec, R., Cugliari, J., Goude, Y., 2014. GEFCom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting* 30, 375–381.

- Neto, G. G., 2014. Métodos Univariados e Multivariados para Previsão da Demanda de Energia Elétrica em Curto Prazo: Um Estudo Comparativo. 2014. Dissertação (Mestrado em Modelagem Computacional) - Universidade Federal de Juiz de Fora. Juiz de Fora. 2014.
- NIST, 2012. Engineering Statistics Handbook. National Institute Of Standards and Technology. Disponível em: <http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm>. Acessado em: 18/05/2016.
- Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. Vol. 1. Morgan Kaufmann.
- Ranaweera, D., Hubele, N., Karady, G., 1996. Fuzzy logic for short term load forecasting. International journal of electrical power & energy systems 18 (4), 215–222.
- Ridgeway, G., 2007. Generalized Boosted Models: a guide to the gbm package. Disponível em: <http://www.saedsayad.com/docs/gbm2.pdf>. Acessado em: 19/06/2016.
- Ripley, B. D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.
- Russell, S. J., Norvig, P., 2003. Artificial intelligence: a modern approach, segunda Edição. Prentice Hall.
- Taylor, J. W., 2003. Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society 54 (8), 799–805.
- Taylor, J. W., De Menezes, L. M., McSharry, P. E., 2006. A comparison of univariate methods for forecasting electricity demand up to a day ahead. International Journal of Forecasting 22 (1), 1–16.
- Wei, W. W., 2006. Time Series Analysis - Univariate and Multivariate Methods, Second edition.
- Xu, Q.-S., Liang, Y.-Z., 2001. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems 56 (1), 1–11.
- Zhou, Z.-H., 2012. Ensemble methods: foundations and algorithms. CRC press.

# APÊNDICE A - Análise

## Exploratória de Dados

Neste apêndice apresentamos uma análise estatística detalhada sobre as duas bases de dados utilizadas neste trabalho. A análise exploratória de dados permite identificar as características gerais do problema e suas eventuais particularidades. Estas informações podem ser muito úteis para direcionar a construção de modelos de previsão otimizados para cada tipo de problema.

Uma parte significativa da metodologia apresentada neste apêndice foi inspirada no trabalho de Ben Taieb e Hyndman (2014).

### A.1 *Rio-96/97*

A Figura A.1 demonstra o nível médio da curva de cargas ao longo do ano<sup>1</sup>. Podemos notar que o período de maior consumo ocorre entre os meses de dezembro a março. Este período corresponde ao verão do hemisfério sul, quando as temperaturas médias são mais elevadas e utilização de equipamentos de refrigeração de ambientes é maior. O consumo de energia é mais baixo nos meses de inverno, indicando que as temperaturas não chegam a níveis tão baixos a ponto de demandar o uso de equipamentos de aquecimento. Este comportamento caracteriza as regiões tropicais localizadas no hemisfério sul.

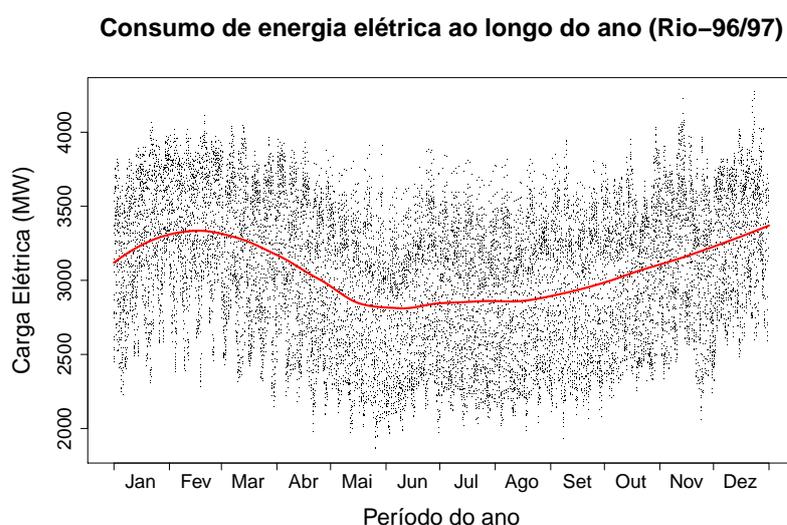


Figura A.1: Consumo de energia elétrica ao longo do ano (*Rio-96/97*).

<sup>1</sup>A suavização representada na linha vermelha foi obtida a partir da técnica LOESS, definindo o parâmetro de amortecimento  $q = 0,3$ , conforme sugere NIST (2012)

O consumo de carga varia bastante entre os dias úteis e finais de semana, conforme podemos observar no *boxplot* da Figura A.2. A demanda é maior nos dias úteis, quando os valores observados possuem uma distribuição muito semelhante. O domingo registra as menores demandas, uma vez que a maior parte das atividades econômicas não funciona neste dia. Em geral, os setores de comércio e serviços funcionam aos sábados no período da manhã ou até o início da tarde, portanto o sábado apresenta um comportamento intermediário entre os dias úteis e o domingo.

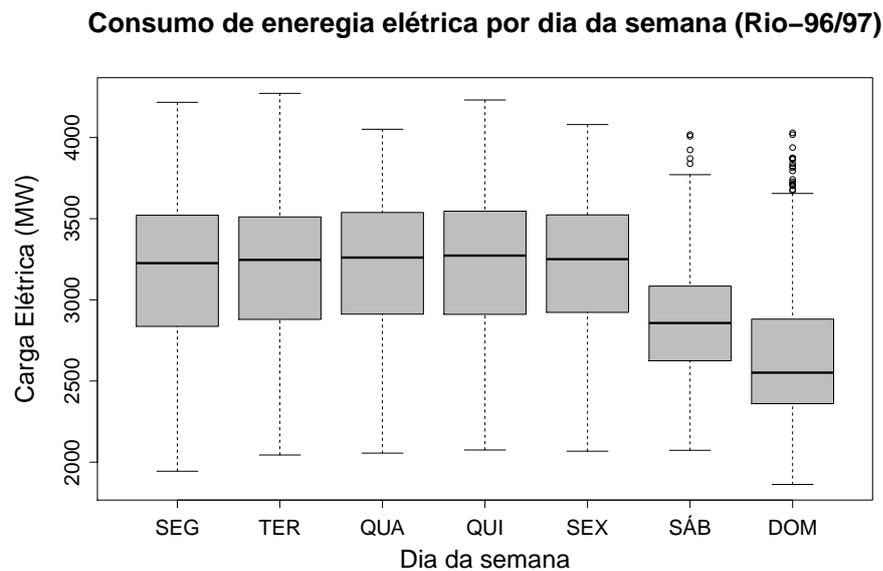


Figura A.2: *Boxplot* do consumo de energia por dia da semana (Rio-96/97).

A Figura A.3 demonstra que o consumo de energia elétrica também varia bastante ao longo do dia (a figura considera somente os dias úteis). O maior consumo é registrado entre 10:00 às 21:00, período que corresponde à atividade comercial e inclui o horário de pico, que ocorre por volta das 20:00, quando uma grande parte da população costuma utilizar o chuveiro elétrico. As cargas mais baixas são observadas durante a noite, quando a maior parte dos equipamentos elétricos está desligada. No entanto, o período noturno registra uma dispersão mais alta em torno dos valores medianos, o que pode ser explicado pela utilização de equipamentos de ar-condicionado em noites muito quentes, visto que as distribuições apresentam uma assimetria negativa.

A correlação entre a temperatura e a carga possui um comportamento diferente, de acordo com o horário do dia, conforme demonstra a Figura A.4. No período noturno, podemos notar uma correlação positiva bastante evidente. Já no intervalo de 16:00 às 20:00, por exemplo, a relação entre as duas variáveis é muito fraca.

### Consumo de energia elétrica por horário (Seg-Sex)

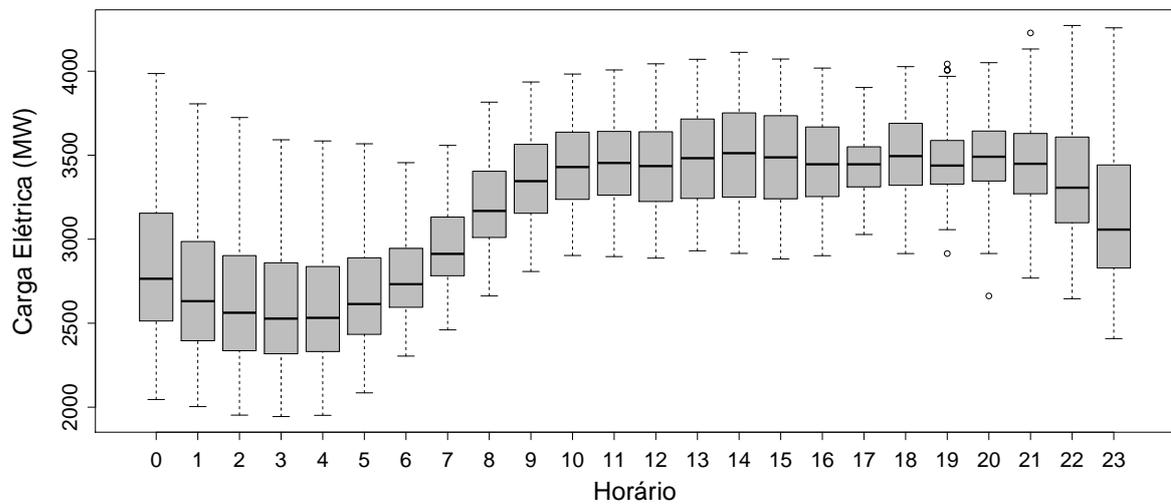


Figura A.3: *Boxplot* do consumo de energia por horário do dia (*Rio-96/97*, seg-sex).

### Temperatura x Carga elétrica (Seg-Sex)

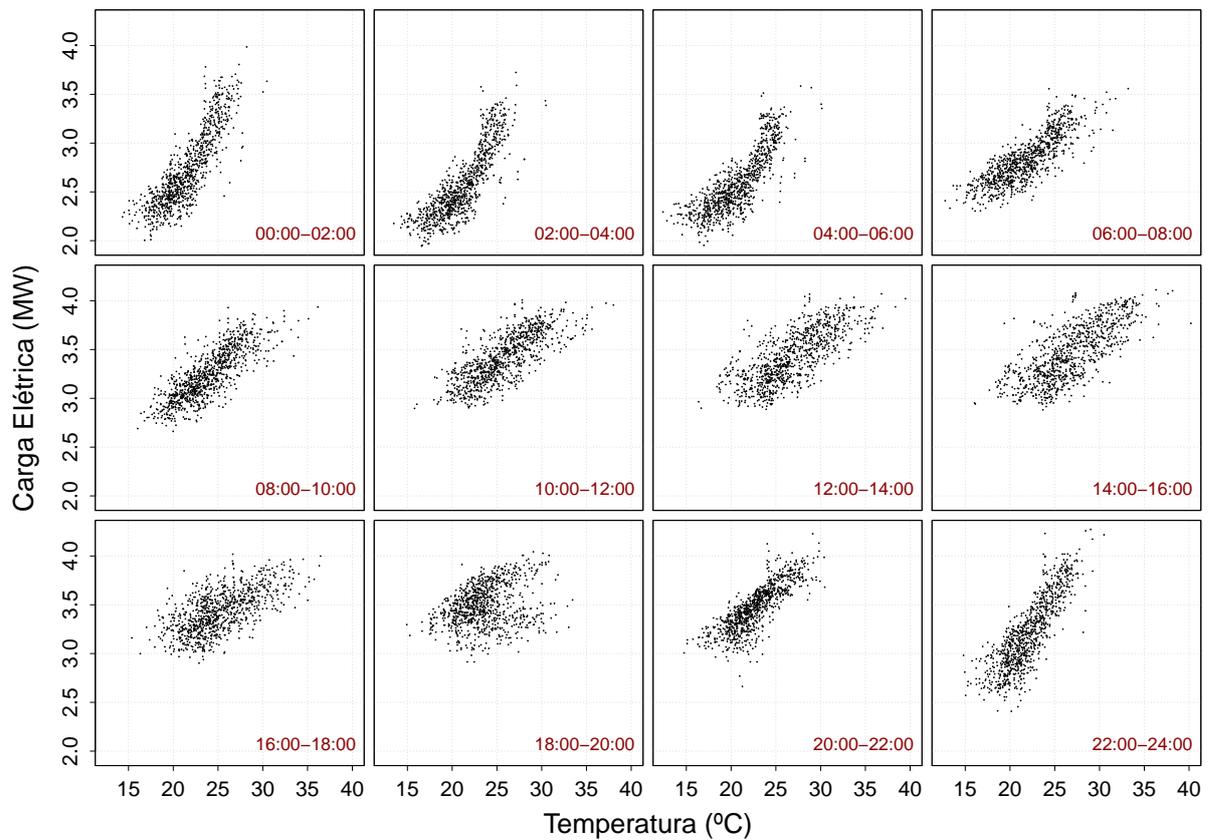


Figura A.4: Diagrama de dispersão 'carga x temperatura' por faixa de horário (*Rio-96/97*, seg-sex).

## A.2 *GEFCom2012*

A base de dados *GEFCom2012* compreende um sistema elétrico formado por 20 zonas de consumo. O gráfico de Pareto, ilustrado na Figura A.5, revela que a participação de cada zona no consumo total do sistema é bastante desigual. As zonas 18, 3, 7, 6 e 2 apresentam as maiores cargas e, juntas, respondem por mais de 50% do consumo total do sistema. A zona 4, por outro lado, tem uma participação praticamente desprezível dentro do sistema como um todo.

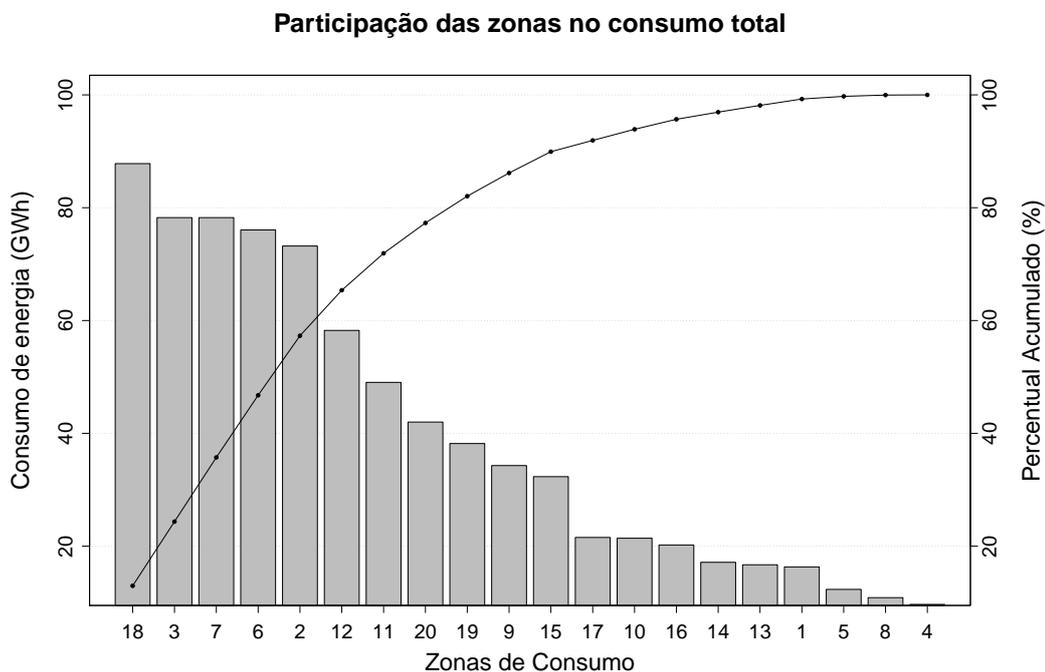


Figura A.5: Participação de cada zona geográfica no consumo total do sistema durante todo o período observado (*GEFCom2012*).

O nível médio da carga ao longo do ano nos dados da base *GEFCom2012* comporta-se de forma muito diferente dos dados da base *Rio-96/97*. Observando a Figura A.6 podemos notar dois picos de demanda durante o ano. O pico entre os meses de janeiro e fevereiro corresponde ao período mais rigoroso do inverno no hemisfério norte, quando a população precisa recorrer aos aquecedores elétricos. O aumento do consumo entre os meses de junho e agosto é referente à utilização de equipamentos de ar refrigerado durante o verão no hemisfério norte. Este comportamento caracteriza regiões de clima temperado.

Evidentemente, a intensidade da relação entre o consumo de energia elétrica com a temperatura pode depender de fatores sociais e econômicos, como por exemplo, a

### Consumo de energia elétrica ao longo do ano (GEFCom2012)

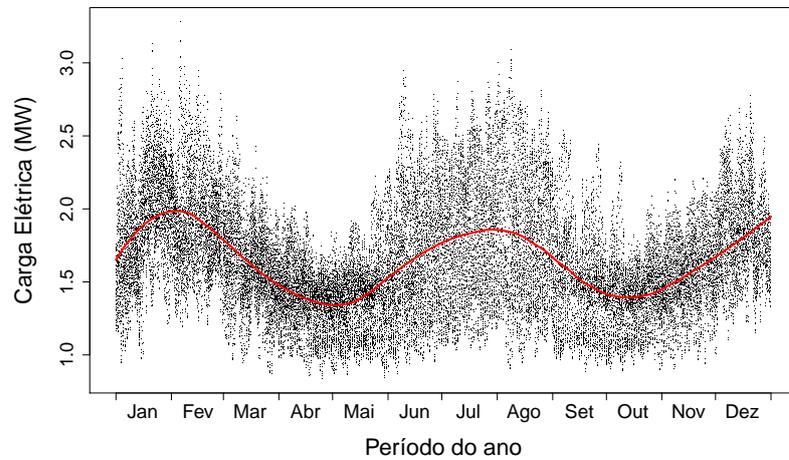


Figura A.6: Consumo de energia elétrica ao longo do ano (*GEFCom2012* - carga total do sistema)<sup>2</sup>.

proporção de residências equipadas com equipamentos de refrigeração e aquecimento, e o preço da energia elétrica. Este último pode sofrer variações sazonais, principalmente se a geração de energia provém de fontes renováveis que dependem de recursos naturais, como o regime de chuvas, ventos ou da intensidade dos raios solares.

A Figura A.7 demonstra que a distribuição das cargas nos dias úteis e finais de semana é muito semelhante. Nota-se apenas que os valores medianos registrados no sábado e domingo são ligeiramente inferiores em relação aos dias de úteis.

<sup>2</sup>Suavização realizada de forma análoga à Figura A.1

### Consumo de energia elétrica por dia da semana (GEFCom2012)

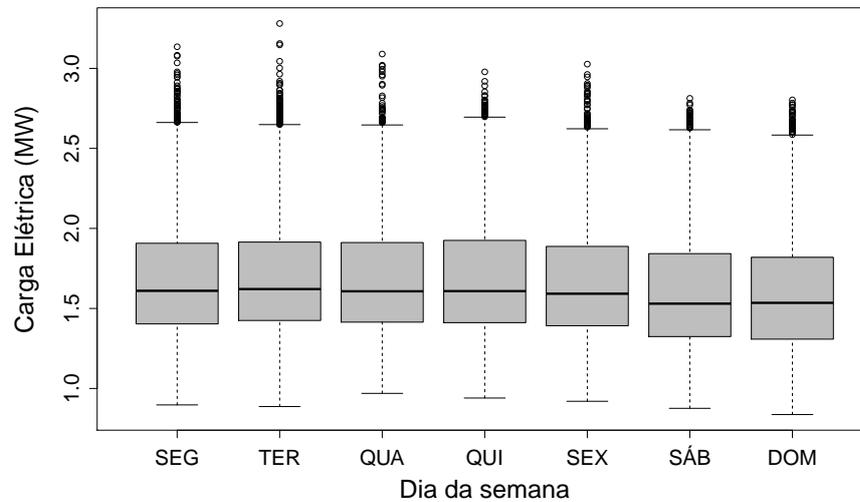


Figura A.7: Consumo de energia elétrica ao longo do ano (*GEFCom2012* - carga total do sistema).

O padrão de consumo de energia ao longo do dia nos dados da competição *GEFCom2012* também é diferente do comportamento observado nos dados da base *Rio-96/97*. Nesta série, verificamos um consumo mediano bastante estável durante o horário comercial de 08:00 às 16:00 e o pico de carga ocorre por volta das 19:00.

### Consumo de energia elétrica por horário (Seg-Sex)

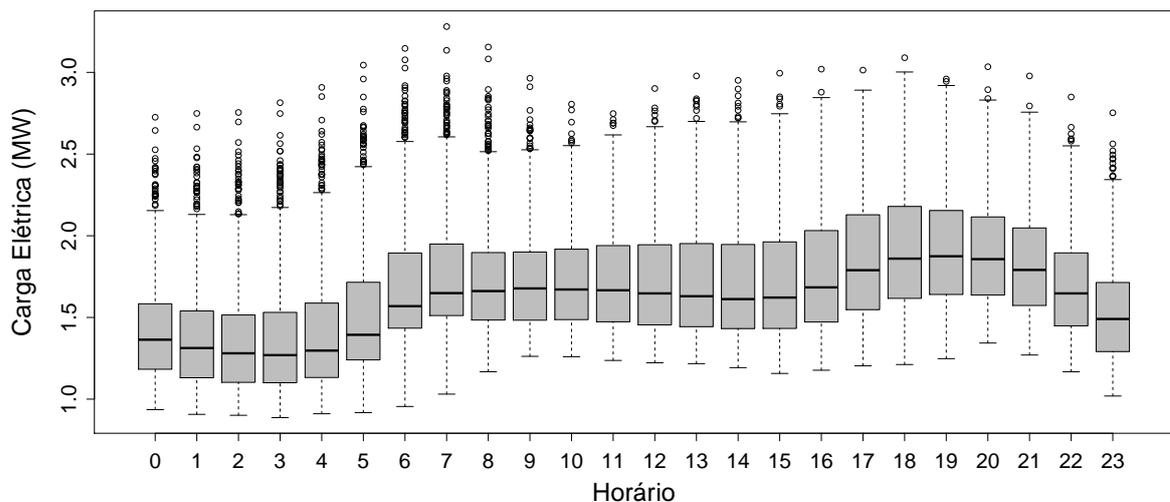


Figura A.8: *Boxplot* do consumo de energia por horário do dia (*GEFCom2012* - carga total do sistema, seg-sex).

Assim como fizemos para a base de dados *Rio-96/97*, a Figura A.9 representa a

correlação entre a carga elétrica e a temperatura em diferentes períodos do dia. Neste caso, podemos notar uma correlação não-linear muito clara. A análise gráfica sugere que existe uma temperatura de conforto, na qual o consumo de energia é mínimo; a carga aumenta à medida que a temperatura se afasta da região de conforto, pois os consumidores recorrem aos equipamentos de aquecimento ou refrigeração.

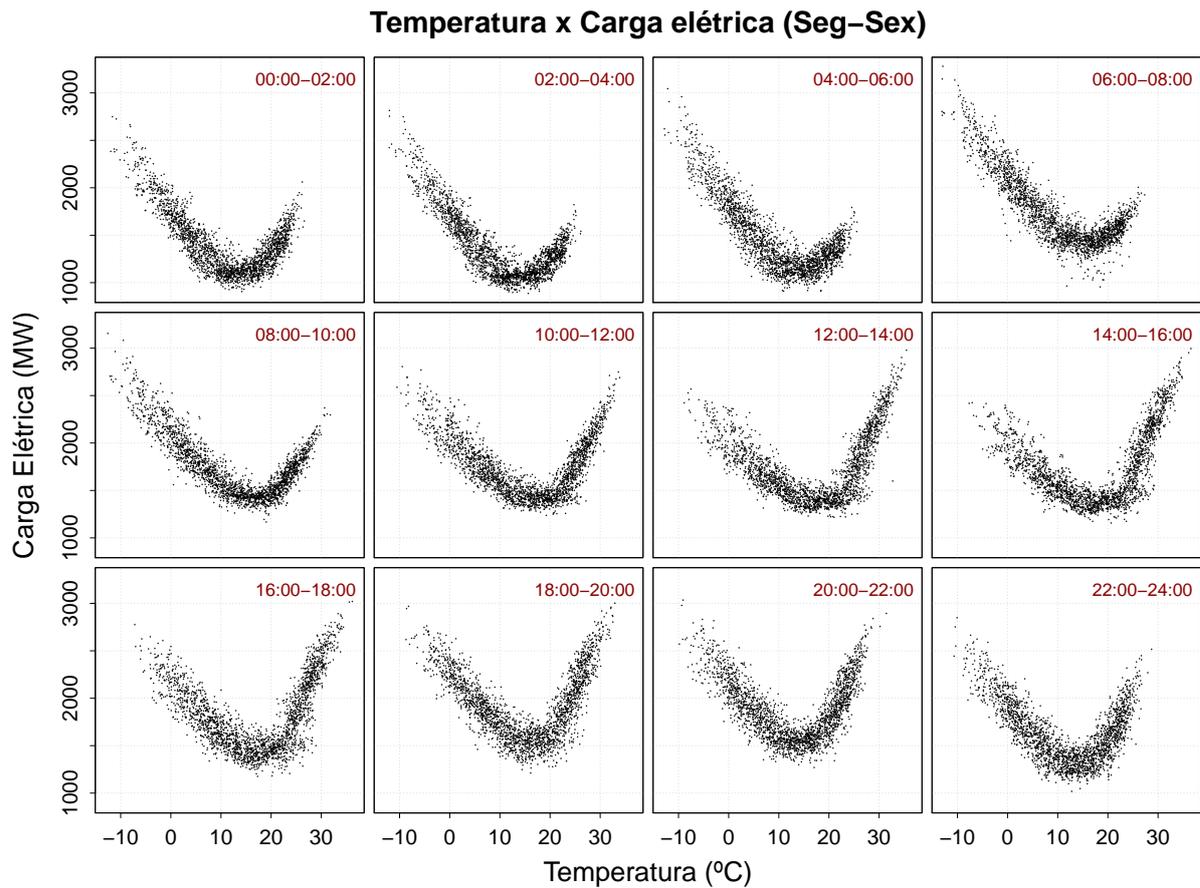


Figura A.9: Diagrama de dispersão ‘carga x temperatura’ por faixa de horário (*GEFCom2012* - carga total do sistema, seg-sex).