UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CAMPUS DE FOZ DO IGUAÇU

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE SISTEMAS DINÂMICOS E ENERGÉTICOS

DISSERTAÇÃO DE MESTRADO

# AGRUPAMENTO DE CURVAS DE CARGA PARA REDUÇÃO DE BASES DE DADOS UTILIZADAS NA PREVISÃO DE CARGA DE CURTO PRAZO

MARCOS RICARDO MÜLLER

FOZ DO IGUAÇU

2014

Marcos Ricardo Müller

# Agrupamento de Curvas de Carga para Redução de Bases de Dados Utilizadas na Previsão de Carga de Curto Prazo

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos como parte dos requisitos para obtenção do título de Mestre em Engenharia de Siste-mas Dinâmicos e Energéticos. Área de concentração: Sistemas Dinâmicos e Energéticos.

Orientador: Edgar Manuel Carreño Franco

Foz do Iguaçu

2014

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca do Campus de Foz do Iguaçu – Unioeste
Ficha catalográfica elaborada por Miriam Fenner R. Lucas - CRB-9/268

M958 Müller, Marcos Ricardo
Agrupamento de curvas de carga para redução de bases de dados utilizadas na previsão de carga de curto prazo / Marcos Ricardo Müller. – Foz do Iguaçu, 2014.
78 p. :tab. : graf.

Orientador: Prof. Dr. Edgar Manuel Carreño Franco.
Dissertação (Mestrado) – Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos - Universidade Estadual do Oeste do Paraná.

1. Sistemas de energia elétrica – Previsão de carga. 2. Método de dias similares. 3. Clausterização de dados. 4. Medidores eletrônicos. I. Título.

CDU 621.317

# Agrupamento de Curvas de Carga para Redução de Bases de Dados Utilizadas na Previsão de Carga de Curto Prazo

Marcos Ricardo Müller

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos e aprovada pela Banca Examinadora:
Data da defesa pública: 21/02/2014.

Prof. Dr. **Dr. Edgar Manuel Carreño Franco** - (Orientador)
Universidade Estadual do Oeste do Paraná - UNIOESTE

Profa. Dra. **Mara Lúcia Martins Lopes**
Universidade Estadual Paulista "Júlio de Mesquita Filho" - UNESP

Profa. Dra. **Huei Diana Lee**
Universidade Estadual do Oeste do Paraná - UNIOESTE

# Resumo

Este trabalho apresenta a utilização de clusterização de curvas de carga do nível menos agregado para o método de dias similares, com o objetivo de obter conjuntos reduzidos de dados que imponham menores cargas computacionais ao algoritmo de previsão, e permitir ainda, desempenhos similares ou superiores quando comparados aos obtidos pelo método de dias similares que faz uso do conjunto original de dados. O método de dias similares permite realizar previsão de carga de curtíssimo prazo a partir de dados históricos de consumo de energia elétrica, além de dados correlatos, que permitem traçar analogias com um dia futuro. Implementações convencionais do mesmo método são utilizadas para comparação de resultados. O cenário que fornece os dados para os estudos, assim como os equipamentos empregados e a etapa de pré-processamento de dados são apresentadas. A análise de silhuetas de cluster foi empregada com o objetivo de validar os agrupamentos. Por meio do cálculo do MAPE foi possível verificar a assertividade das previsões, indicando superioridade daquela baseada nas curvas de carga clusterizadas.

**Palavras-chave:** Método de Dias Similares; Previsão de Carga; Clusterização de Dados; Medidores Eletrônicos.

# Abstract

This work presents the use of clustering techniques in load curves for the similar days method for load forecasting, in order to obtain a reduced data to achieve a faster computational algorithm, while achieving similar or superior performance compared to those obtained by the traditional method that makes use of the original data set. The method allows to perform similar day load forecasting using short-term historical data from the consumption of electricity at consumers level, and related data, which allow tracing analogies to a future day. Conventional implementations of the method are used for comparison and validation. The scenario that provides the data for the studies, as well as the equipment, and data preprocessing stage, are presented. The methodology is validated using the cluster silhoute analysis. With the MAPE values was possible to verify the forecast, indicating superiority of the method based on clustered load curves.

**Keywords:** Similar Days Method; Load Forecasting; Clustering Data; Electronic Meters.

Dedico este trabalho a todos os meus amores.

# Agradecimentos

Agradeço ao Mestre dos mestres, Rei dos reis, detentor da verdade absoluta, Sr. Deus do universo.

Agradeço a minha família por todo o aprendizado, e por tantos momentos felizes e inesquecíveis compartilhados. Em especial gostaria de agradecer a minha mãe, Sra. Marinês Mânica Müller, por todo o amor, amparo e exemplo que tem sido.

Gostaria de agradecer a todas as pessoas que amo e já não estão nesse mundo físico, pois assim como a alma, o amor é imortal.

Thay e Pri, obrigado por vocês existirem na minha vida!

Agradeço a Profa. Huei Diana Lee por toda a inspiração e ajuda nesses anos de mestrado.

Agradeço ao Prof. Edgar Manuel Carreño Franco pela orientação, por ter compartilhado tantos conhecimentos e ter sido exemplo de dedicação ao ensino e a pesquisa.

Agradeço a todos os colegas, amigos e professores do PGESDE, por terem sido parte importante de minha vida nesses anos de mestrado. Devo agradecer em especial a Katiani por todas as horas de estudos compartilhadas, e auxilio oferecido.

A Fundação Parque Tecnológico Itaipu – FPTI, pelo apoio financeiro concedido durante a realização da pesquisa.

Por fim, agradeço a todos que contribuíram, de maneira direta ou indireta, ostensiva ou não, explícita ou implícita, formal ou informal, para que este trabalho pudesse ter chegado ao seu fim a contento.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Símbolos

| | |
|---|---|
| *AG* | Algoritmo Genético |
| *ANEEL* | Agência Nacional de Energia Elétrica |
| *CNAE* | Classificação Nacional de Atividades Econômicas |
| *DE* | Distância Euclidiana |
| *DTW* | *Dynamic Time-Warping* |
| *EE* | Energia Elétrica |
| *EPE* | Empresa de Pesquisa Energética |
| *GPL* | *General Public License* |
| *IC* | Intervalo de Confiança |
| *MAPE* | *Mean Absolute Percentage Error* |
| *MDS* | Método de Dias Similares |
| *MFT* | Modelo de Função de Transferência |
| *PTI* | Parque Tecnológico Itaipu |
| *SC* | Silhueta de Cluster |
| *SPK* | *Service Pack* |
| *ST* | Série Temporal |
| *STP* | *Shielded Twisted Pair* |
| *TM* | *Trade Mark* |
| *Weka* | *Waikato Environment for Knowledge Analysis* |

xx

# Capítulo 1

# Introdução

A perspectiva de aumentar lucros e evitar prejuízos estimula empresas de comercialização de energia elétrica (EE) a investirem em ferramentas de previsão de demanda.

Ao passo que dados estejam disponíveis, diversos estudos podem ser conduzidos com o intuito de criar métodos que permitam produzir estimativas para um determinado horizonte futuro.

Dispor de adequadas estimativas para a demanda de EE pode significar mais lucros para as comercializadoras e distribuidoras de EE, uma vez que o ganho está atrelado ao cumprimento dos limites contratados.

Evitar compras emergenciais de energia de outras distribuidoras, ou o pagamento de multas devido ao descumprimento de obrigações inerentes ao serviço de fornecimento de EE, também constitui o objetivo das empresas de energia.

Comumente dividem-se as previsões de acordo com o horizonte de previsão que abrangem, inicialmente têm-se as de curtíssimo/curto prazo, que abrangem horizontes de algumas horas a poucos dias, as de médio prazo que envolvem períodos de meses a um ano, e por fim as de longo prazo, projetando a demanda para até o limiar de uma década (Leone, 2006).

As previsões de curtíssimo prazo, objeto do presente trabalho, são tidas como fundamentais para orientar o planejamento da operação do sistema, transferências de energia e o gerenciamento da demanda (Rahman e Hazim, 1993).

Obter dados de consumo de EE é fundamental para suprir metodologias de previsão de carga. A aquisição desses dados atualmente é facilitada graças ao uso de modernos medidores digitais. Esses medidores permitem capturar dados de consumo em intervalos pré-definidos, possibilitando ainda que esses sejam transmitidos por uma rede de dados até um servidor de armazenamento.

Com o surgimento das redes inteligentes e a intensificação dos estudos nessa área, é esperado um aumento significativo na aquisição e no armazenamento de dados, trazendo novas

possibilidades e desafios para área, como previsão em níveis cada vez mais desagregados.

No passado não se verificava a atual disponibilidade de dados de consumo de energia, fazendo com que estudos fossem conduzidos com menores amostras de dados.

Apesar da menor disponibilidade de dados de consumo de EE no passado, os métodos de previsão de carga não constituem algo novo, sendo aplicados desde a década de 80. Tais métodos podem ser divididos em dois grupos, o primeiro formado pelos métodos estatísticos, e o segundo, constituído por métodos que têm suas bases em princípios da inteligência artificial. Existem ainda os métodos híbridos, baseados na combinação de dois ou mais métodos de previsão (Guirelli, 2006).

Entre os métodos clássicos, pode-se destacar: Regressão Linear Múltipla, Modelo Auto-regressivo Integrado de Média Móvel - ARIMA, Alisamento Exponencial e Análise Espectral. Entre os métodos baseados em Inteligência Artificial, podem ser citados os Sistemas Especialistas, as Redes Neurais, os que empregam Lógica Fuzzy e os Algoritmos Genéticos (Guirelli, 2006).

O método de dias similares – MDS é um método estatístico que se baseia em dados que alcançam horizontes de um ou mais anos, no qual se procura determinar similaridades entre os dias catalogados com as características conhecidas para um dia futuro (dia da semana, estação do ano e temperatura média prevista, entre outras). Sendo assim, a previsão se dá graças à capacidade do algoritmo em encontrar analogias entre dias passados e futuros (Kadowaki *et al.*, 2004; Senjyu, Higa, e Uezato, 1998).

A abordagem do MDS se destaca por não lidar apenas com a parte não linear da curva de carga, mas também com dias especiais, como os fins de semana e feriados. Esta abordagem também se mostra útil em situações em que os modelos de previsão de carga precisos são difíceis de projetar (Senjyu, Higa e Uezato, 1998).

De maneira simples o método de dias similares pode ser representado por uma tabela atributo-valor, onde são catalogados os dados referentes às curvas diárias de carga e dados externos, como os meteorológicos, servindo de base para consultas parametrizadas, que procuram determinar dias com características similares, que possam compor uma previsão.

O MDS assim como outros métodos de previsão tem como uma de suas fragilidades o tempo computacional, que se eleva de acordo com a quantidade de dados históricos considerados para a previsão. Com base nisso ossível perceber que a redução da base dados utilizada, implica diretamente na redução do tempo computacional.

Com o objetivo de reduzir o conjunto de dados necessário para realizar previsão com o

MDS, este trabalho emprega tarefas de clusterização de dados, possibilitando uma redução de aproximadamente 80% do conjunto original, mantendo índices de assertividade semelhantes e superiores aos obtidos com o conjunto original de curvas de cargas.

Os resultados encontrados a partir das curvas de carga clusterizadas são comparados com outros, obtidos pelo método de dias similares baseado no conjunto original de curvas de carga e com suas médias.

Um diferencial do trabalho é a apresentação de intervalos de confiança para as previsões realizadas, pois tal detalhe agrega informações que permitem realizar análises mais apuradas dos resultados.

Os diferentes usos baseados no conjunto inicial de curvas de carga permitiu suprir resultados que foram mutuamente confrontados. O cálculo do MAPE possibilitou avaliar a assertividade das previsões, apontando superioridade no uso das curvas de carga clusterizadas.

Este trabalho não tem como foco realizar previsão de carga. Ao em vez disso, explora a clusterização de dados, técnica que pode ser empregada com outros distintos métodos de previsão de carga de curto prazo.

## 1.1 Objetivos

Os principais objetivos deste trabalho são apresentados:

*Objetivo Geral:*

- Empregar tarefas de agrupamento de curvas de carga para a redução de bases de dados utilizadas na previsão de carga de curto prazo.

*Objetivos Específicos:*

- Caracterizar a microrrede que supre os dados utilizados no presente trabalho;
- Estudo de métodos de previsão de carga;
- Estudo de medidas de dissimilaridade e de clusterização de dados;
- Estudo compreendendo análise, classificação e tratamento de séries temporais;
- Analisar e preparar os dados que compõem as curvas de carga para que esses possam ser submetidos a tarefas de previsão;
- Codificação do método dias similares;

- Aplicar tarefas de clusterização de dados sobre uma base de curvas de carga de um nível menos agregado;
- Comparar resultados de previsão utilizando o conjunto original de dados e aquele submetido a tarefas de clusterização.

## 1.2 Organização deste Trabalho

**Capítulo 2 – Previsão da demanda**: Apresenta uma introdução sobre a previsão da demanda, principais tipos e modelos de previsão, métodos comuns de previsão, além de caracterizar os principais tipos de carga sobre as quais se realizam tarefas de previsão;

**Capítulo 3 – Séries Temporais**: Este capítulo introduz, descreve e caracteriza as séries temporais. Apresenta os objetivos da análise das séries e os principais desafios e tarefas na mineração desse tipo de dados, com destaque para as tarefas de clusterização. Por fim, são abordadas algumas métricas de dissimilaridade entre séries temporais;

**Capítulo 4 – Cenário de Estudo e Técnicas de Previsão**: Apresenta e caracteriza o cenário que supre os dados utilizados nas tarefas de clusterização e previsão, além de descrever e apresentar os fluxogramas dos algoritmos empregados nas etapas de pré-previsão, previsão e avaliação de resultados;

**Capítulo 5 – Testes e Resultados**: Traz uma breve introdução ao capítulo, descreve o ambiente de desenvolvimento e de testes, apresenta a análise das curvas de carga consideradas no trabalho, apresenta os resultados da validação das clusterizações, descreve os ensaios de previsão e apresenta os resultados obtidos, e em sequencia as considerações finais sobre a tarefa de clusterização e os resultados de previsão;

**Capítulo 6 – Conclusão**: Apresenta as principais conclusões divididas entre os subcapítulos: Principais Contribuições, Limitações e Trabalhos Futuros.

# Capítulo 2

# Previsão da Demanda

## 2.1 Considerações Iniciais

A previsão da demanda tem por objetivo fornecer subsídios para o planejamento da operação e a expansão do sistema elétrico.

O foco das previsões de curto e curtíssimo prazo é permitir o planejamento da operação do sistema, antevendo a necessidade de transferências de energia, e demais manobras que visem evitar a interrupção do fornecimento de EE.

Em um sistema de potência, deve existir um balanceamento entre a energia gerada e a demandada pelos consumidores, ou seja, o sistema deve ser capaz de produzir a mesma quantidade de energia requerida pelos consumidores, considerando ainda, as perdas envolvidas em todo o sistema. Este processo deve ser continuo, de forma a garantir a estabilidade do sistema e o atendimento aos consumidores (Leone, 2006).

Caso a geração exceda a demanda pode ocorrer a sobretensão, já, a geração inferior à demandada, pode causar subtensão no sistema, pois ainda é um desafio armazenar energia em grande quantidade e de forma viável (Miller, 2009). Desta forma, nos processos de operação e controle dos sistemas de potência, é fundamental o acompanhamento da carga, pois atrelado a isso está a necessidade de fornecer energia elétrica de boa qualidade, e ainda, economicamente viável para os envolvidos, para tanto, empresas de energia elétrica precisam dispor de mecanismos que possibilitem a resolução de vários problemas técnicos e operacionais envolvidos (Leone, 2006).

Atualmente existem situações em que o foco da previsão deixa de ser um grande grupo de consumidores, e passa a ser um consumidor específico, com o objetivo, por exemplo, de compreender e balancear as cargas de uma *Smart Microgrid*. As microrredes inteligentes devem ser entendidas como um conceito, passíveis da aplicação de diversas tecnologias de

comunicação de dados, permitindo uma compreensão maior dos estados dos equipamentos da rede, além de permitir atuação sobre os mesmos. Grande parte dos estudos voltados as microrredes tem enfoque na utilização de geração distribuída e seus desafios (Agarwal, Weng, Gupta, 2011; Falcão, 2009).

As curvas de carga representam a evolução do consumo de energia elétrica com base no tempo, de tal maneira que constituem séries temporais, sendo assim, a previsão da demanda pode ser caracterizada como previsão de séries temporais.

## 2.2 Tipos de Previsão em Sistemas de Potência

Os tipos de previsão em sistemas de potência podem ser divididos em três grupos: curto (e/ou curtíssimo), médio e longo prazo:

- **Previsão de Curto Prazo**: neste modelo de previsão, o horizonte de tempo envolvido é bastante curto, sendo de algumas horas a poucos dias. De acordo com Rahman e Hazin (1993), esse tipo de previsão é fundamental para orientar o planejamento da operação, transferência de energia e gerenciamento de demanda, sendo desta forma, fundamental para se planejar de que maneira deve-se conduzir a operação do sistema. Uma boa previsão de curto prazo pode permitir otimizar os recursos de produção, influenciando diretamente na diminuição dos custos de produção de EE.

- **Previsão de Médio Prazo**: compreende horizontes de estimação que se estendem entre meses a um ano. Este tipo de previsão permite que se programe o suprimento de combustível, operações de manutenção e o planejamento de intercâmbio de energia. Caracteriza-se assim, como orientada a otimização dos recursos disponíveis. Um exemplo encontra-se na operação de usinas hidrelétricas, que dependem diretamente dos níveis dos reservatórios hídricos e das afluências que serão recebidas em determinados períodos, de tal maneira, é desejável otimizar a utilização das hidroelétricas em detrimento de termoelétricas, que infundem custos maiores para geração de energia. Estudos que subsidiam informações para o planejamento da operação das hidroelétricas, precisam lidar com incógnitas, que são substituídas por estimativas, entre estas, pode-se destacar: o comportamento do mercado futuro e as afluências.

- **Previsão de Longo Prazo**: neste tipo de previsão o horizonte parte de anos atingindo até uma década. Este tipo de previsão apresenta um princípio distinto das previsões de curto e médio prazo, seu papel é de servir como referência para orientar investimentos no setor energético e decisões comerciais. Quando no contexto da distribuição, esse tipo de previsão também é de grande importância, pois decisões relacionadas à compra e a venda de energia elétrica são tomadas normalmente com prazos de cinco anos de antecedência. Grandes problemas estão envolvidos neste tipo de previsão, como a necessidade de estimar diversos valores, dentre os quais a disponibilidade de geração, preços do petróleo, câmbio e taxa de juros, para um horizonte bastante distante.

## 2.3 Consumidores de Energia Elétrica

No Brasil classifica-se como consumidor qualquer pessoa física ou jurídica que solicite a concessionária o fornecimento de energia elétrica, e assuma a responsabilidade pelo pagamento das faturas e demais obrigações fixadas em regulamentos pela ANEEL (ANEEL, 2012).

Entre os consumidores de EE é possível observar a existência de cargas das mais distintas naturezas, sendo que tipicamente são classificadas entre os tipos: residencial, comercial, industrial, rural, iluminação pública, poder público, serviço público e especiais (Francisquini, 2006).

Conhecer os consumidores de energia elétrica implica em conhecer suas curvas de carga. Por meio da análise das curvas de carga de um consumidor residencial, por exemplo, é possível até mesmo indagar sobre sua classe social, sua rotina e costumes.

Ter acesso e analisar as curvas de carga de uma empresa pode permitir identificar e traçar estratégias importantes, tornando possível compreender o processo produtivo e até mesmo considerar meios de amenizar o consumo em determinados períodos.

### 2.3.1 Tipos de Consumidores

Ao se referir a consumidores de EE é importante ter em mente que consumidores de um mesmo grupo podem possuir perfis de consumo muito distintos, porém, é possível selecionar a partir de uma grande amostra, consumidores que apresentam perfis semelhantes dentro de uma

determinada faixa de variação. É quando se compara consumidores de diferentes grupos que se encontram as maiores dissimilaridades.

Este subcapítulo apresenta sessões dedicadas a apontar algumas características dos principais tipos de consumidores de EE.

#### 2.3.1.1 Consumidor Residencial

É possível observar no setor energético nacional em 2011 a grande participação da classe residencial, respondendo pelo consumo de aproximadamente 25,85% de toda energia demandada, quando observados os consumidores cativos e livres (EPE, 2012).

As cargas dos consumidores residenciais de energia elétrica são tipicamente formadas pelo uso de sistemas de aquecimento de água, climatização, refrigeração, iluminação, ferro elétrico, televisão e outros equipamentos elétricos.

No Brasil o aquecimento de água nas residencias ocorre tipicamente pela utilização de chuveiros elétricos, bastante utilizados no horário de ponta, respondendo por uma parte expressiva do consumo neste período crítico.

A Figura 1 apresenta com base na pesquisa de campo presente no Relatório Brasil - Classe Residencial (2007) a participação típica (%) de determinados equipamento elétricos nas cargas residenciais, os equipamentos que compõem a lista são: chuveiro, equipamentos para condicionamento ambiental, televisor, som, ferro de passar, geladeira, freezer e lâmpadas.

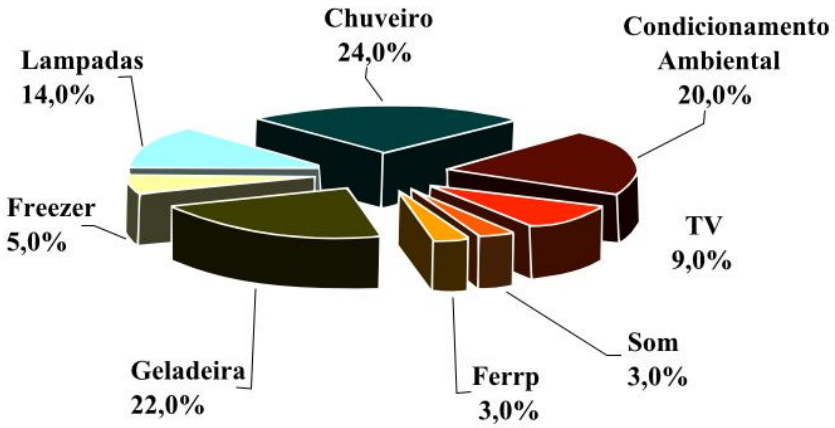



Figura 1 – Influência dos Eletrodomésticos nas Cargas Residenciais (Adaptado de Procel e Eletrobrás, 2007)

É necessário observar que diversos fatores subjetivos moldam o perfil de consumo de cada unidade consumidora. O número de integrantes da família, a quantidade de equipamentos elétricos, o tamanho dos ambientes da residência, o tipo de iluminação empregada e a existência

ou não de sistemas de climatização, juntamente com outros fatores, que abordam desde a qualidade da isolação térmica empregada, o aproveitamento de iluminação e o aquecimento solar entre outros, que acabam por diferenciar o consumo entre unidades consumidoras. Entretanto, os períodos de ocorrência de maior e menor consumo podem se mostrar semelhantes para uma determinada parcela das unidades consumidoras, caracterizando-as.

Além dos fatores econômicos e de consciência ambiental, existem diferentes costumes e rotinas, que acabam por moldar diversos perfis de consumo. A variação na utilização de determinados equipamentos elétricos, mudanças no tempo de banho (com chuveiro elétrico), alterações na rotina e no número de integrantes de uma determinada residência, também contribuem com que as curvas de carga residenciais sejam bastante distintas.

A Figura 2 apresenta a curva de carga real de um consumidor da classe residencial que consome cerca de 330 kWh/mês, na qual é possível observar a ocorrência de um consumo praticamente constante entre 2 e 6 horas, apresentando um crescimento que inicia a partir das 6 horas e atinge seu pico local as 8 horas, voltando a decrescer até as 9 horas, desta última hora até as 17 horas o consumo se mantém estável, apresentando leve aumento e diminuição no consumo entre as horas deste período, a partir das 17 horas se inicia um consumo abrupto, atingindo o pico em torno das 21h, voltando a decrescer até se estabilizar em torno das 2 horas da manhã.

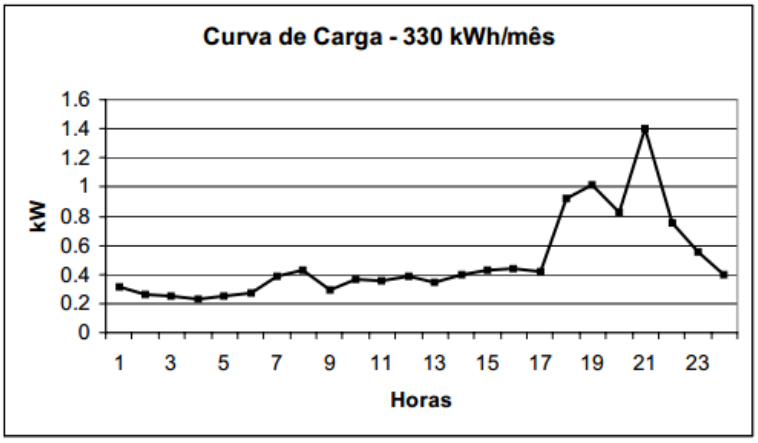



Figura 2 - Curva de Carga de um Consumidor Residencial (Francisquini, 2006).

### 2.3.1.2 Consumidor Comercial

Os consumidores comerciais foram responsáveis por aproximadamente 16,96% do consumo de energia elétrica no Brasil no ɔ de 2011. Esses consumidores geralmente são

classificados com base em seus ramos de atividade e consumo mensal de energia elétrica (EPE, 2012).

Segundo EPE (2012) consumidores da classe comercial são aqueles que exercem atividade comercial ou de prestação de serviços, à exceção dos serviços públicos ou de outra atividade não prevista nas demais classes. Esta classe é subdividida nas seguintes subclasses: comercial, serviços de transporte (exceto tração), elétrica, serviços de comunicações e telecomunicações, associação e entidades filantrópicas, templos religiosos, administração condominial, iluminação em rodovias, semáforos, radares e câmeras de monitoramento de trânsito e outros serviços/atividades.

Devido à diversidade das atividades comerciais, diferentes infraestruturas, tamanho dos empreendimentos e tecnologias envolvidas, como lâmpadas mais ou menos eficientes, sistemas de refrigeração ou aquecimento com diferentes níveis de eficiência energética, e, diferentes escalas de trabalho, por exemplo, fazem com que o setor comercial apresente por vezes, curvas de cargas bastante distintas entre si, mesmo para um determinado segmento de mercado ou ramo de atividades em comum.

A Figura 3 apresenta uma curva de carga real de um consumidor da classe comercial. Ao observar a referida curva é possível identificar certa estabilidade no consumo entre 1h e 7h, após esse horário, período que comumente iniciam as atividades de abertura do comércio, ocorre um ligeiro aumento até as 10h, continuando com a observação da curva nota-se um declínio do consumo no horário de almoço, voltando a crescer em seguida. Ao verificar a curva de carga com atenção, é possível identificar um pico local logo após as 11h e o pico diário entre 18 e 19h, decaindo a partir deste horário até as 24h, se estabilizando a 1 hora da manhã.

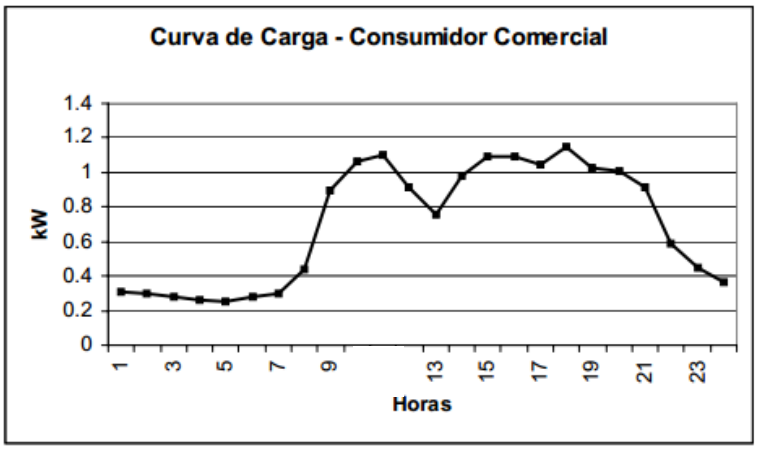



Figura 3 - Curva de Carga de Consumidor Comercial (Francisquini, 2006).

### 2.3.1.3 Consumidor Industrial

Entre os grupos que dividem os tipos de consumidores de energia elétrica, o industrial é o maior em demanda, correspondendo a 42,4% do consumo de energia elétrica em 2011 no Brasil (EPE, 2012).

A EPE (2012) caracteriza os consumidores da classe industrial como aqueles que desenvolvem atividade industrial de acordo com a Classificação Nacional de Atividades Econômicas – CNAE, que caracteriza as indústrias em duas macro esferas, sendo a primeira correspondente as Indústrias Extrativas e a segunda as Indústrias de Transformação (CNAE, 2012).

O setor industrial abrange uma grande gama de atividades, e, indústrias do mesmo ramo podem variar expressivamente em tamanho, tecnologias empregadas e localização, apresentando desta maneira, curvas de cargas bastante distintas.

A Figura 4 apresenta as curvas de cargas diárias de quatro indústrias, na qual é possível verificar diferentes perfis e patamares de consumo. Algumas indústrias apresentam consumo quase constante, outras consomem relativamente pouco durante a maior parte do dia e apresentam um consumo elevado em um período menor de tempo, desta maneira é possível perceber que dependendo do ramo, do tamanho, das tecnologias empregadas, da localização e de outras variáveis envolvidas, suas curvas de carga são moldadas.

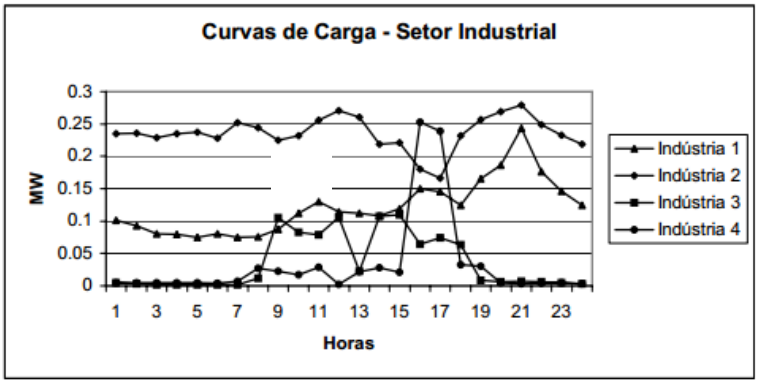



Figura 4 - Curvas de Carga do Setor Industrial (Francisquini, 2006).

### 2.3.1.4 Consumidores Especiais

Consumidores especiais não fazem parte de uma classe apresentada no relatório da EPE (2012), porém são consumidores que podem ser enquadrados em diferentes categorias, mas que

apresentam características de consumo ou de necessidade da não interrupção do fornecimento de energia elétrica, que os fazem especiais.

São classificados como consumidores especiais aqueles de grande sazonalidade, como parques de exposições que podem ficar grandes períodos de tempo consumindo pouca energia e, quando da realização de um evento, podem apresentar uma demanda muito grande, que deve ser atendida de forma segura pelo sistema; existem indústrias que dependem do fornecimento contínuo e de qualidade para realizarem suas atividades, nesta categoria podem se enquadrar indústrias de metais e fabricantes de fibras óticas, entre outras.

Entre os consumidores especiais é importante destacar os hospitais, uma vez que estes dependem do fornecimento de energia elétrica para que equipamentos que dão suporte a manutenção a vida possam funcionar.

É importante perceber que apesar do nome similar, os consumidores citados na resolução ANEEL Nº 247/2006 não apresentam as características aqui utilizadas para definir um consumidor especial.

#### 2.3.1.5 Outras Classes de Consumidores

A EPE (2012) divide os consumidores entre outras cinco classes, abrangendo os consumidores rurais, iluminação pública, poder público, serviço público e consumo próprio, conforme apresentado nos seguintes tópicos:

- **Consumidores Rurais**: são aqueles que desenvolvem atividade relativa à agropecuária, incluindo o beneficiamento ou a conservação dos produtos agrícolas oriundos da mesma propriedade. A caracterização de um consumidor na classe rural está sujeita à comprovação perante a distribuidora. Para os consumidores rurais consideram-se as seguintes subclasses: agropecuária rural, agropecuária urbana, rural residencial, cooperativa de eletrificação rural, agroindustrial, serviço público de irrigação rural, escola agrotécnica e aquicultura.
- **Iluminação Pública**: é caracterizada como aquela sob responsabilidade de pessoa jurídica de direito público ou por esta indicada mediante concessão ou autorização. A classe se caracteriza pelo fornecimento de energia elétrica para iluminação de ruas, praças, avenidas, túneis, passagens subterrâneas, jardins, vias, estradas, passarelas, abrigos de usuários de transportes coletivos, logradouros de uso comum e livre acesso,

iluminação de monumentos, fachadas, fontes luminosas e obras de arte de valor histórico, cultural ou ambiental, localizadas em áreas públicas e definidas por meio de legislação específica, exceto o fornecimento de energia elétrica que tenha por objetivo qualquer forma de propaganda ou publicidade, ou para realização de atividades com objetivos econômicos.

- **Poder Público**: independentemente da atividade desenvolvida, caracteriza-se pelo fornecimento de energia elétrica a unidades consumidoras mediante solicitação realizada por pessoa jurídica de direito público que assuma as responsabilidades inerentes à condição de consumidor. Nesta classe se inclui a iluminação em rodovias e semáforos, radares e câmeras de monitoramento de trânsito, exceto nos casos que possam ser classificados como serviço público de irrigação rural, escola agrotécnica, iluminação pública ou serviço público. A classe é dividida em três subclasses, de acordo com o nível do poder público, sendo em: Poder Público Federal, Poder Público Estadual ou Distrital e Poder Público Municipal.
- **Serviço Público**: é caracterizada pelo fornecimento exclusivo para motores, máquinas e cargas necessárias para realização de serviços públicos de água, esgoto, saneamento e tração elétrica urbana ou ferroviária, explorados diretamente pelo Poder Público ou mediante concessão ou autorização, existem duas subclasses consideradas, sendo a primeira relacionada a tração elétrica e a outra a água, esgoto e saneamento.
- **Consumo Próprio**: também denominada como outros consumos, diz respeito ao fornecimento de energia elétrica destinada ao consumo das instalações da distribuidora. A classe correspondeu a 0,8% de todo consumo de energia elétrica no Brasil no ano de 2011, sendo esta, a classe com menor consumo.

## 2.4 Níveis de Desagregação

O consumo de energia elétrica normalmente é apresentando com grande nível de agregação, isto é, fornece informações referentes a um grande grupo de consumidores, ou referentes a grandes consumidores de EE.

Normalmente os consumidores são agregados em nível de alimentadores, transformadores e subestações de energia elétrica.

Para Mota, Mota e França (2004) a agregação das cargas se dá pela integralização das demandas individuais de um alimentador. Os Autores salientam a existência de duas abordagens para o problema de agregação de cargas, conforme apresentado nos tópicos seguintes:

- **Primeira abordagem**: consiste basicamente em modelar a curva de demanda total de um determinado alimentador, por meio de um método adequado de identificação, como, a partir de séries temporais de medições. Essa abordagem requer um esforço computacional correspondente apenas ao método de identificação analítico utilizado.

- **Segunda abordagem**: consiste na modelagem individual de cada equipamento, efetuando a simulação de seu funcionamento durante o dia, totalizando a demanda por meio da soma dos perfis de consumo obtidos; esta abordagem se destaca pela possibilidade de que diferentes modelos sejam empregados em distintos equipamentos pertencentes a um mesmo alimentador, possibilitando que se obtenham resultados mais precisos para descrever o comportamento da carga agregada. Essa abordagem se mostra mais trabalhosa, e por não fornecer explicitamente um modelo para a demanda agregada, exige que se empregue um método de identificação analítico adequado.

A respeito da curva de carga diária dos transformadores de distribuição, Meffe (2001) salienta que ela é obtida pela agregação das curvas de carga diárias dos consumidores correspondentes.

Com a agregação das curvas de carga de diversos consumidores, especialmente de uma mesma classe, as curvas obtidas são suavizadas. Isso ocorre em parte por mudanças individuais e coletivas no consumo de EE, onde devido ao grande número de consumidores, ocorre a compensação parcial de mudanças significativas originárias de consumidores específicos.

Porém quando se trata do nível mais desagregado, isto é, de uma única unidade consumidora, podem ser encontrados os maiores níveis de incertezas, e, portanto, as maiores oscilações no consumo de EE.

## 2.5 Modelos de Previsão

Segundo Esteves (2003), para que se possam estabelecer valores futuros para uma dada série temporal, é necessário que se consiga abstrair as características e o comportamento de determinada série, sendo que, tais informações devem ser extraídas dos dados disponíveis. A literatura dispõe de diversos modelos de previsão adequados para exercer tal tarefa, sendo esses modelos normalmente classificados em uma das três categorias apresentadas a seguir:

- **Modelos Univariados**: são modelos que utilizam apenas valores passados de uma determinada série para explicá-la. Isso ocorre em geral com dados de consumo de energia elétrica. Esta classificação abrange os métodos de amortecimento exponencial, métodos de decomposição e os modelos Box&Jenkins.

- **Modelos Causais**: também conhecidos como modelos de função de transferência - MFT, se diferenciam dos modelos univariados por considerarem também outras séries correlacionadas. Para explicar, consideraremos o consumo de energia elétrica, no qual também pode-se considerar uma série com os preços relativos, ou ainda, as condições climáticas verificadas para determinados dias, para então efetuar previsões.

- **Modelos Multivariados**: São modelos que permitem realizar múltiplas previsões ao mesmo tempo, caracterizando-se assim, como um modelo único capaz de prever o que se dará com diversas séries. Um exemplo, com foco no setor elétrico, seria um modelo que permitisse prever ao mesmo tempo o consumo de energia elétrica em distintas concessionárias de energia no Brasil.

### 2.5.1 Seleção de Métodos de Previsão

A escolha de um método de previsão se dá ao observar diversos fatores, como a natureza do problema, os conhecimentos prévios do pesquisador, a precisão esperada, os custos computacionais, entre outros.

É importante ter ciência que os métodos de previsão apresentam vantagens e desvantagens, seja no campo da qualidade dos resultados ou na facilidade de uso, de tal maneira que, o critério de finalidade de uso deve exercer importante peso sobre a escolha de um modelo, uma vez que a aplicação em determinados setores podem exigir diferentes níveis de qualidade.

Moretin e Toloi (1981) apresentam cinco critérios organizados em ordem decrescente de importância, que podem ser considerados na seleção de métodos de previsão:

- Precisão;
- Comportamento (exibido pela variável a prever);
- Horizonte temporal;
- Custos;
- Facilidade de aplicação.

## 2.6 Métodos Comuns de Previsão

Segundo Guirelli (2006), na década de 80 já se aplicavam uma série de métodos para a previsão de carga, sendo possível dividi-los em dois grupos, o primeiro formado pelos métodos estatísticos (ou paramétricos), e o segundo pelos métodos baseados em inteligência artificial (ou não paramétricos). Designam-se como híbridos os métodos que se baseiam na combinação de dois ou mais métodos de previsão de carga. Os tópicos seguintes apresentam alguns integrantes dos grupos mencionados:

- Entre os métodos clássicos, podem ser citados:
  - Regressão Linear Múltipla;
  - ARIMA;
  - Alisamento Exponencial;
  - Análise Espectral.
- Entre os métodos Baseados em Inteligência Artificial, é possível destacar:
  - Sistemas Especialistas;
  - Redes Neurais;
  - Lógica Fuzzy;
  - Algoritmos Genéticos.

### 2.6.1 Regressão Linear Múltipla

Guirelli (2006) explica que na Regressão Linear Múltipla, a carga pode ser modelada como uma função linear de múltiplas variáveis, tal que:

$$y(t) = a_0 + a_1 x_1(t) + \cdots + a_n x_n(t) + a(t) \qquad (2.1)$$

onde: $y(t)$ corresponde a carga no instante $t$; $a_0$, $a_1$ e $a_n$ são coeficientes da regressão; $a(t)$ é uma variável do tipo aleatória de média zero e variância constante; $x_1(t)$,... $x_n(t)$ são variáveis explicatórias.

Variáveis explicatórias são fatores que exercem influência sobre a carga do sistema, como por exemplo, a temperatura ambiente. Análises estatísticas podem apresentar a significância de cada variável explicatória para previsão de carga. É possível encontrar os coeficientes de regressão por meio do métodos dos mínimos quadrados (Guirelli, 2006).

O método é pouco utilizado, uma vez que comparações realizadas apontam que ele apresenta maiores erros, quando comparado com outros métodos (Moghram e Rahman, 1989).

### 2.6.2 Autoregressive Integrated Moving Average - ARIMA

O modelo auto-regressivo integrado de média móvel - ARIMA é a generalização do modelo auto-regressivo de média móvel - ARMA. O modelo é bastante utilizado na modelagem e previsões de séries temporais.

Segundo Guirelli (2006) a teoria para séries temporais lida com processos estacionários, séries temporais sem tendências, onde não há mudança sistemática na sua variância. Processos não estacionários devem ser convertidos em estacionários, com a remoção de suas tendências e sazonalidades.

O modelo ARIMA pode ser descrito pela equação (Box e Jenkins, 1976; Felipe, 2012):

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \qquad (2.2)$$

onde: $a_0$ representa uma constante no modelo estimado, $a_1, \ldots, a_p$ ajustam os valores passados de $y_t$ do instante imediatamente anterior até o mais distante que é denotado por $p$. $\varepsilon$ denota o componente errático da série, uma sequência de choques aleatórios e independentes uns dos outros, e, $\varepsilon_t$ é uma porção não controlável do modelo, conhecido como ruído branco se a série for não estacionária. $\beta_1, \ldots, \beta_q$ permitem escrever a série em função de choques passados. Normalmente $\varepsilon_t$ é considerado como tendo distribuição normal, média zero, variância constante e não correlação.

Em termos estatísticos, considerando que $E[x]$ denota a média teórica do valor de $x$, a sequência $\varepsilon_t$ é considerada um processo de ruído branco se para cada período de tempo $t$ tem-se (Felipe, 2012):

(i) $E[\varepsilon_t] = E[\varepsilon_{t-1}] = \cdots = 0$, média zero;

(ii) $E[\varepsilon_t^2] = E[\varepsilon_{t-1}^2] = \cdots = \sigma^2$, variância constante;

(iii) $E[\varepsilon_t . \varepsilon_{t-s}] = E\left[\varepsilon_{t-j} . \varepsilon_{t-j-s}\right] = \cdots = 0$, covariância nula para todo valor de $s$.

### 2.6.3 Alisamento Exponencial

Guirelli (2006) aponta que no Alisamento Exponencial, considera-se que cada elemento de uma determinada série temporal seja formado por uma constante e um componente de erro, de tal forma que:

$$x_t = b + \varepsilon_t \tag{2.3}$$

onde: $b$ é o componente constante, e $\varepsilon$ é o componente de erro.

A parte $b$ é constante ao longo de cada segmento de uma determinada série temporal, mas é passível de variar no tempo, pode-se então definir que uma medição atual é função das anteriores, sendo que as mais antigas apresentam um peso exponencialmente menor (Guirelli, 2006):

$$s_t = \alpha \quad \vdash (1 - \alpha) s_{t-1} \tag{2.4}$$

onde: $s$ corresponde aos valores previstos nas medições; $x$ denota as medições; $\alpha$ corresponde a constante de alisamento e $t$ é o instante em que se calcula a série temporal.

### 2.6.4 Análise Espectral

De acordo com Esteves (2003), na análise espectral são situadas as características de um processo estocástico em termos de frequências, sendo que, no caso das séries temporais, pode determinar as periodicidades existentes na mesma. É necessário estimar o espectro de um processo, uma vez que ele não é conhecido. Normalmente é estimado por meio do periodograma de janelas espectrais, uma vez que possui boas propriedades estatísticas.

### 2.6.5 Sistemas Especialistas

Guirelli (2006) define um sistema especialista como “um programa que possui uma grande base de dados sobre um domínio específico e usa um complexo raciocínio por inferência para realizar tarefas que podem ser feitas por um especialista humano”. Segundo o autor, a aplicação de sistemas especialistas não é muito disseminada devido à necessidade de se ter um especialista capaz de prever a carga, e depois é preciso converter o conhecimento do especialista em regras matemáticas, processo esse, moroso. É comum encontrar esse tipo de sistema associado a outras técnicas, como Lógica Fuzzy e Redes Neurais, possibilitando que se encontrem resultados melhores, comparados a aqueles que seriam verificados caso as técnicas fossem empregadas separadamente.

### 2.6.6 Redes Neurais Artificiais

As Redes Neurais Artificiais podem ser entendidas como uma técnica de processamento que se baseia nos sistemas nervosos biológicos, portanto, formadas por neurônios interconectados. Elas são treinadas para problemas específicos através de um processo de aprendizado, que consiste no fornecimento de um conjunto de dados de entrada e saída, sendo no caso, as entradas compostas pelas cargas anteriores, e a saída pela carga que se deseja prever. Uma vez treinada, é capaz de prever valores futuros com base em dados passados. É possível incluir variáveis nas redes neurais, sem que seus relacionamentos matemáticos com a carga sejam conhecidos. Existe uma série de técnicas para redes neurais artificiais, onde é possível variar a estrutura da rede e o método de tre mento (Guirelli, 2006).

As redes neurais muitas vezes são vistas como *caixas pretas*, uma vez que conduzem a soluções de problemas complexos, sem que seja entendido ou apresentado de maneira explícita a evolução do processo que conduziu até a solução apresentada.

### 2.6.7 Lógica Fuzzy

A Lógica Fuzzy ou lógica difusa foi desenvolvida originalmente por Zadeh (1965). Diferentemente da lógica convencional, não utiliza valores binários, 0 e 1 (não ou sim). A lógica difusa pode assumir infinitos valores entre 0 e 1, incorporando dessa maneira, imprecisões e imperfeições do mundo real. A aplicação da lógica difusa na previsão de carga ocorre devido a

sua capacidade aproximar qualquer função não-linear com grande precisão, além de encontrar padrões em grandes conjuntos de dados (Guirelli, 2006).

### 2.6.8 Algoritmos Genéticos

Os Algoritmos Genéticos - AG normalmente são utilizados na previsão de carga como uma ferramenta de auxílio para outros métodos, como por exemplo, na otimização da estrutura de redes neurais. Os AG são poderosas ferramentas para grandes problemas de otimização combinatória, se baseiam na sobrevivência dos melhores indivíduos, passagem de características aos descendentes e a aplicação de mutações (Guirelli, 2006).

## 2.7 Método de Dias Similares

O método de dias similares (MDS) é um método estatístico causal que busca prever a carga futura traçando analogias com dados do passado, para tanto, diversas características podem ser consideradas, dentre as quais se destacam os tipos de dias e dados meteorológicos.

O MDS é útil por não lidar apenas com a parte não linear da curva de carga, mas também com os dias especiais, como os fins de semana e feriados. Esta abordagem também se mostra adequada para situações em que os modelos de previsão precisos são difíceis de projetar (Senjyu, Higa e Uezato, 1998).

Dentre os estudos que utilizam a abordagem de dias similares com o objetivo de realizar previsões de carga, pode-se citar: (Mu *et al*., 2010) que aborda o uso de dias similares para previsão de carga de curto prazo, incorporando pesos nos dias selecionados, de maneira que os mais semelhantes possam exercer maior influência sobre a previsão, o trabalho trata ainda da seleção de dias similares, além de discorrer sobre situações em que esses não estejam disponíveis; (Barghinia *et al*., 2008) aborda a utilização combinada de redes neurais bayesianas, Neuro-Fuzzy e busca de dias similares, com o objetivo de aprimorar previsões de carga de curto prazo; (Mandal *et al.*, 2008) apresenta uma análise da sensibilidade dos parâmetros de dias similares, que são utilizados em uma rede neural artificial baseada no método de dias similares, com o objetivo de efetuar predição horária de preços para os mercados da Pennsylvania, New Jersey e Maryland; (Senjyu *et al*., 2005) propõem uma abordagem que faz uso de lógica fuzzy, onde ela é utilizada e com base nos dias similares corrige a saída da rede neural artificial

utilizada para predição de curto prazo, o trabalho apresenta ainda a utilização da norma euclidiana com fatores ponderados na seleção dos dias similares. (Kadowaki *et al*., 2004) apresenta um sistema especialista baseado em dias similares, tal sistema tem por objetivo realizar previsão de carga de curtíssimo prazo para o período da ponta. Os trabalhos apresentados não tratam explicitamente dos níveis menos agregados, onde existem os maiores níveis de incertezas, e grandes oscilações nas curvas de carga.

Para Kadowaki et al. (2004), um conjunto de dias similares é formado por agrupamentos de dias que possuem, para cada intervalo de tempo analisado, curvas de carga semelhantes. Para que os dias similares possam ser escolhidos, fatores como as demandas de carga, condições climáticas, tipo de dia, horário de verão, horário do pôr do sol, temperatura entre outros podem ser utilizados na forma de regras heurísticas.

Para a seleção dos dias similares Senjyu, Higa e Uezato (1998) abordam a utilização da norma euclidiana com fatores ponderados para avaliar a semelhança entre os dias, sendo que os fatores ponderados são utilizados por causa da diferença de unidade entre os elementos. A norma euclidiana também é aplicada para selecionar os dias que serão utilizados na previsão, quanto mais a norma euclidiana diminui, melhor se torna a avaliação da similaridade. Antes de aplicar a norma, verificou-se a forte correlação entre alguns dados climáticos e as curvas de carga, com destaque para a temperatura máxima, mínima, tipo de dia e a carga, sendo assim, foram considerados dias semelhantes aqueles selecionados através da temperatura máxima, mínima, tipo de dia e da norma euclidiana.

Os trabalhos que se baseiam no método dos dias similares costumam se diferenciar quanto a diversificação, a qualidade e a quantidade de dados utilizados, uma vez que determinados dados não estejam disponíveis, é possível adaptar o método para essa realidade. A utilização ou não utilização de determinados dados externos às curvas de carga, pode influenciar de maneira mais ou menos significativa a qualidade das previsões. Para determinar quais informações são mais significativas, uma das alternativas é verificar o grau de correlação entre as variáveis, ou ainda, através de testes, variando a utilização dos dados externos, observando a assertividade das previsões, com o objetivo de determinar quais características devem ser consideradas no cenário em questão.

O MDS se baseia em dados que alcançam horizontes de um ou mais anos, de onde se procura determinar similaridades entre os dias. As características peculiares aos dias

semelhantes levam em consideração a data, de onde é possível verificar a existência de feriados e a estação do ano a que cada dia pertence.

A demanda verificada de um dia semelhante pode ser considerada uma previsão. Normalmente a previsão vai além de considerar apenas um dia semelhante, em vez disso, costuma envolver análises mais complexas, que podem considerar coeficientes de tendência, entre outras características, possibilitando previsões mais apuradas (Senjyu, Higa e Uezato, 1998).

### 2.7.1 Fatores que Influenciam a Curva de Carga

A curva de carga de um consumidor específico, ou de um conjunto de consumidores, apresenta todas as informações relativas ao comportamento da carga e à sua solicitação ao sistema que a supre (Kagan, Oliveira e Robba, 2010).

É possível observar diferenciações entre as curvas de carga diária dos dias compreendidos em uma semana, sendo possível observar uma variação ainda mais acentuada quando no decorrer de um ano, onde, passam as estações e ocorrem períodos de férias entre outras datas.

Existem diversos fatores que influenciam como os usuários consomem energia elétrica, quando se fala em usuários residências os níveis de incertezas são ainda maiores, sendo assim, tais fatores são difíceis de considerar em um sistema de previsão, desta maneira, a alternativa é a utilização de fatores alheios a vontade do consumidor, dentre eles, destacam-se as condições climáticas, os tipo de dia, o horário de verão, o horário do pôr do sol e a temperatura, conforme abordados a seguir (Kadowaki *et al.*, 2004).

- **Condição climática**: apresenta elevada influência sobre a curva de carga, devido principalmente a seu impacto sobre a carga de iluminação, que é aumentada em dias nublados, por exemplo;

- **Tipo de dia**: normalmente são divididos entre dias úteis, finais de semana e feriados, porém em outras situações são admitidos outros tipos, como sábados, domingos, pré-feriados, pós-feriados e feriados nacionais. Poderiam ser admitidos dezenas de subgrupos de tipos de dias, porém nesse caso, o objetivo de possuir alguns perfis de carga representativos se perderia. É importante destacar que cada cenário de estudo pode se mostrar único, e exigir diferentes tipos de dias a serem considerados;

- **Horário de verão**: apresenta influência sobre a curva de carga, normalmente causando um deslocamento e uma pequena redução na demanda máxima;

- **Horário do pôr do sol**: no sudeste brasileiro o horário de pôr do sol em Junho (inverno) é aproximadamente às 17:30h, se deslocando para 19:00h em Janeiro (verão), desta maneira é possível perceber que tal fenômeno causa importante impacto sobre a carga de iluminação principalmente, fazendo com que a rampa da ponta seja cada vez mais deslocada para a direita;

- **Temperatura**: apresenta influência sobre a curva de carga, em parte devido ao uso de determinados equipamentos elétricos, dentre eles destacam-se os sistemas de climatização. Pode apresentar grande variação, principalmente entre as estações do ano e de acordo com a umidade do ar.

# Capítulo 3

# Séries Temporais

## 3.1 Considerações Iniciais

Uma série temporal - ST pode ser definida com um conjunto de observações ordenadas no tempo. Os valores mensais de consumo de energia ativa de uma determinada unidade consumidora é um exemplo de série temporal (Morettin e Toloi, 2006).

A relação temporal entre as observações compõe a característica mais importante dos dados temporais. Justamente por apresentarem essa relação temporal, as análises empregadas devem observar as peculiaridades intrínsecas, diferentes na análise de dados tradicionais. A descoberta de novos conhecimentos pode ser auxiliada pela análise dos dados temporais em diferentes prismas, não contempláveis com conjuntos tradicionais de dados (Aikes Junior, 2012).

Trabalhar com séries temporais permite de maneira mais imediata, verificar eventos ocorridos no passado. Porém, o mais desejado talvez seja explorar a capacidade dessas séries de explicarem acontecimentos do presente, e permitirem ainda, estimar valores para um futuro, caracterizando previsão.

Neste capítulo são abordadas características, definições, objetivos da análise e outras peculiaridades das séries temporais.

## 3.2 Definições

As séries temporais são geradas por meio da observação de ocorrências que podem ser quantificadas numericamente, gerando uma sequência de dados distribuídos no tempo. Uma série temporal pode ser expressa por (Souza, 1989):

$$S_t = \{S_t \in \Re | \, t = 1, 2, 3 \ldots n \} \tag{3.1}$$

onde: $t$ é um índice temporal, e $n$ representa o número de observações.

Com base no intervalo em que os dados são observados classificam-se as séries temporais como discretas ou contínuas. As primeiras são observadas de maneira discreta, isto é, em um determinado período de tempo e com observações equidistantes. No caso das séries temporais continuas, são observadas em intervalos de tempo não necessariamente equidistantes, buscando um período de observação quase que contínuo no tempo (Morettin e Toloi, 2006; Brockwell e Davis, 2002).

Em diversas situações séries temporais continuas são convertidas em discretas, para tanto é necessário realizar uma amostragem da série continua em intervalos de tempo equidistantes, $\Delta t$ (Morettin e Toloi, 2006).

## 3.3 Componentes de Séries Temporais

Em uma série temporal é possível observar a existência de três componentes básicos, tendência, sazonalidade e ruído, este último também designado por resíduo. A Equação 3.2 apresenta os referidos componentes de uma série temporal (Morettin e Toloi, 2006):

$$S_t = J_t + K_t + L_t \tag{3.2}$$

onde: $S$ representa a série, $J$ a tendência, $K$ a sazonalidade e $L$ representa o ruído, todos para um instante $t$ da série.

Cada um dos três componentes é abordado separadamente nos tópicos seguintes (Pellegrini e Fogliatto, 2001):

- **Tendência**: essa característica é observada quando a série apresenta comportamento ascendente ou descendente por um longo período de tempo;
- **Sazonalidade**: essa característica é verificada quando padrões cíclicos de variação se repetem em intervalos relativamente constantes de tempo;
- **Ruído**: compreende variações que não podem ser explicadas pelas características de comportamento da série temporal (como as descritas nos tópicos supracitados), são ruídos aleatórios ocorridos no processo gerador dos dados.

Existem trabalhos que consideram outras características de comportamento além da tendência e sazonalidade, compreendendo ainda, média e ciclo. A característica de média ocorre quando os valores da série temporal flutuam em torno de uma média constante, e o ciclo é observado quando a série exibe variações ascendentes e descendentes, entretanto, ocorrem em intervalos não regulares de tempo (Pellegrini e Fogliatto, 2001).

## 3.4 Análise de Séries Temporais

As séries temporais podem ser estudadas por diversos motivos e com distintos objetivos. Dentre os interesses possíveis ao estudo de ST, podem-se citar (Morettin e Toloi, 2006; Liu, 2009):

- Investigar e caracterizar o mecanismo gerador da série temporal;
- Efetuar previsões de valores futuros para a série;
- Descrever o comportamento da série;
- Analisar inter-relacionamentos de variáveis;
- Verificar a existência de periodicidades relevantes nos dados;
- Controlar e otimizar processos em sistemas.

De acordo com os interesses de estudo de ST supracitados, e de maneira geral, diversos autores dividem o objetivo da análise de séries temporais em quatro categorias (Chatfield, 2003; Liu, 2009):

- **Descrição**: objetiva verificar comportamentos da ST, descrevendo a existência ou não de tendência, sazonalidade, *outliers*, mudas estruturais entre outros;
- **Explicação**: baseia-se em duas ou mais variáveis e se dedica ao estudo do relacionamento entre elas em um determinado sistema. O processo de formulação de modelos estatísticos que representem as relações entre as variáveis em um sistema, também é conhecido como análise estrutural;
- **Previsão**: é uma das mais universais e importantes aplicações da análise de séries temporais. Dados os valores passados de uma determinada série temporal, busca-se prever seus possíveis valores futuros;
- **Controle**: também é designado como controle e otimização. Sendo possível representar o relacionamento entre variáveis em um sistema por meio de um modelo matemático,

logo se assume que seja possível efetuar previsão, portanto, permitindo controlar determinadas aplicações.

Existem diversas ferramentas computacionais que dão apoio a análise de séries temporais, entre elas pode-se destacar o Weka (*Waikato Environment for Knowledge Analysis*), um software livre de código aberto com licença GPL (*General Public License*) para mineração de dados, que pode ser utilizado para realizar tarefas de classificação, minerar regras de associação, clusterizar dados, além de permitir a análise e até mesmo previsão de séries temporais (Gonçalves, 2011).

## 3.5 Mineração de Dados de Séries Temporais

A economia moderna tem se baseado cada vez mais em informações, característica que vem alterando o ambiente operacional das organizações e dos negócios modernos. O modo como os dados são coletados e analisados também foi alterado, e graças ao amplo uso de tecnologia de informação, gigantescas quantidades de dados têm sido coletadas em ambientes on-line e de tempo real (Liu, 2009).

Os dados ordenados no tempo, normalmente são agregados de acordo com um intervalo apropriado, produzindo um grande volume de dados de séries temporais equiespaçadas, passíveis de análise e exploração por diversas ferramentas e metodologias modernas, desenvolvidas para a análise desse tipo de dados (Liu, 2009).

Devido a diversas características que incluem a dificuldade em se analisar diretamente séries temporais muito extensas, a subjetividade em se comparar distintas séries temporais, que podem apresentar distintas taxas de amostragem, ruídos, valores ausentes entre outros, que a mineração de dados de séries temporais tem sido largamente utilizada, permitindo extrair conhecimentos relevantes a partir de tais conjuntos de dados.

De acordo com o objetivo da análise a ser realizada, determinadas tarefas de mineração de dados de séries temporais podem ser empregadas, entre elas podem ser destacadas a de pré-processamento, recuperação de conteúdo, agrupamento, classificação, detecção de anomalias e previsão.

### 3.5.1 Pré-Processamento

À tarefa de pré-processamento cabe tratar os dados de maneira que ao medir-se a distância entre duas séries temporais, os resultados obtidos sejam os mais próximos a realidade, uma vez que as séries já foram verificadas quanto a presença de distorções em seus dados.

Durante a obtenção dos dados temporais, que inicia com a medição junto aos sistemas observados, o subsequente transporte por uma rede de dados, e que termina com a interceptação pelo computador servidor, para que finalmente possam ser armazenados, diversos processos são efetuados e protocolos de comunicação de dados observados, onde erros podem ocorrer, além daqueles causados por possíveis problemas e limitações nos sistemas de medição.

À etapa de pré-processamento cabe preparar os dados para que estes possam ser adequadamente analisados. Nesta etapa deve ser observada a existência de valores faltantes, informações discrepantes, amostragem irregular e tendência (Pyle, 1999).

#### 3.5.1.1 Valores Faltantes

A ausência de dados pode ser decorrente de mau funcionamento de equipamentos de medição, erros no preenchimento de formulários, inconsistência com outros registros que acabam por levar a supressão de determinados dados, entre outras situações possíveis.

Para evitar incoerências nos dados temporais, é necessário realizar estudos que permitam identificar esse tipo de ocorrência, e possibilitem a aplicação de tratamentos adequados.

Existem diversas maneiras de se lidar com dados ausentes, uma das técnicas baseia-se na substituição dos dados faltantes pela média dos vizinhos mais próximos, outra, pelo vizinho anterior mais a média de crescimento constatada para determinado intervalo de tempo, entre outras abordagens possíveis (Neves e Alvares, 2003).

É importante observar que ao se introduzir informação onde não existia, sendo essa informação concebida de modo estatístico ou por qualquer outro meio que não seja a medição convencional adotada, os resultados obtidos podem ser distanciados da realidade.

Introduzir dados pode gerar tendência, por tal motivo, em diversas situações pode-se inserir ruído aos valores faltantes preenchidos, ou ainda, escolher não preencher valores faltantes, e lidar com séries de dados sem essa característica.

### 3.5.1.2 Valores Discrepantes

Valores discrepantes, aberrantes ou *outliers*, como as próprias designações sugerem, são informações que fogem do esperado. Tais informações devem ser tratadas com importância, e quando constatado que de fato são informações errôneas, devem ser tratadas.

Informações tidas como discrepantes que tenham sido adequadamente verificadas e dadas como errôneas, podem ser abordadas de maneira similar aos dados ausentes, efetuando sua substituição com base em alguma técnica escolhida.

Existem diversos meios de verificar valores discrepantes, uma abordagem possível seria a tabulação dos dados, a verificação da média, do desvio padrão, máximo e mínimo, somado ao conhecimento dos limiares tangíveis, de maneira que os valores fora desse espaço possam ser considerados como discrepantes.

O uso de clusterização (*clustering*) de dados também pode ser um meio adequado para verificar a existência de valores discrepantes, feita a clusterização de determinados dados, elementos isolados, não pertencentes a nenhum dos agrupamentos formados, poderiam ser melhor analisados e eventualmente considerados como discrepantes. A Figura 5 ilustra um *outlier*.

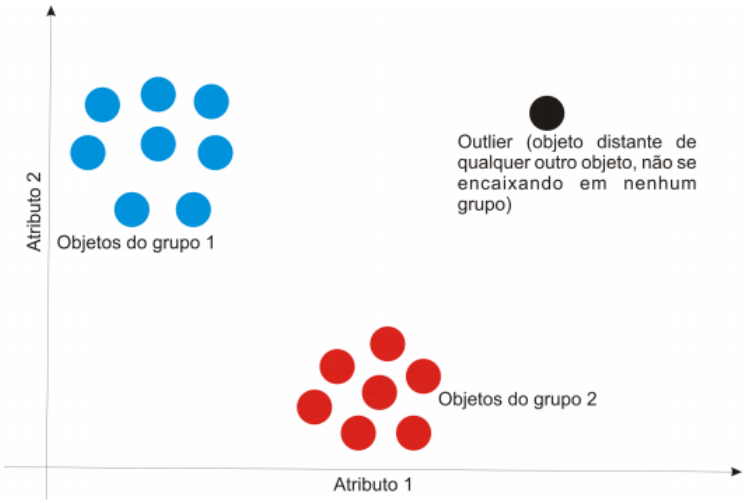



Figura 5 - Representação de um Outlier em uma Base de Dados de Duas Dimensões (Fontana e Naldi, 2009)

### 3.5.1.3 Amostragem Irregular

Em grande parte dos casos as séries temporais são formadas por valores equiespaçados, isso é, a mensuração dos valores são realizadas em espaços idênticos de tempo.

Alguns métodos de análise de séries temporais lidam apenas com séries equiespaçadas, e nos casos em que essa característica não é observada nas séries, se faz necessário um ajuste em seus valores, de modo que reflitam os valores caso o intervalo de amostragem tivesse sido regular (Pyle, 1999).

#### 3.5.1.4 Tendência

O componente tendencial, também conhecido como tendência ou tendência secular, diz respeito a um fator evolutivo que demonstra a influência de fatores que fazem com que um fenômeno tenha sua intensidade aumentada ou diminuída com o passar do tempo. Caracteriza-se como um movimento ascendente ou descendente de longa duração. Uma série temporal que não apresenta tendência ascendente ou descendente é dita como uma série estacionária (Scottá e Fonseca, 2013).

### 3.5.2 Recuperação de Conteúdo

A recuperação de conteúdo em Séries Temporais – ST é um tema de crescente interesse, e tem como objetivo recuperar conteúdos de bases de dados, para tanto, faz uso de uma determinada ST passada como exemplo, de tal forma que os conteúdos recuperados sejam os mais semelhantes a essa ST (Mörchen, 2006).

De acordo com Chen e Ozsu (2003) a recuperação por conteúdo em ST pode ser dividida em duas categorias principais, sendo elas:

- ***Pattern Existence Queries***: Diz respeito a busca por séries que apresentem determinado padrão em seus dados;
- ***Exact Match Queries***: Nessa categoria são especificados os valores exatos que devem ser apresentados pelas ST recuperadas;

Tal recurso permite que especialistas possam encontrar acontecimentos similares prévios, ou até mesmo, verificar que tal acontecimento (ST) admitido como referência, trata-se de uma situação nova, anômala para determinado sistema.

### 3.5.3 Classificação de Dados Temporais

A área de classificação de dados é um tópico importante no campo da mineração de dados, e se destaca pela grande aplicabilidade. Diversos métodos vem sendo propostos com base em

modelos conhecidos de aprendizado de máquina, como as redes neurais e as árvores de decisão. Apesar da área de classificação de dados ser bastante estudada, menos trabalhos abordam a classificação de dados temporais (Tseng e Lee, 2009).

De acordo com Larose (2005) a tarefa de classificação de dados consiste em classificar entradas desconhecidas com base nos valores de seus atributos, para isso, inicialmente toma-se um conjunto de exemplos compostos por vários atributos e com classes conhecidas.

As tarefas de classificação de ST podem ser divididas entre Classificação de Séries Temporais e Classificação de Pontos Temporais, em que no primeiro caso, pretende classificar as séries temporais inteiras, para tanto, aplica-se uma etiqueta a cada série de treino, já no segundo caso, o objetivo é classificar determinados pontos da série, e para isso são atribuídas etiquetas aos pontos de interesse nas séries de treino (Morchen, 2006).

### 3.5.4 Agrupamento

As tarefas de agrupamento ou clusterização (clustering) consistem na divisão de determinados conjuntos de objetos em grupos (clusters) onde as similaridades entre os objetos que compõem um cluster devem ser as maiores possíveis, e a diferenciação entre os clusters deve ser maximizada (Cluto, 2010).

Para que objetos possam ser divididos em grupos, é necessário utilizar algum critério de medida de similaridade entre eles. Uma abordagem convencional para a comparação de dois objetos é a associação de uma função de distância (calculada através dos valores de seus atributos) ao conceito de dissimilaridade, onde dois objetos podem ser classificados como altamente dissimilares se a distância entre eles for alta, e baixamente dissimilares se a distância for pequena. Entre as medidas de similaridade existentes destacam-se as distâncias Manhatan e Euclidiana, exemplificadas na Figura 6 (Fontana e Naldi, 2009):

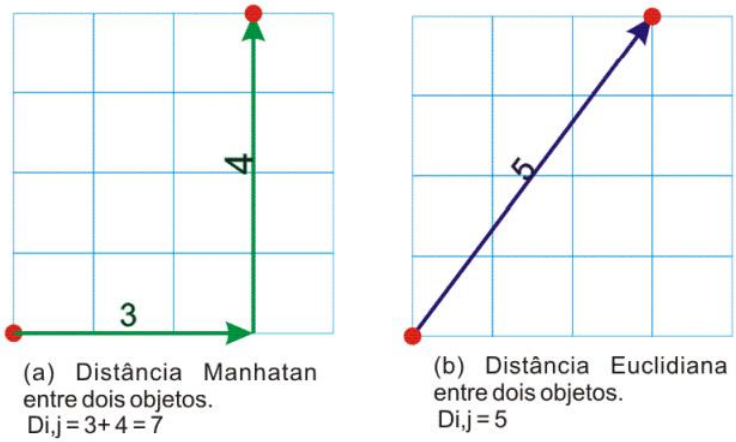



Figura 6 - Representação Gráfica das Distâncias Manhatan e Euclidiana (Fontana e Naldi, 2009)

Para distinguir as tarefas de classificação e agrupamento de dados, a Tabela 1 apresenta a comparação entre agrupamento e classificação (Fontana e Naldi, 2009):

Tabela 1- Comparações entre Agrupamento e Classificação (Fontana e Naldi, 2009)

| **Classificação** | **Agrupamento** |
|---|---|
| Número de categorias definido | Número de categorias não definido |
| Supervisionado | Não supervisionado |
| Categorias previamente definidas | Categorias não são previamente definidas |
| Classificar um novo objeto, entre as categorias já definidas. | Agrupar objetos conforme as suas similaridades: objetos de mesmo grupo têm alta similaridade, enquanto objetos de grupos diferentes têm baixa similaridade. |

Os algoritmos de agrupamento geralmente são classificados como particionais, hierárquicos ou híbridos. Os algoritmos hierárquicos utilizam aglomerações ou divisões de grupos para produzir um conjunto de partições. Algoritmos do tipo aglomerativo efetuam a união de grupos de maneira iterativa, até que um único grupo contenha todos os objetos da base de dados, já os algoritmos divisivos separam iterativamente os grupos, em dois, até que cada objeto forme seu próprio grupo. A decomposição resultante é uma árvore de grupos que é conhecida como dendrograma, que permite visualizar a maneira que um algoritmo gera uma saída, além de permitir analisar essa saída, é capaz de classificá-la como boa ou não (Fontana e Naldi, 2009).

Para Morchen (2006) existem três categorias principais em que podem ser divididas as abordagens de agrupamento de séries temporais, sendo elas:

- ***Whole Series Clustering***: Esta abordagem diz respeito ao agrupamento de um conjunto de séries temporais numéricas, com base em alguma medida de similaridade obtida de algoritmos de agrupamento convencionais;
- ***Sub-series Clustering***: Nesta abordagem uma série temporal longa possui segmentos selecionados, permitindo a formação de um conjunto de séries temporais menores, dando sentido ao nome Agrupamento de Subséries;
- ***Time Point Clustering***: Consiste no processo de agrupar determinados pontos de uma ST com base na combinação de seus valores e na proximidade temporal.

Existem diversos algoritmos para clusterização de dados, que variam na abordagem, mas possuem um propósito em comum, gerar agrupamentos significativos. Entre os algoritmos de clusterização podem ser citados (Fontana e Naldi, 2009):

- **Divisive Analysis** - DIANA: um algoritmo divisivo hierárquico proposto por Kaufman e Rousseeuw (1990), ele inicia com um único grupo composto por todos os objetos, e realiza sua função por meio da heurística de dividir o maior grupo iterativamente, o diâmetro de um agrupamento é tido como a maior distância entre dois objetos que o compõem;

- **Bisecting k-means**: consiste em uma variação hierárquica do algoritmo k-means, onde em cada iteração seleciona um grupo e o divide, formando uma hierarquia;

- **X-means**: foi proposto por Pelleg e Moore (2000) com o objetivo de gerar uma partição do conjunto de dados fazendo uso de k-means, como entradas o algoritmo deve receber a base de dados a ser particionada e um intervalo de grupos;

- **K-means**: utiliza o conceito de centróides, sendo apresentado em detalhes a seguir.

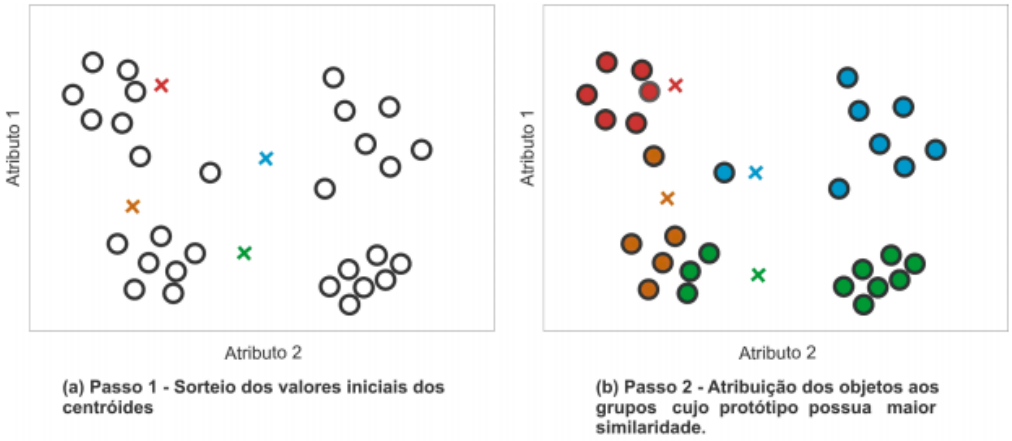
Um dos algoritmos mais utilizados para realizar tarefas de agrupamento (clusterização) de dados é o *K-means*, também é conhecido como K-médias. Esse algoritmo utiliza o conceito de centróides como protótipos representativos dos grupos (clusters), sendo calculados pela média de todos os objetos do grupo que representa.

O objetivo do algoritmo K-means "[...] é encontrar a melhor divisão de $P$ dados em $K$ grupos $C_i, i = 1, \dots, K$, de maneira que a distância total entre os dados de um grupo e o seu

respectivo centro, somada por todos os grupos, seja minimizada" (Pimenteli, França e Omar, 2003).

O algoritmo K-means pode ser descrito em quatro passos, ilustrados pelas Figuras 7, 8 e 9 (Fontana e Naldi, 2009; Beltrame e Fonseca, 2010):

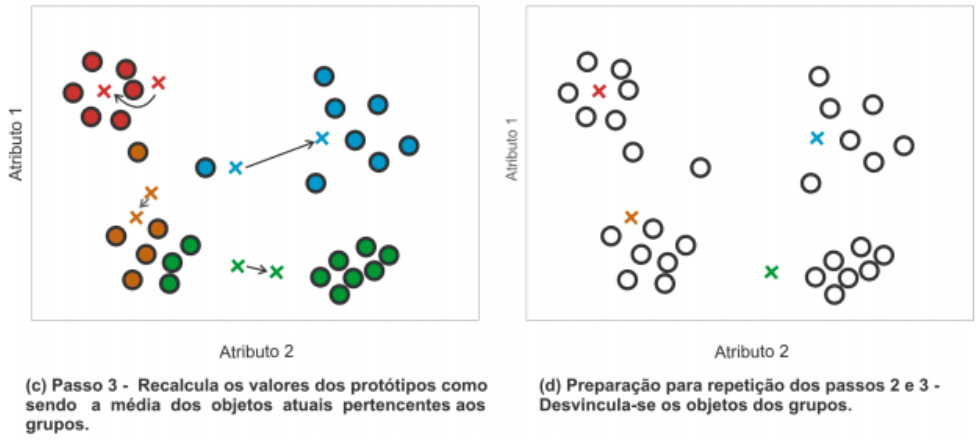
- **Passo 1**: Inicialmente atribuem-se valores para os protótipos seguindo algum critério, um exemplo seria o sorteio aleatório desses valores dentro dos limites de domínio de cada atributo;
- **Passo 2**: Atribui-se cada objeto ao grupo cujo protótipo possua maior similaridade com o objeto;
- **Passo 3**: Recalcula-se o valor do centróide (protótipo) de cada grupo, como sendo a média dos atuais objetos do grupo;
- **Passo 4**: Repete-se os passos 2 e 3 até o momento em que os grupos se estabilizem;




(a) Passo 1 - Sorteio dos valores iniciais dos centróides

(b) Passo 2 - Atribuição dos objetos aos grupos cujo protótipo possua maior similaridade.

Figura 7 - Passo 1 e 2 do Algoritmo K-Means (Beltrame e Fonseca, 2010).




(c) Passo 3 - Recalcula os valores dos protótipos como sendo a média dos objetos atuais pertencentes aos grupos.

(d) Preparação para repetição dos passos 2 e 3 - Desvincula-se os objetos dos grupos.

Figura 8 - Passo 3 do Algoritmo K-Means (Beltrame e Fonseca,2010).

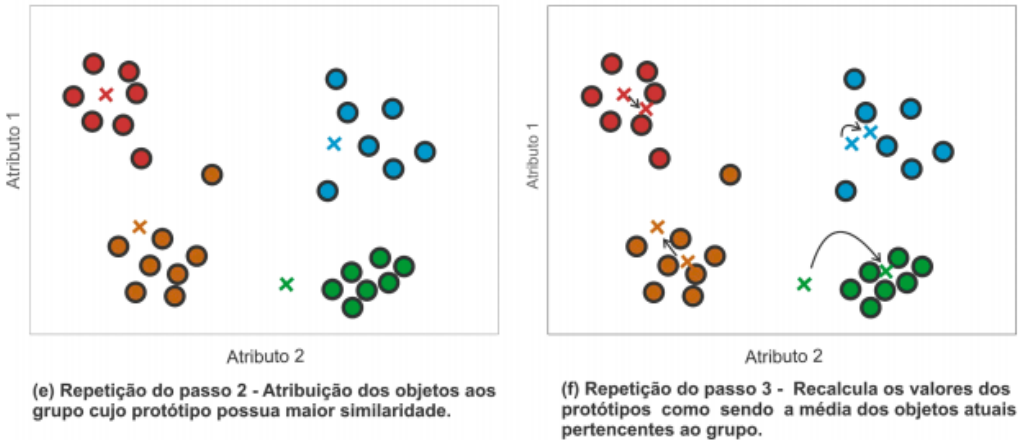



Figura 9 - Passo 4 (Repetição dos Passos 2 e 3) do Algoritmo K-Means (Beltrame e Fonseca, 2010).

O algoritmo K-means é popular devido a sua facilidade de implementação e sua ordem de complexidade linear, O(n) (Jain, Murty e Flynn, 1999).

#### 3.5.4.1 Avaliação da Qualidade de um Agrupamento

As medidas de avaliação de agrupamentos normalmente são dividas em dois grupos, que se diferenciam pela maneira que abordam o problema, o primeiro corresponde as medidas não-supervisionadas, que verificam a qualidade de um agrupamento sem utilizar medidas externas aos dados, para isso verificam a coesão e a separação conforme Figura 10, já a abordagem do segundo grupo é referente as medidas supervisionadas, que fazem uso de medidas externas aos dados para verificar a qualidade de um agrupamento, um exemplo é a utilização de um conjunto de dados de teste etiquetados com um atributo classe (Amo, 2012; Liu *et al.*, 2010).

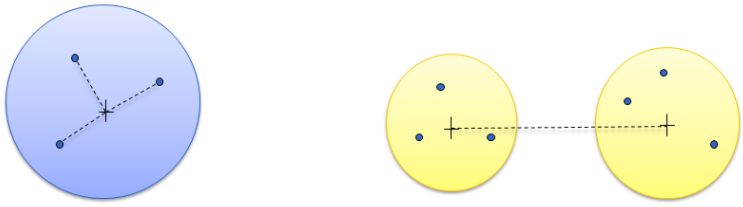



Figura 10 - Medidas de Avaliação: Coesão e Separação (Adaptado de Amo, 2012).

A coesão e separação podem ser consideradas para melhorar a clusterização, uma vez que um cluster com baixo grau de coesão pode ser dividido em 2 subclusters, e dois clusters que

apresentam boa coesão porém baixo grau de separação podem ser juntados para formar um único cluster (Amo, 2012).

Uma abordagem do tipo não supervisionada que pode ser utilizada para verificar a qualidade de um agrupamento é a Silhueta de Cluster – SC, (Rousseeuw, 1987). Na SC cada cluster é representado por uma silhueta que apresenta a maneira que objetos se posicionam dentro do cluster, podendo esses se posicionar bem (silhueta próximo a 1), de maneira intermediária ou com pouca ou quase nenhuma significância (próximo a 0), conforme ilustrado na Figura 11.

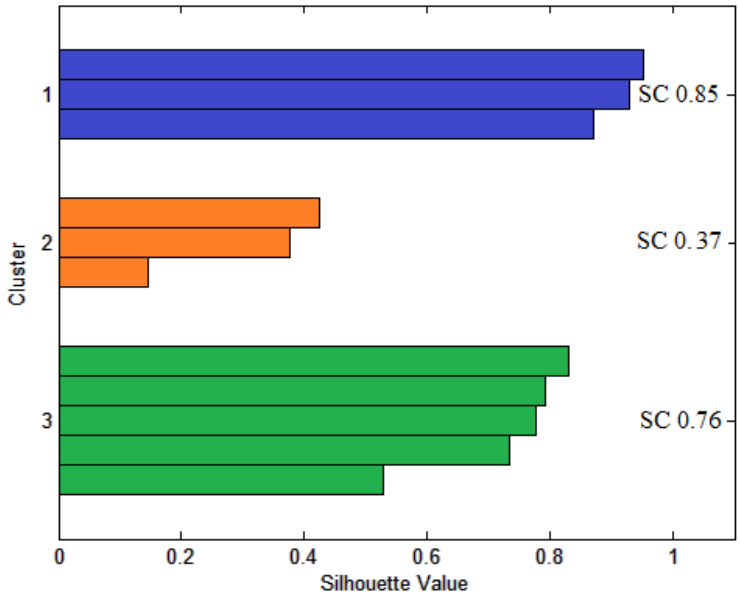



Figura 11 – Exemplo de Silhueta de Cluster para Três Clusters.

Para verificar o coeficiente de silhueta de um ponto $i$ de um determinado cluster devemos calcular $a$, que corresponde a média da distancia deste ponto $i$ até os outros pontos que compõem o cluster, e depois devemos calcular $b$, que é a distância média de $i$ até todos os pontos de outro cluster, desta maneira o coeficiente de silhueta de um ponto é dado por $s = 1 - a/b$ se $a < b$ ou de maneira não usual, $s = b/a - 1$ se $a \geq b$. Tipicamente o resultado é entre 0 e 1, sendo que quanto mais próximo for de 1 melhor (Liu *et al.*, 2010).

A média do coeficiente de silhueta de todos os pontos que compõem um determinado cluster é tida como a Silhueta do Cluster, e esse valor pode ser interpretado da seguinte maneira (Rousseeuw, 1987):

- **0.71 - 1.00**: Significa que uma estrutura forte foi encontrada;
- **0.51 - 0.70**: Corresponde a uma estrutura razoável;

- **0.26 - 0.50**: A estrutura encontrada é fraca e pode ser artificial;
- < **0.25**: Indica que nenhuma estrutura substancial foi encontrada.

### 3.5.5 Detecção de Anomalias

A detecção de anomalias ou raridades pretende identificar fenômenos não esperados, ou sem um antecedente similar, daí também conhecida como detecção de novidades.

É importante ter em mente que antes de classificar um determinado fenômeno como uma raridade, é necessário observar o tamanho da base de dados considerada, pois a ocorrência de 1% de determinado fenômeno pode ser considerado uma raridade em uma base de dados pequena, porém pode não ser em outra, composta por milhares de exemplos (Weiss, 2004).

### 3.5.6 Previsão

A previsão de séries temporais é uma das aplicações mais difundidas e importantes nessa área. Esse tipo de previsão por vezes exige um alto nível de acurácia em suas respostas, uma vez que importantes decisões podem ser baseadas nela. A previsão desempenha um papel importante em diversas áreas, dentre elas podem ser citadas: negócios, economia, gestão da operação, planejamento empresarial e políticas públicas (Liu, 2009).

Segundo Lütkepohl (2005), a previsão de um determinado momento $Z_{m+1}$ de uma série $Z = (Z_1, Z_2, \dots, Z_m)$ pode ser descrita segundo a equação:

$$Z_{m+1} = f(Z_m, Z_{m-1}, Z_{m-2}, \dots) \tag{3.3}$$

onde: $f()$ representa uma função de previsão que faz uso dos valores passados $Z_m, Z_{m-1}, Z_{m-2}$, ... para efetuar a previsão.

Um exemplo da aplicação de previsão de séries temporais é o das companhias aéreas comerciais, onde as previsões do volume futuro de passageiro são importantes para o órgão responsável pelo planejamento e pela administração de rotas aéreas, para os fabricantes de aeronaves, e por fim, para as companhias aéreas, para que possam definir o número de voos e aviões necessários para atender a demanda prevista (Liu, 2009).

## 3.6 Métricas de Dissimilaridade entre Séries Temporais

Grande parte das tarefas de mineração em séries temporais é dedicada a quantificar a semelhança entre séries temporais e subsequências, trabalho que requer o uso de uma função para medir a similaridade/dissimilaridade entre duas sequências (Alencar, 2007).

Segundo Agrawal *et al*. (1995) as principais críticas que se observam contra abordagens tradicionais, como a distância Euclidiana ou de Manhattan, dizem respeito a alta sensibilidade a ruídos, a translações, variações de fase entre as sequências, escalamentos horizontais entre outros.

Diversas técnicas surgiram com o objetivo de lidar com um ou mais dos aspectos negativos apresentados pelas abordagens tradicionais, por vezes, buscam ainda abordar o problema de maneiras diferenciadas (Drago e Varejão, 2007).

Atualmente existem diversas métricas de dissimilaridade, sendo a distância Euclidiana e o *Dynamic Time-Warping* - DTW duas das mais utilizadas por pesquisadores da área, métricas que são apresentadas a seguir. Também são comentadas métricas não convencionais.

### 3.6.1 Distância Euclidiana

Diversos trabalhos na área de séries temporais utilizam a distância Euclidiana - DE como medida de dissimilaridade. A distância Euclidiana pode ser medida apenas em séries que apresentem um mesmo número de mensurações, isso é, tenham o mesmo tamanho. O cálculo da DE consiste em verificar a distância entre pares de pontos de duas sequencias (Alencar, 2007).

Dadas duas séries temporais $P = (p_1, \dots, p_n)$ e $Q = (q_1, \dots, q_n)$ de mesmo tamanho $n$, a distância Euclidiana entre essas duas séries é definida como:

$$D_{istânciaEuclidiana}(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3.4}$$

A complexidade computacional do cálculo da distância Euclidiana é $O(n)$, porém caso se observe que um determinado limiar não tenha sido atendido durante a execução da sub-rotina da DE, é possível realizar um abandono precoce, conforme apresentado no parágrafo seguinte.

Quando utiliza-se a distância Euclidiana como sub-rotina de algoritmos de classificação ou de indexação, o interesse por vezes está em obter-se a distância exata enquanto ela permanecer abaixo de um limiar $r$. Caso seja verificado que o limiar $r$ não foi atendido durante o cálculo da DE, é possível realizar um abandono precoce (*Early Abandon*), desde que a soma das diferenças correntes ao quadrado entre cada par de pontos de determinadas séries temporais ultrapassem $r^2$, dessa maneira, o cálculo pode ser interrompido com a certeza de que a distância Euclidiana exata que havia sido calculada excederia $r$ (Keogh *et al.*, 2006).

### 3.6.2 Dynamic Time-Warping – DTW

A distância Euclidiana apesar de ser bastante simples de calcular pode produzir resultados errôneos em certos casos, como em séries temporais que apesar de similares, apresentam distorções no eixo do tempo (Alencar, 2007).

O DTW é uma técnica que busca encontrar um alinhamento ótimo entre duas sequências (dependentes no tempo), observando certas restrições. Intuitivamente e de maneira não linear as sequências são distorcidas para que coincidam com outra sequência, conforme ilustrado na Figura 12 (Müller, 2007).

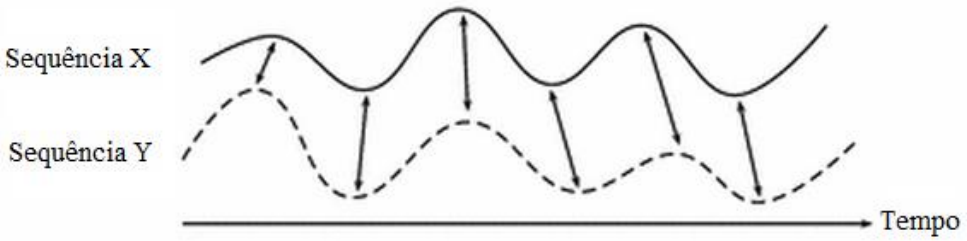



Figura 12 - Técnica DTW (Adaptado de Müller, 2007).

### 3.6.3 Métricas Não Convencionais

A técnica de Alinhamento de Strings é um exemplo de abordagem distinta, que se baseia na conversão de séries temporais para sequência distintas das originais, formadas por elementos de um determinado alfabeto. O cálculo da similaridade fica baseado no número de modificações necessárias para se igualar uma série com outra. Cada operação realizada para se igualar as séries pode receber distintos valores, entre as operações possíveis pode-se citar a edição, a soma e a remoção de caracteres (Theodoridis, Koutroumbas, 2006).

## 3.7 Intervalo de Confiança – Bootstrap

Um intervalo de confiança - IC descreve um conjunto ou intervalo de valores, dentro do qual é possível que esteja um parâmetro da população.

A técnica bootstrap trata uma amostra $X = (X_1, X_2, \dots, X_N)$ como se ela representasse toda a população. Para o cálculo do bootstrap, consideraremos que $X = (X_1, X_2, \dots, X_N)$ seja uma amostra que contenha $N$ observações, deve-se então construir $B$ amostras $X^{*(1)}, X^{*(2)}, \dots, X^{*(B)}$ independentes e identicamente distribuídas (iid) de comprimento $B$, que para técnica bootstrap, corresponde a amostrar com substituição a partir do conjunto $X$. Consideremos $\hat{\mu}$ a estimativa de uma variável aleatória $X$ (DiCiccio e Efron, 1996):

$$\hat{\mu} = \frac{X_1, X_2, \dots, X_N}{N} \tag{3.5}$$

para calcular o intervalo de confiança com bootstrap, pode-se seguir os seguintes passos:

- **Passo 1**: Verificar $N$ e a $\hat{\mu}$ de um conjunto de amostra $X$;
- **Passo 2**: Utilizando um gerador de números aleatórios selecionar $N$ amostras a partir de $X$;
- **Passo 3**: Obter a estimativa bootstrap $\hat{\mu}_1^*$;
- **Passo 4**: Repetir os passos 1 e 2 um número $B$ de vezes ($B$ deve ser um número grande, exemplo 1000) para obter as estimativas $\hat{\mu}_1^*, \hat{\mu}_2^*, \dots, \hat{\mu}_B^*$;
- **Passo 5**: Aproximação da distribuição de $\hat{\mu}^*$, devendo ordenar as estimativas por ordem crescente $\hat{\mu}_1^* \leq, \hat{\mu}_2^* \dots, \hat{\mu}_B^*$, $\hat{\mu}_k^*$ é o $k$ésimo menor valor de $\hat{\mu}_1^*, \hat{\mu}_2^*, \dots, \hat{\mu}_B^*$;
- **Passo 6**: O intervalo de confiança $(1-\alpha)100\%$ é dado por $(\hat{\mu}_{q1}^*, \hat{\mu}_{q2}^*)$ onde $q_1 = parte\_inteira(B\alpha/2)$ e $q2 = B - q_1 + 1$. Exemplo: para $\alpha = 0.005$ e $B = 1000$, $q_1 = 25$ e $q_2 = 976$.

## 3.8 Avaliação de Previsões

O índice de Erro Percentual Absoluto Médio – MAPE permite avaliar a assertividade de previsões.

Este índice é utilizado em diversos trabalhos e considerado por concessionárias de energia elétrica como uma forma padronizada de avaliar o desempenho de predições de carga. O MAPE é calculado de acordo com a Equação:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{3.6}$$

onde $A_t$ é o valor real, $F_t$ é o valor de previsão e $n$ é o tamanho da série temporal. O MAPE retorna um valor que deve ser pequeno, indicando maior precisão do método de previsão.

# Capítulo 4

# Cenário de Estudo e Técnicas de Previsão

## 4.1 Considerações Iniciais

Neste capítulo são apresentadas algumas características da codificação e do funcionamento dos algoritmos de previsão baseados no método de dias similares. Os dados de entrada, os processos envolvidos e as saídas esperadas são caracterizadas. A tarefa de clusterização de dados utilizada para permitir a redução do conjunto original é descrita e comentada. Por fim, o algoritmo utilizado para a representação gráfica e a avaliação de resultados é apresentado. Os testes realizados e os resultados obtidos são apresentados no próximo capítulo.

## 4.2 Cenário de Estudo

O Parque Tecnológico Itaipu - PTI é um polo científico e tecnológico criado pela Itaipu Binacional em 2003. O PTI está instalado nos antigos alojamentos dos operários que construíram a usina de Itaipu, na cidade de Foz do Iguaçu - PR, onde reúne diversas entidades públicas e privadas, com destaque para as instituições de ensino e as empresas na área de tecnologia. O parque está situado a 180 metros acima do nível do mar, com clima subtropical temperado super úmido. Nesse cenário se insere uma microrrede inteligente de onde são provenientes os dados utilizados para os estudos

As medições das grandezas elétricas provenientes dos blocos da microrrede do PTI são realizadas por medidores eletrônicos. Os dados são coletados e transportados por uma rede

dedicada de par trançado blindado (*Shielded Twisted Pair* - STP) até um servidor, onde são armazenados.

Dados de consumo estão sendo obtidos desde abril de 2012, entretanto, devido a problemas técnicos, a medição tornou-se continua a partir da segunda metade daquele ano.

As aferições são realizadas com intervalos de quinze minutos, de tal modo que noventa e seis aferições compõem uma curva de carga diária.

As curvas diárias de consumo foram divididas de acordo com as estações do ano e dias da semana, sendo apresentadas as que compreendem o inverno do ano de 2012 e 2013.

Para o presente trabalho foram considerados os dados provenientes do bloco em que está instalado o restaurante do PTI, destacado na Figura 13. Iniciando o atendimento ao público as 11:30h e encerrando as 14:00h.

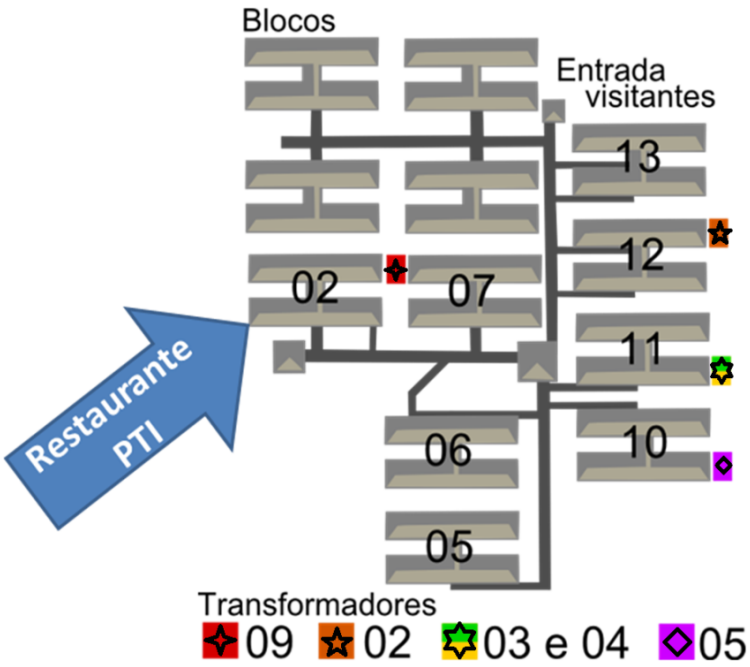



Figura 13 - Blocos e Transformadores do PTI.

O restaurante serve diariamente aproximadamente 800 refeições, sendo que essas variam para mais quando se realiza algum evento no PTI, e para menos na temporada de férias universitárias. A louça é lavada manualmente, com exceção aos pratos, limpos em lava-louças elétrico. A área destinada aos consumidores é superior a 1.200 m², sendo climatizada no verão.

O restaurante opera com um efetivo entre 30 a 40 funcionários, das mais distintas áreas, cozinheiros, auxiliares administrativo, caixas, equipe de limpeza, entre outros. O expediente inicia entre 8:00 e 8:30h, e finaliza entre 17:30 e 18:00h.

Juntamente ao bloco do restaurante opera uma agência bancária de porte pequeno, com aproximadamente 4 funcionários, e com expediente convencional entre 10:00 e 15:00h. No mesmo bloco funciona uma agência dos correios, com dois funcionários e horário de atendimento entre 08:00 e 17:30h, fechando para almoço entre 12:00 e 13:30h. Existe ainda um café, com horário de atendimento entre 8:00 e 18:00 h.

## 4.3 Algoritmo de Previsão: Método de Dias Similares

O algoritmo toma por base abordagens convencionais do método de dias similares, e faz uso do conjunto total de curvas de carga, formado por aferições entre 2012 e 2013.

Como entradas o algoritmo espera receber informações relativas a estação do ano, o tipo de dia (dias da semana ou feriado) e a temperatura prevista para o dia que se quer prever, além destas informações, é possível inserir o crescimento médio anual do consumo de energia elétrica.

Uma vez que as informações de entrada são inseridas, o algoritmo busca as curvas de carga que estejam de acordo com a estação do ano e o dia da semana pretendido. Aos dados relativos ao ano anterior ao corrente, aplica-se o crescimento médio anual do consumo de EE.

O algoritmo calcula a temperatura média de todos os dias selecionados. A temperatura destes dias é composta a partir de aferições igualmente espaçadas e concomitantes com aquelas relativas ao consumo de energia ativa, sendo assim, 96 aferições da temperatura descrevem a evolução desta no decorrer de um dia.

Alguns parâmetros internos do algoritmo devem ser ajustados para que os dias a serem considerados na previsão sejam mais ou menos parecidos com a entrada relativa à temperatura média prevista, como padrão, o algoritmo admite uma variação de 10 por cento para mais ou para menos.

Dentre as curvas que atenderam os parâmetros anteriores do algoritmo, admite-se uma determinada quantidade das mais similares. Esta quantidade pode ser ajustada a partir de parâmetros no algoritmo, sendo que, uma vez que não estejam disponíveis curvas suficientes para atender esta etapa, os parâmetros relativos à temperatura média são relaxados, fazendo com que mais curvas sejam admitidas, e dentre estas as mais similares são selecionadas.

A distância euclidiana foi utilizada para verificar a similaridade entre as curvas, para tanto, todas as curvas foram comparadas entre si, sendo que as menores distâncias euclidianas correspondem a aquelas mais similares.

Uma vez que as curvas similares foram selecionadas, a obtenção da média simples entre elas é retornada como previsão. A Figura 14 apresenta o fluxograma simplificado do funcionamento do algoritmo.

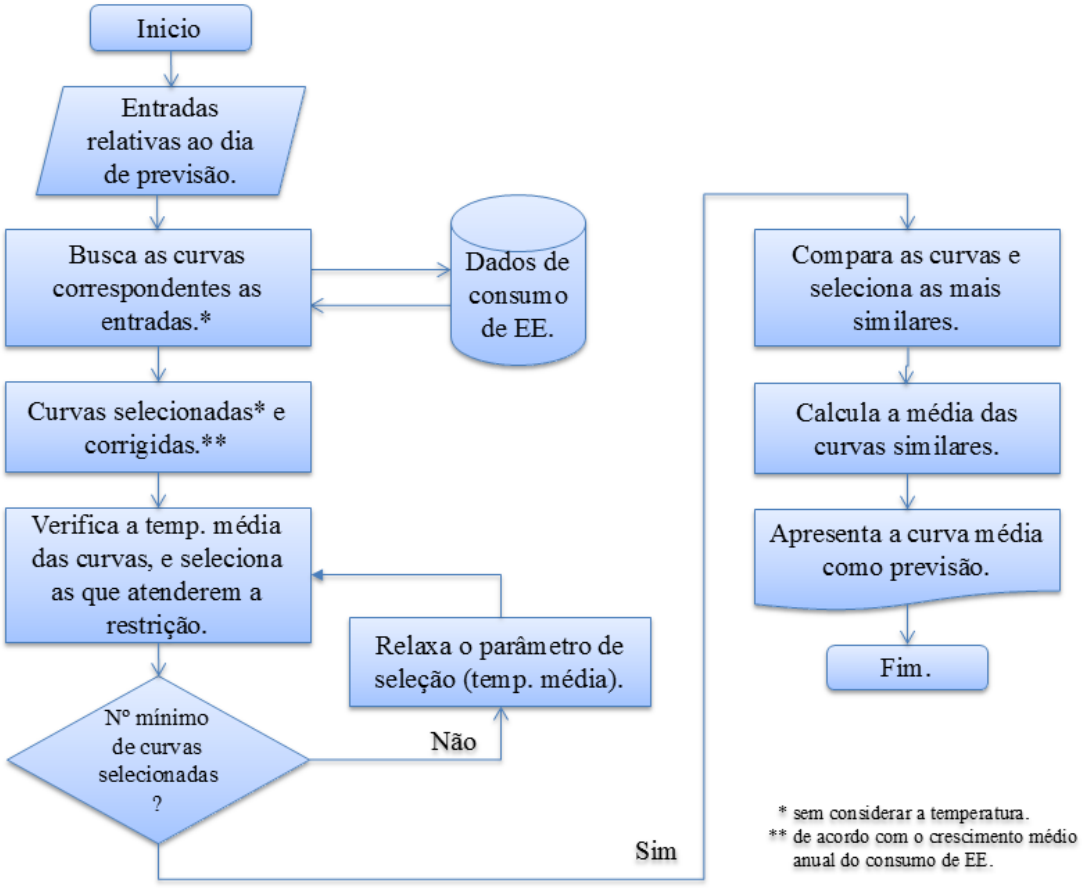



Figura 14 - Fluxograma do Funcionamento do Algoritmo.

## 4.4 Algoritmo de Previsão: Método de Dias Similares com Curvas de Carga Clusterizadas

O MDS aqui apresentado é antecedido por tarefas de clusterização e avaliação das curvas de carga, fazendo com que a execução do algoritmo corresponda a uma tarefa de busca em uma base reduzida de dados, atendendo determina parâmetros de entrada.

Como entradas o algoritmo espera receber a estação do ano, o dia da semana (ou feriado)

e a temperatura média prevista para o dia de previsão.

O índice de crescimento anual do consumo de EE é utilizado na tarefa de pré-processamento das curvas de carga, para que seja aplicado aos dados do ano anterior ao último empregado.

Na tarefa de pré-processamento, as curvas de carga de uma determinada estação do ano são divididas por dia da semana para serem submetidas a tarefas de clusterização, utilizando o algoritmo *Simple K-means* com a distância euclidiana.

Para as tarefas de clusterização os dados foram divididos entre os anos de 2012 e 2013, de acordo com o tipo de dia e com a estação do ano que pertenciam, (todas as segundas do inverno foram submetidas a tarefas de clusterização, depois todas as terças dessa mesma estação, e assim por diante atingindo todos os tipos de dias e as quatro estações do ano) considerou-se ainda dois cenários, sendo que, no primeiro foi considerada apenas a característica de temperatura media para realizar a clusterização, no segundo cenário além da temperatura foi considerado o valor mínimo, o máximo, a média e o desvio padrão de cada curva de carga.

Após serem criados os clusters que representam as curvas de carga, foram executadas tarefas de comparação entre eles, com o objetivo de encontrar clusters semelhantes, que pudessem ser representados por apenas uma curva de carga, reduzindo ainda mais o conjunto final.

Para que as curvas de carga pudessem ser submetidas a tarefas de comparação, foi necessário calcular seus mínimos, suas médias e seus máximos, que permitiram uma pré-triagem rápida, para que uma quantidade menor de comparações fossem realizadas. De maneira a permitir que apenas a forma (*shape*) das curvas fossem comparadas, seus dados foram normalizados.

Todas as curvas normalizadas são comparadas entre si, com o objetivo de se verificar a distância absoluta entre elas. As que possuem menores distâncias absolutas (dentro de uma faixa estabelecida) são verificadas quanto a seus mínimos, médias, máximos e desvios padrão, sendo que se forem semelhantes dentro de uma faixa de *Y* (5%) por cento (considerando todos os pontos pares entre duas séries temporais – curvas de carga) são tidas como bastante semelhantes, e a média simples entre elas passa a representa-las no algoritmo. O valor de cinco para *Y* foi estabelecido com base em experimentos piloto para o conjunto de dados considerado, devendo ser adequado para outras situaçõe . Figura 15 apresenta o fluxograma simplificado do funcionamento do algoritmo de pré-processamento das curvas de carga, responsável pela

clusterização dos dados.

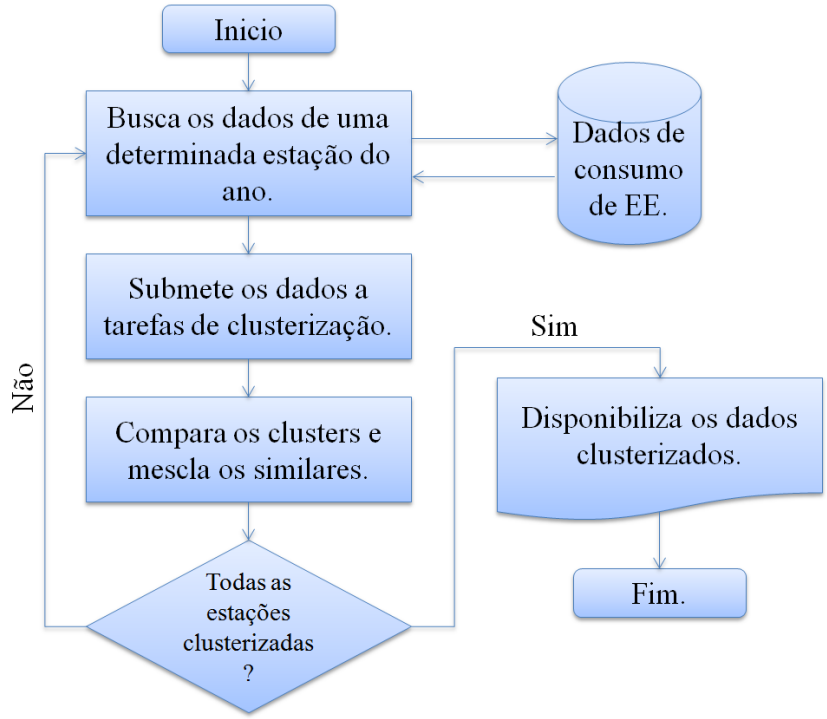



Figura 15 - Fluxograma do Algoritmo de Pré-processamento (clusterização).

Com base nas entradas do usuário, o algoritmo busca em uma base de curvas pré-processadas, a que apresenta temperatura média mais próxima a aquela informada pelo usuário (em caso de empate, escolhe randomicamente entre elas). Uma vez tendo selecionado uma curva, o algoritmo permite corrigí-la de acordo com a última curva registrada para o mesmo dia da semana, para tanto, a influência desta deve ser ajustada no algoritmo, podendo ir de zero (nenhuma influência) a 100 (substituindo a curva selecionada pela última registrada), para o presente trabalho, tomando por base experimentos piloto e comparando os resultados obtidos, o valor de influência foi fixado em 15.

Um diferencial aplicado as respostas baseadas em clusters, é o cálculo de seu intervalo de confiança – IC (por padrão utiliza-se 95%, valor que pode ser ajustado para mais ou para menos no algoritmo). Devido a pequena quantidade de curvas que compõem cada cluster, foi adotada a técnica de bootstrap para o cálculo do IC, uma vez que esta técnica se mostra especialmente útil em casos em que o número de amostras é reduzido.

A Figura 16 apresenta o fluxograma básico do funcionamento do algoritmo de previsão, sem abordar a etapa de pré-processamento, previamente apresentada.

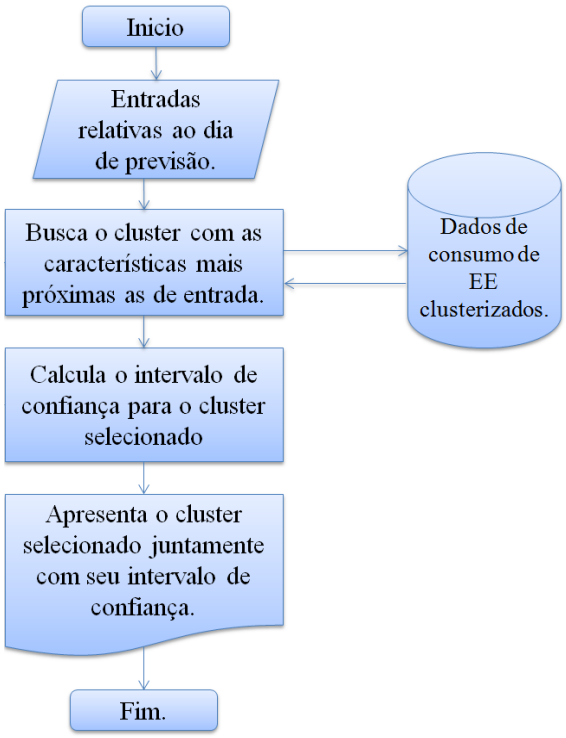



Figura 16 - Fluxograma do Algoritmo de Previsão.

## 4.5 Algoritmo de Apresentação e Avaliação de Previsões

É responsável por apresentar graficamente e calcular a assertividade dos resultados apresentados pelos algoritmos de previsão. Para tanto, captura as respostas apresentadas pelos referidos algoritmos e as compara com a curva real verificada para o determinado dia.

Como entradas o algoritmo espera receber o dia da semana referente à previsão, sendo que esta informação será utilizada para selecionar a curva correspondente verificada (curva de carga real constatada), permitindo assim a realização de testes, comparando as respostas dos algoritmos de previsão com as curvas reais verificadas para outros tipos de dias, permitindo constatar o quão ajustada a curva de previsão se mostra para um determinado tipo de dia.

O cálculo do MAPE é utilizado para avaliar a assertividade das previsões, sendo apresentado juntamente com o esboço gráfico da curva prevista e a real verificada, e ainda, no caso da utilização dos dados clusterizados, é exibido o limite superior e inferior do intervalo de confiança calculado. A Figura 17 apresenta o fluxograma do algoritmo de avaliação de previsão e apresentação de resultados.

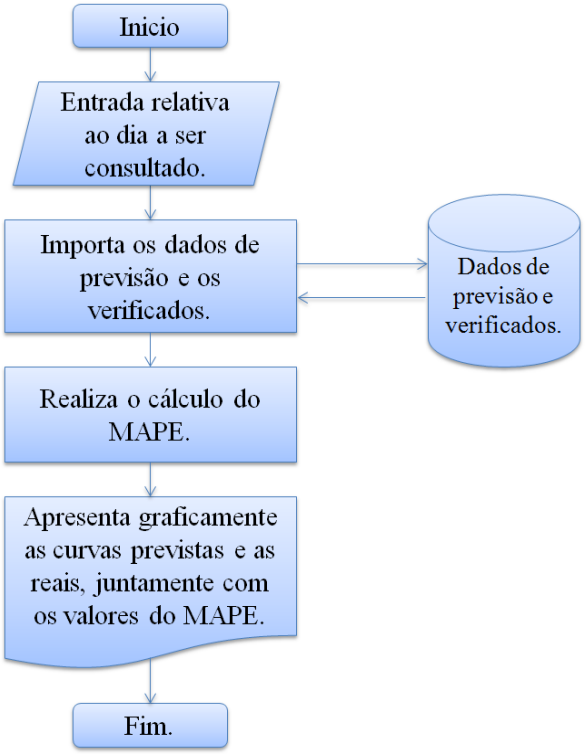



Figura 17 - Fluxograma do Algoritmo de Apresentação e Avaliação de Previsões.

## 4.6 Considerações Finais

Ao comparar os algoritmos descritos no presente capítulo com outros também baseados no MDS, é possível que se verifiquem diferenças, por vezes explicadas pela necessidade de se adaptar-se o algoritmo aos diferentes cenários de estudo e a restrita disponibilidade de dados.

A utilização da distância Euclidiana em detrimento da distância Manhattan foi estabelecida após testes, em que os clusters construídos a partir da segunda distância apresentaram resultados inferiores, para a avaliação foram verificadas a qualidade da previsão que permitiam e também a qualidade dos clusters, utilizando para isso, a avaliação da Silhueta de Cluster.

O principal diferencial está na clusterização das curvas de carga e na apresentação de um intervalo de confiança para cada previsão, permitindo que a carga de processamento imposta pelo segundo algoritmo (de previsão) seja menor, e ainda suas respostas mais completas, uma vez que sua interpretação pode levar em conta o IC.

# Capítulo 5

# Testes e Resultados

## 5.1 Considerações Iniciais

Este capítulo aborda questões relacionadas a qualidade dos agrupamentos realizados, observações quanto a real representatividade dos clusters, além de verificações quanto aos níveis de assertividade obtidos com as implementações dos métodos de dias similares, sendo o MAPE escolhido para realizar tal indicação.

Em relação a alguns resultados apresentados neste capítulo, sobretudo aqueles referentes a qualidade das previsões, salienta-se a natureza diversa dos dados utilizados, originários de um nível de consumo menos agregado, no qual os índices de incerteza são consideravelmente aumentados. De tal maneira, a tarefa de prever o comportamento de um consumidor específico não é trivial, e os índices de assertividade por vezes se fazem aumentados.

É importante destacar que o objetivo principal deste trabalho foi apresentar o uso de clusterização de dados, e não realizar previsões de carga, uma vez que a técnica pode ser utilizada com outros métodos de previsão.

## 5.2 Ambiente de Desenvolvimento e de Testes

A codificação dos métodos apresentados foi realizada no ambiente interativo para computação numérica MATLAB® versão 7.10.0, nome que também designa a linguagem utilizada.

Para permitir a visualização, a análise computacional e estatística dos dados utilizados nos métodos de previsão, além do MATLAB® utilizou-se o pacote de software Weka, versão 3.6.

A implementação e os testes foram realizados num mesmo computador, configurado com Windows 7 Home Premium (SPK 1), processador AMD Athlon (tm) II P340 Dual-Core 2.20GHZ e 4GB de memória RAM.

## 5.3 Análise das Curvas de Carga

Uma tarefa que deve anteceder a de previsão é a verificação e validação dos dados a serem utilizados. Recursos gráficos permitem que os dados sejam apresentados de maneira que facilitem sua interpretação, possibilitando a identificação de possíveis padrões e anomalias.

A Figura 18 apresenta as curvas referentes ao inverno de 2012, agrupadas de acordo com os dias da semana. Ao observar as curvas de carga agrupadas por dias da semana é possível identificar alguns dias que se apresentam como atípicos, um gesto instintivo seria retirá-los, entretanto antes de se tomar qualquer decisão é necessário verificar de maneira aprofundada tais dias.

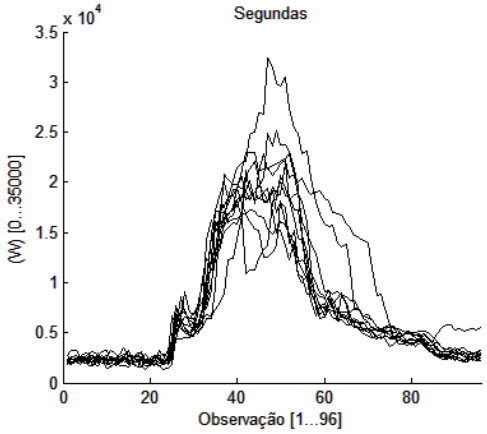



Figura 18a – Curvas de Carga: Segundas.

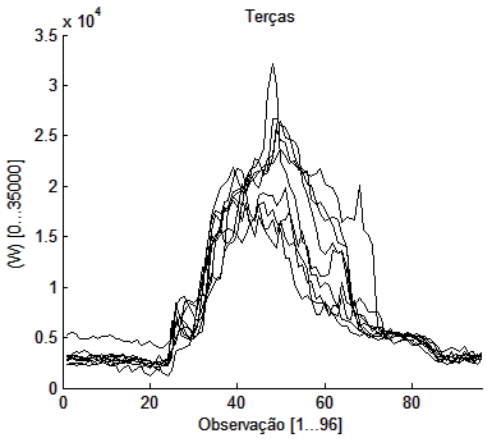



Figura 18b – Cur . _ de Carga: Terças.

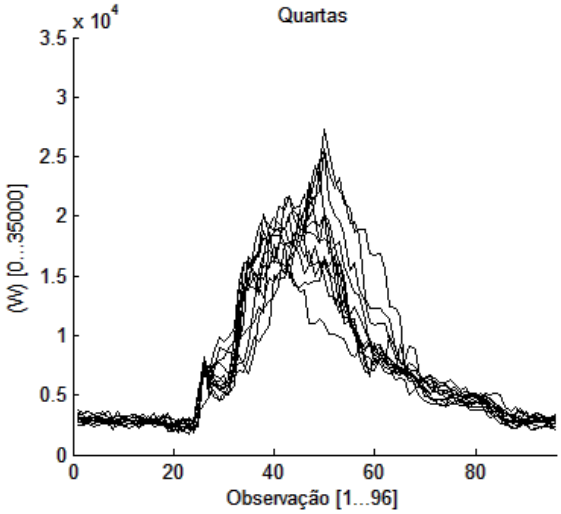



Figura 18c – Curvas de Carga: Quartas.

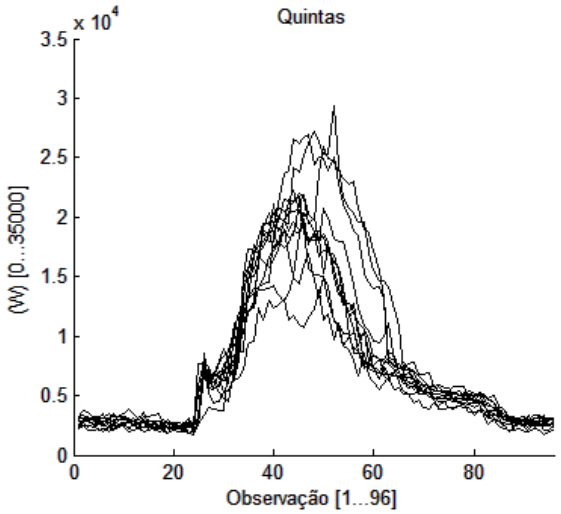



Figura 18d – Curvas de Carga: Quintas.

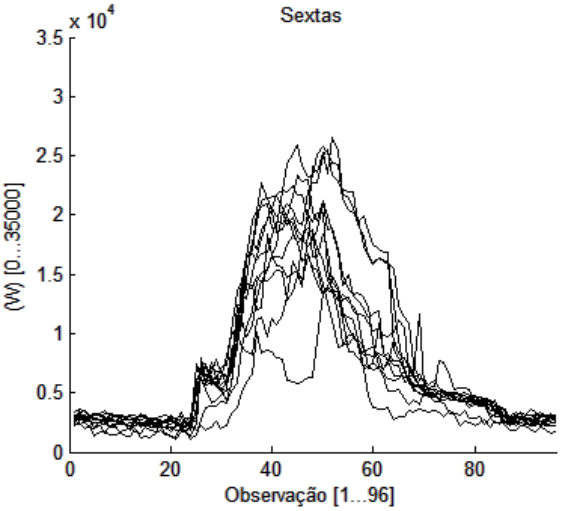



Figura 18e – Curvas de Carga: Sextas.

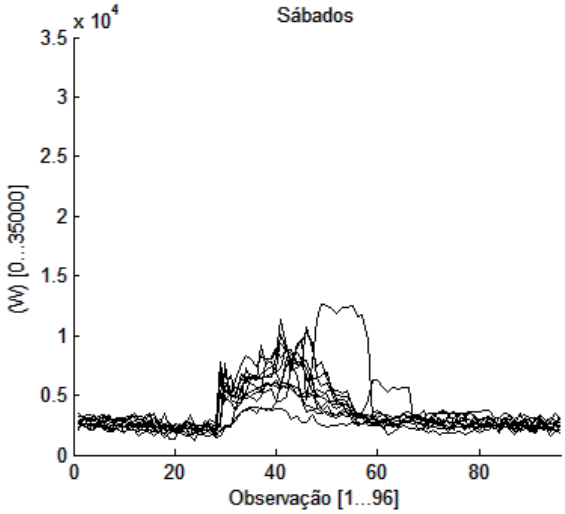



Figura 18f – Curvas de Carga: Sábados.

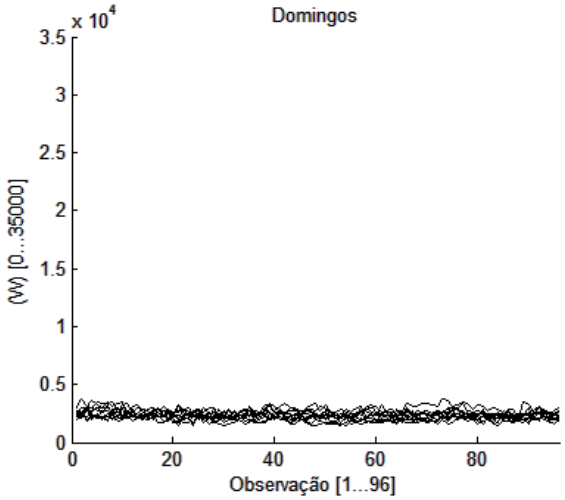



Figura 18g – Curvas de Carga: Domingos.

A ocorrência de eventos festivos, feriados, paralisações, mudanças bruscas de temperatura entre outros episódios, podem ionar variações que, à princípio, poderiam parecer atípicas. Atentar para estas questões permite evitar que dados consistentes sejam descartados.

Uma vez que as curvas de carga tenham sido analisadas, e possíveis valores incorretos tenham sido corrigidos, a média das curvas agrupadas de acordo com os dias da semana pode ser utilizada para representar os referidos conjuntos de dias, conforme ilustrado na Figura 19.

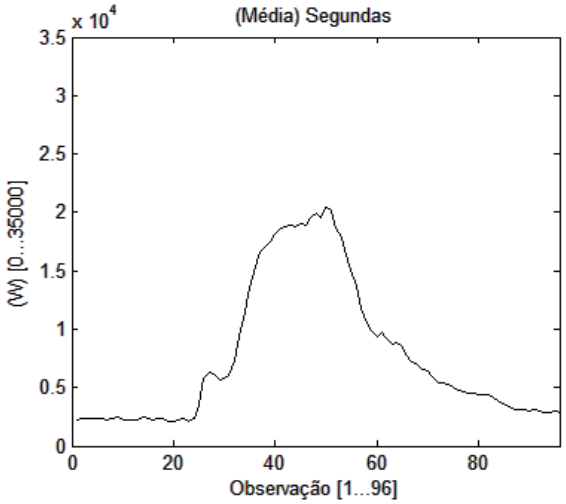



Figura 19a – Curva Média: Segundas.

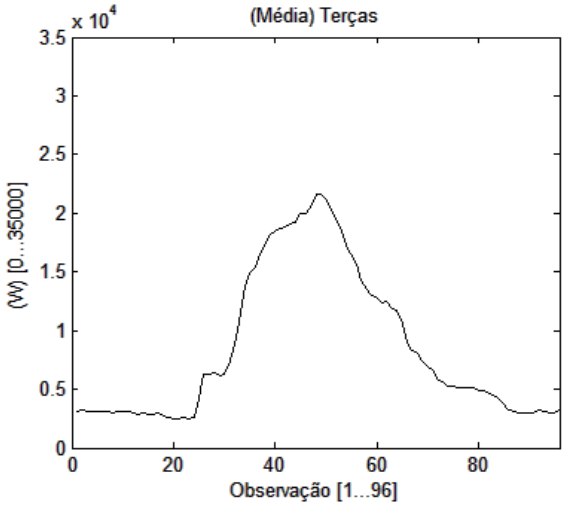



Figura 19b – Curva Média: Terças.

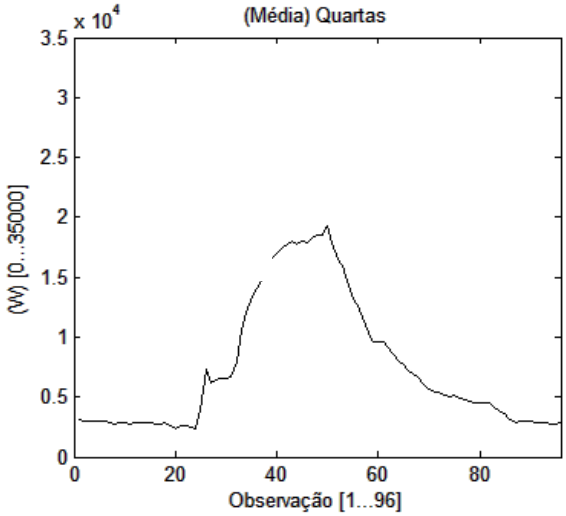



Figura 19c – Curva Média: Quartas.

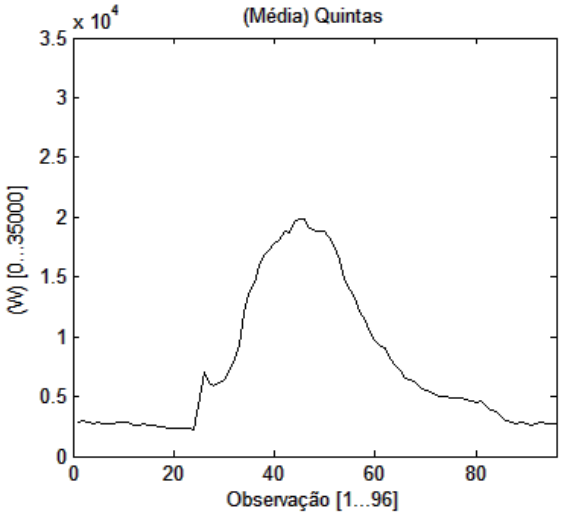



Figura 19d – Curva Média: Quintas.

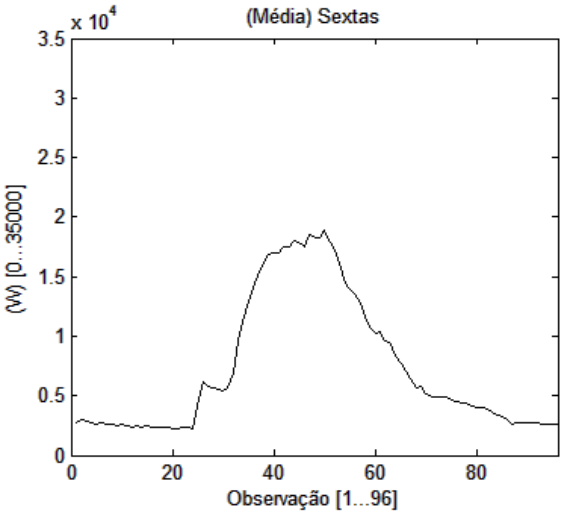



Figura 19e – Curva Média: Sextas.

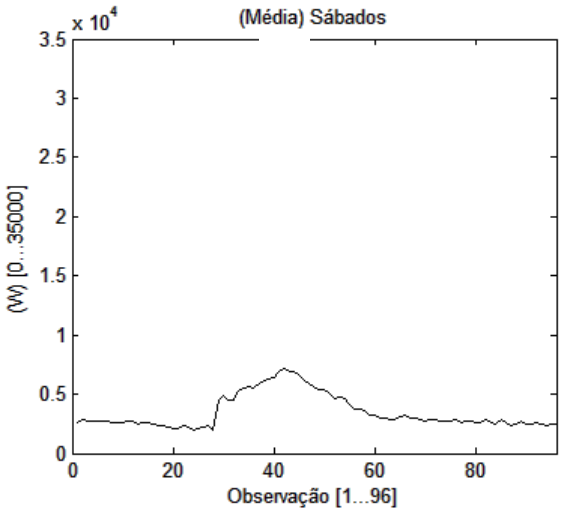



Figura 19f – Curva Média: Sábados.

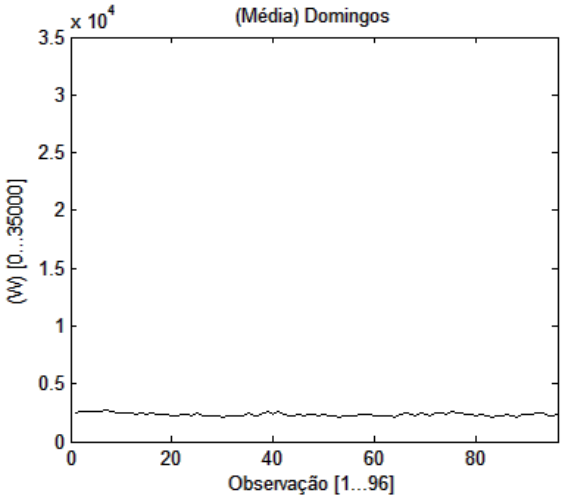



Figura 19g – Curva Média: Domingos.

A Figura 19 permite verificar uma característica abordada neste trabalho, a de que em níveis mais agregados ocorre a suavização das curvas de carga, essa característica pode ser percebida ao se comparar a Figura 19 com a Figura 18. É importante observar que as curvas médias para os dias úteis são bastante similares, característica que não é facilmente verificada nem entre as curvas de um mesmo tipo de dia apresentado na Figura 18. Tal característica permite supor que, utilizando os recursos apresentados no presente trabalho, seria possível obter resultados mais assertivos (com menores MAPEs) caso fossem utilizados dados do nível mais agregado.

## 5.4 Validação de Clusters

Com o objetivo de validar a qualic dos agrupamentos utilizados nos algoritmos de previsão, e verificar a melhor abordagem para clusterização, foi utilizada a medida não supervisionada Silhueta de Cluster - SC, que deve retornar valores positivo e próximos de 1 para indicar alto nível de coesão na estrutura formada.

Mais precisamente, segundo Rousseeuw (1987), valores menores que 0.25 indicam que nenhuma estrutura significativa foi encontrada, valores entre 0.26 e 0.50 indicam estruturas fracas e potencialmente artificiais, enquanto valores entre 0.51 e 0.70 correspondem a uma estrutura razoável, por fim, valores entre 0.71 e 1 indicam a ocorrência de estruturas fortes.

Para esta versão do trabalho a quantidade de clusters gerados para cada tipo de dia foi fixada em três, uma vez que os resultados obtidos com números maiores de clusters não foram

expressivos, melhorando os resultados em média menos que 5%, portanto, resultando em MAPEs similares aos encontrados utilizando o conjunto formado por três clusters por tipo de dia.

Para permitir a avaliação dos agrupamentos utilizados nas tarefas de previsão, foram calculados os valores das SC, além de serem gerados os gráficos correspondentes a cada tipo de dia. A clusterização baseada em múltiplos parâmetros (máximo, mínimo, média, desvio padrão e temperatura média) para a representação das curvas de carga apresentou os valores de SC mais altos, apontando a ocorrência de dezesseis estruturas fortes, quatro razoáveis e apenas uma fraca, conforme a Figura 20.

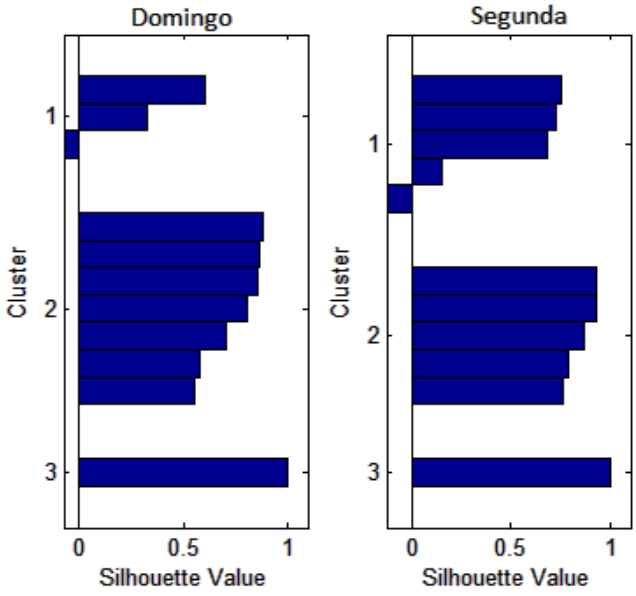



Figura 20a - Gráfico de SC (Múltiplos Parâmetros) – Domingo e Segunda.

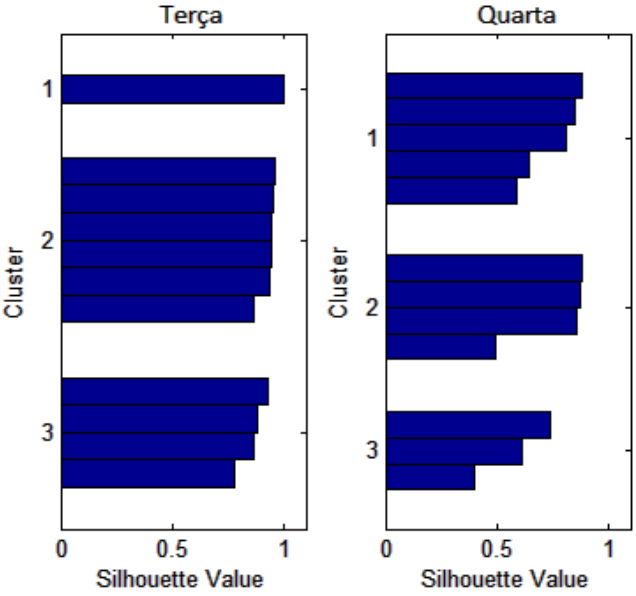


Figura 20b - Gráfico de SC (Múltiplos Parâmetros) – Terça e Quarta.

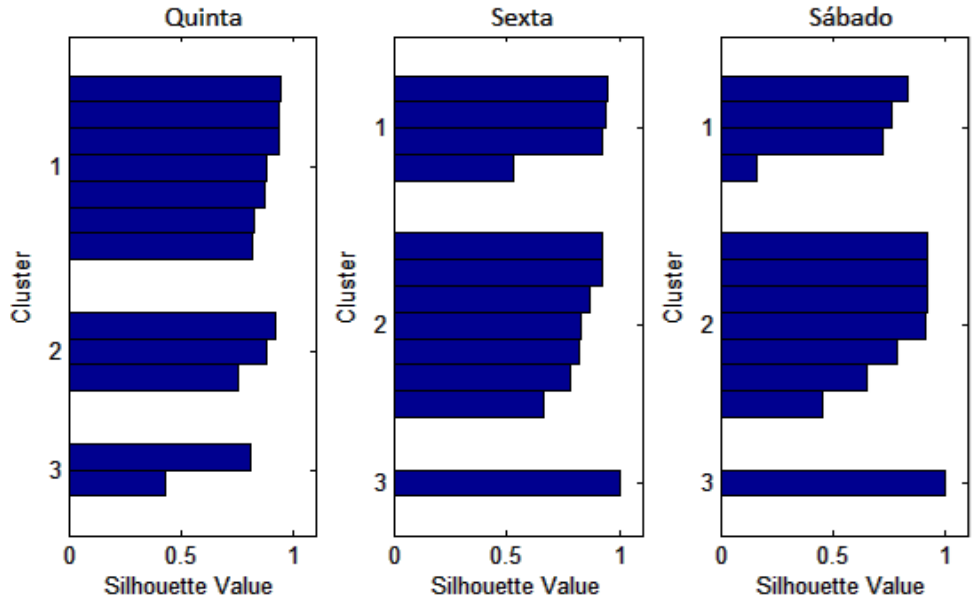



Figura 20c - Gráfico de SC (Múltiplos Parâmetros) – Quinta, Sexta e Sábado.

A clusterização baseada apenas na temperatura obteve os piores resultados, onde dez dos vinte e um clusteres se mostram sem nenhuma estrutura significativa, outros cinco se mostraram fracos e potencialmente artificiais e seis se apresentaram fortes. É importante destacar que dos seis que se apresentaram fortes apenas um era formado por mais que um objeto, uma vez que, sempre que um objeto compõem de maneira isolada um cluster, seu valor de SC é um, a Figura 21 apresenta as SC para os dias da semana.

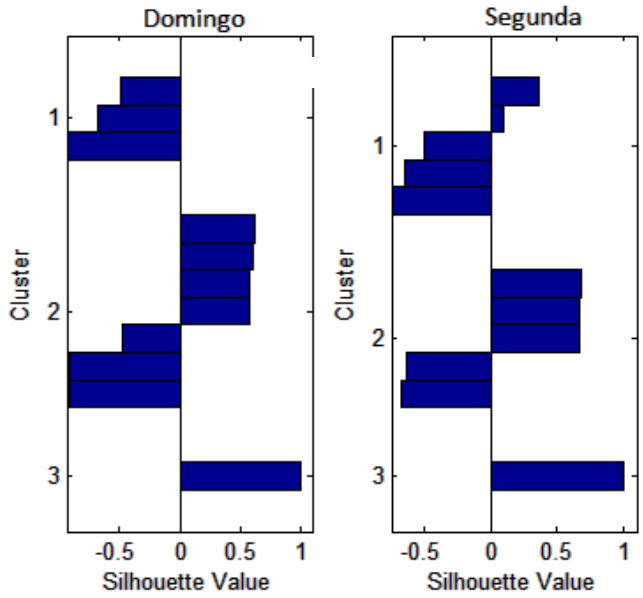



Figura 21a - Gráfico de SC (Temperatura) Domingo e Segunda.

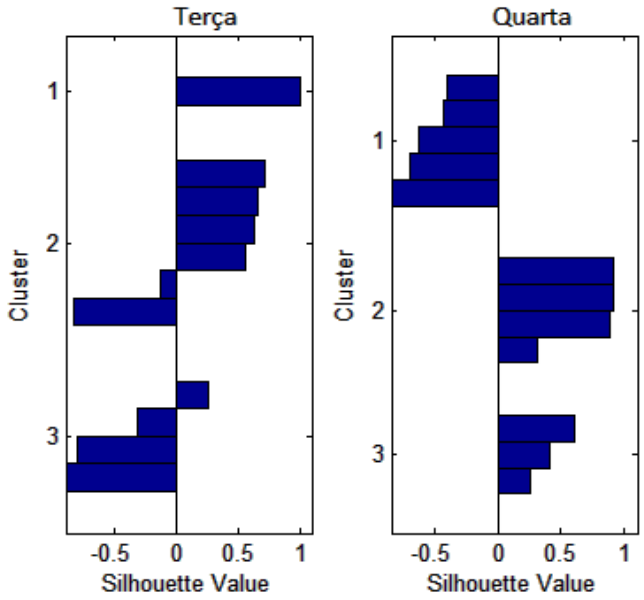



Figura 21b - Gráfico de SC (Temperatura) Terça e Quarta.

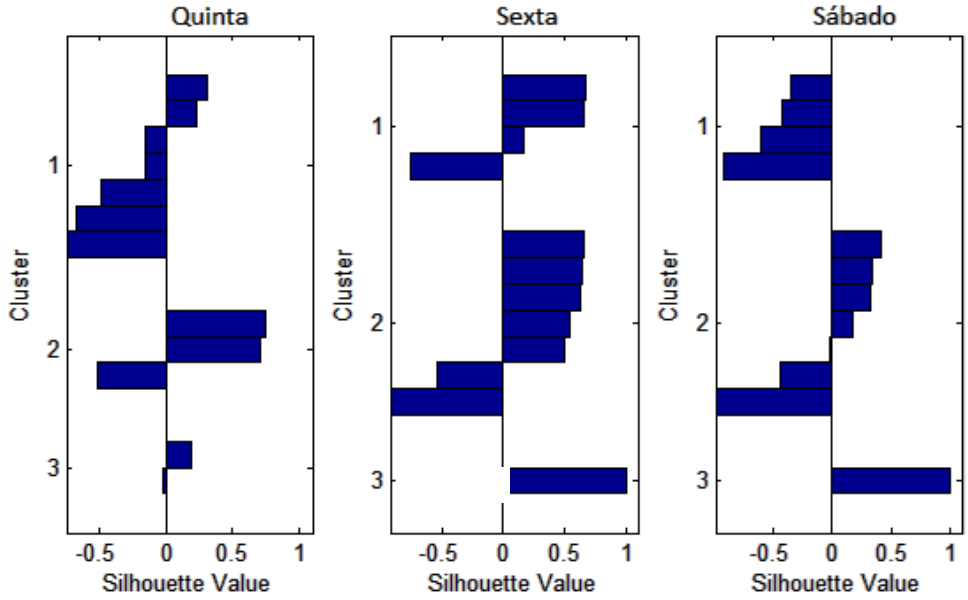



Figura 21c - Gráfico de SC (Temperatura) Quinta, Sexta e Sábado.

Ao observar a Figura 20 e a Figura 21, nota-se que 10 dos 42 clusters são formados por apenas um objeto, situação que sinalizou a possibilidade da existência de dados incorretos, para sanar tais dúvidas, as curvas foram reanalisadas individualmente e no contexto geral, utilizando para isso recursos gráficos, constatou-se, porém, que as curvas não apresentavam indícios de *outliers*, e tratavam-se sim, de curvas particularmente distintas.

Os valores de SC obtidos são apresentados na Tabela 2, sendo que aqueles correspondentes a abordagem de clusterização baseada em múltiplos parâmetros são destacados

em negrito (Abord. 1), e os resultantes da clusterização com base na temperatura são apresentados em itálico (Abord. 2).

Tabela 2 - Valores Obtidos para a Silhueta de Cluster.

| **Medida de Silhueta de Cluster** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dia | | Domingo | Segunda | Terça | Quarta | Quinta | Sexta | Sábado |
| Abord. 1 | Cluster 1 | **0,46** | **0,70** | **1** | **0,74** | **0,87** | **0,72** | **0,70** |
| | Cluster 2 | **0,81** | **0,79** | **0,88** | **0,74** | **0,83** | **0,82** | **0,79** |
| | Cluster 3 | **1** | **1** | **0,86** | **0,60** | **0,66** | **1** | **1** |
| Abord. 2 | Cluster 1 | *-0,65* | *-0,61* | *1* | *-0,67* | *-0,46* | *0,42* | *-0,37* |
| | Cluster 2 | *0,01* | *0,20* | *0,32* | *0,78* | *0,33* | *0,49* | *0,08* |
| | Cluster 3 | *1* | *1* | *-0,60* | *0,49* | *0,19* | *1* | *1* |

Conforme é possível verificar na Tabela 2, a abordagem que utiliza vários parâmetros para representar as curvas de carga a serem clusterizadas obteve os melhores resultados, indicando que tal estratégia é potencialmente boa para clusterizar este tipo de dados. Em contrapartida, a clusterização baseada apenas na temperatura apresentou resultados bastante ruins, no qual apenas uma estrutura (não formada por um único objeto) foi classificada como forte, indicando assim, uma capacidade pobre de representar isoladamente este tipo de dados, não sendo indicada para realizar tal tarefa.

## 5.5 Ensaios de Previsão

Este subcapítulo descreve como foram organizados, realizados e apresentados os testes e resultados.

Foram criados quatro grupos que representam as previsões realizadas, o primeiro foi denominado GA, responsável por retornar uma curva de carga média como previsão, sendo esta formada por todas as curvas de um mesmo tipo de dia em uma determinada estação do ano, já os resultados do GB são aqueles retornados pela implementação do método de dias similares convencional, que faz uso do conjunto completo de curvas de carga, por sua vez, o GC corresponde aos resultados obtidos por meio do MDS que faz uso das curvas de carga clusterizadas de acordo com a temperatura média, o máximo, o mínimo, a média e o desvio padrão, por fim, o grupo GD corresponde ao que utiliza as curvas de carga clusterizadas de acordo com suas temperaturas médias. A Tabela 3 sintetiza os grupos de previsão.

Tabela 3 – Grupos de Previsão.

| **Grupos de Previsão** | |
|---|---|
| Grupo | Descrição |
| GA | Curva média de um mesmo tipo de dia |
| GB | MDS considerando o conjunto total de curvas de carga |
| GC | MDS considerando as curvas de carga clusterizadas de acordo com múltiplos parâmetros |
| GD | MDS considerando as curvas de carga clusterizadas de acordo com a temperatura média. |

Inicialmente excluiu-se do conjunto total de curvas de carga aquelas referentes à última semana registrada (inverno de 2013) posteriormente foram executadas as tarefas de clusterização de dados, de maneira que o objetivo passou a ser prever as curvas de carga que foram excluídas. Apesar de se ter acesso aos dados de temperatura dos dias excluídos, estes foram ignorados, e todas as informações utilizadas foram as disponibilizadas no sítio da web do Instituto Tecnológico SIMEPAR, agência paranaense que tem como uma de suas finalidades prover à sociedade informações de natureza meteorológica.

Os resultados obtidos pelos quatro g os foram confrontados com aqueles reais, previamente excluídos do conjunto original, para então serem analisados e apresentados em forma de tabela, com seus respectivos valores de MAPE.

No caso dos grupos GC e GD, que fazem uso dos dados agrupados, a escolha do cluster que será apresentado como previsão é realizada com base na temperatura média informada como entrada e a verificada para cada cluster, sendo que aquele com a temperatura média mais próxima a informada é selecionado.

## 5.6 Tempos Computacionais

Durante a execução dos algoritmos de previsão foram verificados os tempos decorrentes a partir da entrada de dados até a apresentação dos resultados, foram realizados três testes com os mesmos parâmetros de entrada para cada algoritmo, os tempos registrados são apresentados na Tabela 4.

Tabela 4 – Tempos Computacionais para Previsão (s).

| **Tempos Computacionais (s)** | | | | | |
|---|---|---|---|---|---|
| Grupo | Teste 1 | Teste 2 | Teste 3 | Média | Desvio Padrão |

| | | | | | |
|---|---|---|---|---|---|
| GA | 0,043 | 0,039 | 0,046 | 0,0426 | 0,0028 |
| GB | 7,141 | 6,782 | 6,762 | 6,895 | 0,1741 |
| GC | 3,554 | 3,890 | 2,972 | 3,472 | 0,3792 |
| GD | 4,128 | 3,238 | 3,568 | 3,664 | 0,3673 |

É possível verificar que o algoritmo que retorna a curva média é o que registra os menores tempos, tal resultado era esperado, devido sua simplicidade. Porém como é possível observar, os algoritmos que fazem uso dos dados clusterizados (GC e GD) apresentaram menores tempos que aquele que utiliza o conjunto total de curvas de carga (GB) para realizar previsão.

Tal característica apresentada por GC e GD se deve em maior parte a execução do algoritmo ter de buscar suas respostas em bases de dados 87,5 a 91,66% menores que a original (2.016 registros na base reduzida - 21 curvas de carga, contra 14.784 na base original - 154 curvas de carga), fazendo com que as respostas sejam encontradas mais rapidamente, uma vez que o número de testes e de comparações é diminuído.

É importante salientar que nos tempos de GC e GD não estão contidos os tempos de pré-processamento e clusterização das curvas de carga, uma vez que essas tarefas são realizadas de maneira prévia (*offline*).

## 5.7 Resultados de Previsão

Os resultados de previsão estão divididos em quatro grupos previamente apresentados (GA, GB, GC e GD), correspondentes à previsão baseada na curva média, no MDS com o conjunto total de curvas de carga, com o MDS que faz uso do conjunto clusterizado a partir de múltiplas características da curva de carga e pelo MDS baseado nos dados clusterizados com base na temperatura, respectivamente.

A Tabela 5 apresenta os valores do MAPE para os quatro grupos de previsão, e o cluster escolhido como previsão nos grupos GC e GD são apresentados. Os melhores resultados são destacados em negrito, e os segundos melhores resultados para cada dia da semana foram apresentados com um asterisco (*).

Tabela 5 - Resultados do Cálculo do MAPE para os Quatro Grupos.

| **Cálculo do MAPE (%)** | |
|---|---|
| | Grupos |

| Dia da Semana | GA | GB | GC | | GD | |
|---|---|---|---|---|---|---|
| | MAPE | MAPE | Cluster | MAPE | Cluster | MAPE |
| Segunda | 18,48 | **12,64** | 1 | 16,75 | 3 | 16,66* |
| Terça | 22,74 | **9,28** | 3 | 16,14 | 1 | 16* |
| Quarta | 17,47 | 17,8 | 1 | **15,29** | 1 | 17,39* |
| Quinta | 16,27 | 17,1 | 3 | **14,02** | 2 | 15,35* |
| Sexta | 18,77 | 15,05* | 1 | **14,84** | 1 | 17,95 |
| Sábado | 32,15 | 29,99* | 3 | **26,11** | 2 | 31,56 |
| Domingo | 15,81 | 17,14 | 1 | 13,49* | 2 | **12,26** |

Os resultados do grupo GA foram os menos expressivos, não obtendo nenhum resultado classificado como o melhor ou o segundo melhor para os dias da semana, até certo ponto isto era esperado, uma vez que seu nível de generalização é bastante alto.

Os resultados apresentados por GB para a previsão da segunda e da terça feira foram os melhores, com MAPEs 4.11 e 6.86 unidade enores que os apresentados pelos segundos melhores resultados, respectivamente, obtendo ainda, o segundo melhor resultado para o sábado.

O GC apresentou os melhores resultados, acumulando quatro dos menores MAPEs verificados, e um segundo menor.

O grupo GD apresentou um dos melhores resultados e quatro dos segundos melhores, quando comparado apenas com GB (segundo colocado de maneira geral) ele é capaz de superar três de seus resultados apenas, o mesmo ocorre quando comparado com GC (o melhor colocado geral).

É importante destacar que em nenhum dos casos o cluster selecionado como previsão foi superado pelos demais (para um mesmo grupo), isso quer dizer que em nenhum dos casos os valores de MAPE verificados para os clusteres não selecionados superou o do selecionado. Desta maneira é possível constatar que a característica de temperatura se mostrou adequada para a seleção do cluster a ser utilizado como previsão, porém não se revelou viável para caracterizar as curvas de carga durante as tarefas de clusterização, conforme visto anteriormente.

No caso da segunda feira, a curva de carga prevista é inferior em alguns pontos e em outros ultrapassa a verificada, porém em maior parte fica contida no intervalo de confiança gerado a partir das curvas que compuseram o referido cluster, conforme Figura 22.

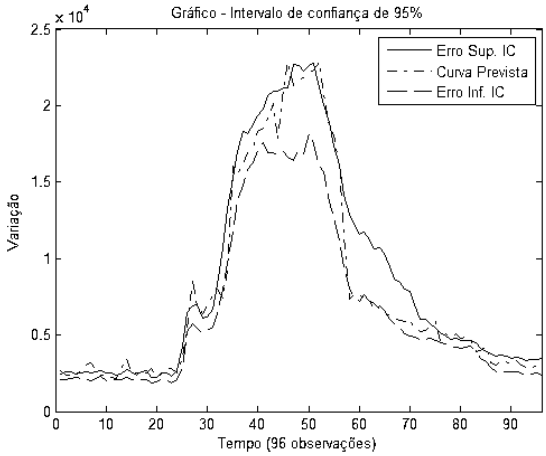



Figura 22 – Intervalo de Confiança: Segunda Feira.

O experimento de previsão foi repetido em outras quatro oportunidades, atualizando a base de dados e excluindo a última semana registrada, e os resultados obtidos se mostraram bastante semelhantes aos apresentados na ɔela 5 quanto ao número de melhores respostas obtidas por cada grupo, sendo que, o GB foi o que mais oscilou neste quesito, a Tabela 6 apresenta a quantidade de menores e segundo menores MAPEs obtidos por cada grupo, onde o grupo com as melhores respostas para cada experimento está destacado com negrito, e o com a maior quantidade das segundas melhores com um asterisco (*).

Tabela 6 - Resultados Obtidos Pelos Grupos em Quatro Ocasiões de Previsão.

| **Melhores Resultados (menores valores) do Cálculo do MAPE** | | | | | |
|---|---|---|---|---|---|
| Respostas | Experimentos | Grupos | | | |
| | | GA | GB | GC | GD |
| Melhores | 1 | 1 | 1 | **4** | 1 |
| Segundas M. | | 1 | 1 | 2 | 3* |
| Melhores | 2 | 0 | **3** | **3** | 1 |
| Segundas M. | | 0 | 2 | 2 | 3* |
| Melhores | 3 | 0 | 2 | **3** | 2 |
| Segundas M. | | 1 | 3* | 3* | 0 |
| Melhores | 4 | 0 | 1 | **5** | 1 |
| Segundas M. | | 1 | 3* | 1 | 2 |

Em todos os ensaios de previsão, o grupo GC apresentou o maior número de melhores respostas, no primeiro experimento apresentou quatro delas de um total de sete possíveis, no segundo e terceiro experimento apresentou três, com destaque para o quarto, onde das sete melhores respostas, o grupo acumulou cinco. O grupo GB, por sua vez, ficou em segundo lugar, com o maior número de melhores respostas em um caso e as segundas melhores em dois.

É importante observar que no segundo experimento, GB e GC empataram no número de melhores respostas, se esse caso não fosse adicionado a conta de melhores respostas dos dois grupos, GB não acumularia nenhuma das melhores respostas, GC perderia uma, porém os dois continuariam em suas posições, como sendo a segunda melhor e a melhor alternativa para realizar as previsões, respectivamente.

## 5.8 Considerações Finais

Neste capítulo foram apresentadas diversas etapas que devem anteceder uma tarefa de previsão, sobretudo quando baseada em dados clusterizados, dando importância para a validação dos dados e dos clusters.

A qualidade dos dados utilizados nos estudos exerce influência direta na capacidade de assertividade das previsões. Questões como a precisão dos medidores, a análise e o pré-processamento dos dados devem ser tratadas com atenção.

Os parâmetros de entrada também exercem influência direta nas previsões obtidas, desta maneira, a temperatura média prevista utilizada é capaz de alterar os resultados de previsão, sendo necessário tratar com atenção esta questão, utilizando dados de fontes confiáveis.

Foi possível observar superioridade do grupo (GC) que fez uso dado dados clusterizados a partir de diversas características (temperatura, consumo máximo, mínimo, média e desvio padrão) para descreverem as curvas de carga.

O local em que se insere a microrrede que forneceu os dados para este trabalho apresenta alta variação de temperaturas, registrando em uma mesma semana de inverno amplitudes térmicas próximas de 20°C. Acredita-se que as altas variações de temperatura, sobretudo no inverno, tenham influenciado diretamente na capacidade de assertividade das previsões, causando MAPE’s altos.

É importante destacar ainda que o local em que estava alojado o medidor responsável por registrar a variação da temperatura ao longo de um dia, era de metal, e que ficava exposto ao tempo, fazendo com que o calor e o frio absorvidos pelo metal pudessem contaminar as medições realizadas.

# Capítulo 6

# Conclusões

Neste capítulo são apresentadas as principais conclusões e considerações realizadas, sendo essas divididas entre os subcapítulos: Principais Contribuições, Limitações e Trabalhos Futuros.

## 6.1 Principais Contribuições

Este trabalho apresentou a utilização de técnicas de agrupamento com o objetivo de reduzir um determinado conjunto de dados utilizado em tarefas de previsão de carga, indicando para isso, o uso de algoritmos de clusterização de dados.

Um diferencial deste trabalho é a utilização de um grande conjunto de curvas de carga diárias (correspondentes a dois anos) referentes a um nível menos agregado, onde se encontram maiores incertezas e variações, variações essas, entre duas ou mais curvas de cargas para um mesmo tipo de dia.

A proposta de utilizar tarefas de clusterização de curvas de carga com o objetivo de reduzir o conjunto original, impactando com a redução de 40 a 50% no tempo computacional despendido para realizar determinada tarefa de previsão, também é destacada.

A forma de caracterizar as curvas de carga, para que essas fossem submetidas a tarefas de clusterização, permitiram agrupamentos significativos, avaliados por meio da técnica de análise de Silhueta de Cluster.

A utilização da técnica Bootstrap permite o cálculo do intervalo de confiança dos clusters gerados e também deve ser destacada, uma vez que viabiliza agregar novas característica aos resultados, possibilitando diferentes interpretações para as previsões e para a verificação da similaridade das curvas que compõem cada cluster.

Por fim, é importante destacar a utilização da última curva semelhante armazenada na base de dados com o objetivo de corrigir as previsões realizadas, tal técnica permitiu em alguns casos considerável melhora nas curvas de previsão, tanto para o algoritmo do MDS que fazia uso do conjunto total de curvas de carga, quanto para o algoritmo que fazia uso do conjunto de curvas clusterizadas.

## 6.2 Limitações

A quantidade limitada de informações externas referentes aos dias (curvas de carga diária) utilizados nos ensaios pesa negativamente, podendo exercer considerável diminuição na qualidade dos agrupamentos e das previsões realizadas.

A utilização de dados clusterizados com quaisquer outras técnicas de previsão de carga pode ser questionada quanto ao seu impacto na qualidade das previsões resultantes, é certo que a redução do número de curvas de cargas, por meio da combinação das mais similares, resulta na simplificação e perda de algumas características. Mas é importante perceber que tal curva representativa é formada por outras bastante semelhantes a ela, e desta forma, apesar de perder algumas características, pode ganhar outras, úteis ou não a previsão.

Considerando ainda o impacto do uso de dados clusterizados é importante perceber que, a quantidade de clusters, e os níveis de similaridade utilizados para sua composição, assim como os tipos de medidas utilizadas entre outras características intrínsecas a tarefa, podem ser alteradas e aperfeiçoadas dependendo de cada situação.

Não foram realizados testes com outros conjuntos de dados de consumo, em níveis mais agregados, experiência que permitiria uma perspectiva maior para a avaliação do emprego de clusterização de dados neste tipo de tarefa. Por se tratar de uma limitação factível de ser ultrapassada, ela é descrita no tópico de trabalhos futuros.

## 6.3 Trabalhos Futuros

A utilização de outras técnicas de clusterização, como a abordagem hierárquica, poderia resultar em agrupamentos ainda mais significativos, que poderiam ser utilizados com algum algoritmo

de previsão e posteriormente comparados com a abordagem utilizada neste trabalho, ficando essa verificação como trabalho futuro.

Com a disponibilidade de mais informações referentes as curvas de carga utilizadas (como dados meteorológicos mais precisos, incluindo índice pluviométrico, umidade relativa do ar, e intensidade de iluminação solar) é possível modificar os algoritmos de previsão e clusterização, uma vez que mais informações explicativas estejam disponíveis, mais significativas podem ser as seleções das curvas sua clusterização. Porém, apesar destes dados poderem ser obtidos para determinados ambientes, sobretudo de estudos, é necessário indagar se tal tarefa seria praticável em cenários reais, como o de uma cidade, onde a intensidade de iluminação, e o índice de precipitação podem variar dentro de um pequeno espaço demográfico, graças a presença ou não de nuvens, por exemplo. Portando, a necessidade de utilizar muitas informações externas às curvas de carga, também podem influenciar para a não aplicabilidade do método em cenários reais.

Utilizar a técnica de clusterização de dados descrita com outros métodos de previsão, além de utilizar diferentes conjuntos de dados, com diferentes níveis de agregação, que permitam verificar resultados semelhantes ou não aos descritos no presente trabalho, compõe a lista de trabalhos futuros.

Existem trabalhos que indicam que o formato e o ângulo do inicio da rampa de uma curva de carga, por vezes está relacionado com padrões verificáveis, desta maneira, a combinação desta abordagem com a dos dias similares e os dados clusterizados pode possibilitar melhores resultados de previsão.

# Referências Bibliográficas


Agarwal, Y.; Weng, T.; Gupta, R.K.(2011). Understanding the role of buildings in a smart microgrid. Design, *Automation & Test* in Europe Conference & Exhibition, Pages 1-6. Grenoble.

Agrawal, R.; Lin, K.; Sawhney, H. S.; Shim, K. (1995). Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. Proceedings of the *21th International Conference on Very Large Data Bases*, pages 490–501. San Jose - CA.

Aikes Junior, J. (2012). Estudo da Influência de Diversas Medidas de Similaridade na Previsão de Séries Temporais Utilizando o Algoritmo KNN-TSP. Dissertação de Mestrado. Universidade Estadual do Oeste do Paraná - UNIOESTE, Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos, Foz do Iguaçu - PR.

Aikes Junior, J.; Lee, H. D.; Ferrero, C. A.; Wu, F. C. (2012). Estudo da Influência de diversas Medidas de Similaridade na Previsão de Séries Temporais utilizando o Algoritmo kNN-TSP. Proc. *Encontro Nacional de Inteligência Artificial* in Series Brazilian Conference on Intelligent System, pp. 1 - 12.

Alencar, A. B. (2007). Mineração e visualização de coleções de séries temporais. Dissertação - Instituto de Ciências Matemáticas e de Computação - ICMC da Universidade de São Paulo - USP. São Carlos - SP.

Amo, S. (2012). Avaliação de Clusteres. Data Mining. Arquivos de Aula. Faculdade de Computação, Universidade Federal de Uberlândia, Uberlândia - MG.

ANEEL - Agência Nacional de Energia Elétrica. (2012). Espaço do Consumidor. Disponível em: < http://www.aneel.gov.br/area.cfm?id_area=19 > Acesso em 05 de Novembro de 2012.

Barghinia, S.; Kamankesh, S.; Mahdavi, N.; Vahabie, A. H.; Gorji, A. A. (2008). A Combination Method for Short Term Load Forecasting Used in Iran Electricity Market by NeuroFuzzy, Bayesian and Finding Similar Days Methods. *Electricity Market, 2008. EEM 2008 5th International Conference on European*, pp. 1 - 6. Lisboa - Portugal.

Beltrame, W. A. R; Fonseca, F. C. S. (2010). Aplicações Práticas dos Algoritmos de Clusterização Kmeans e Bisecting K-means. Departamento de Informática, *I Seminário de Informática*, pp. 10 – 8, Universidade Federal do Espírito Santo - UFES.

Box, G. E. P.; Jenkins, G. M. (1976). *Time series analysis forecasting and control*. Edição revisada. San Francisco: Holden Day.

Brockwell, P. J. e Davis, R. A. (2002). *Introduction to Time Series Forecasting*. 2° ed., Springer, New York.

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction*. Chapman and Hall/CRC, Sixth Edition.

Chen, L. e Ozsu, M. T. (2003). Similarity-based Retrieval of Time-Series Data Using Multi-Scale Histograms. Technical Report CS, School of Computer Science - University of Waterloo, Waterloo, Canada.

Cluto. (2010). Clustering High-Dimensional Datasets. Software Informations, disponível em <http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview>, Acesso em 03 de Setembro de 2012.

CNAE - Classificação Nacional de Atividades Econômicas; CONCLA - Comissão Nacional de Classificação. (2012). Tabelas de Códigos e Denominações. Disponível em: <http://www.cnae.ibge.gov.br/estrutura.asp?TabelaBusca=CNAE_200@CNAE%202.1>. Acesso em 06 de Novembro de 2012.

DiCiccio, T. J.; Efron, B. (1996). Bootstrap confidence intervals. *Statist. Sci*., Volume 11, pp. 189 – 228, Number 3.

Drago, I.; Varejão, F. M. (2007). Uma análise experimental de métricas de similaridade na classificação de séries temporais. *VI ENIA - Encontro Nacional de Inteligência Artificial*. Rio de Janeiro - RJ.

EPE - Empresa de Pesquisa Energética. (2012). Anuário Estatístico de Energia Elétrica 2012, Disponível em: < http://www.epe.gov.br/AnuarioEstatisticodeEnergiaEletrica/ >. Acesso em 05 de Novembro de 2012.

Esteves, G. R. T. (2003). Modelos de Previsão de Carga de Curto Prazo. Dissertação de Mestrado apresentada na Pontifícia Universidade Católica do Rio de Janeiro. Disponível em: < http://www2.dbd.puc-rio.br/pergamum/biblioteca/php/mostrateses.php?open=1&arqtese=5000064391_03_Indice.html >. Acesso em 13 de Novembro de 2012.

Falcão, D. M. (2009). Smart Grid e Microredes: O Futuro Já é Presente. *VIII Simpósio de Automação e Sistemas Elétricos*. Rio de Janeiro - RJ.

Felipe, I. J. S. (2012). Aplicação de modelos ARIMA em séries de preços de soja no norte do Paraná. *Tekhne e Logos*, v.3, n.3. Botucatu - SP.

Fontana, A., Naldi, M. C. (2009). Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados. Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, Relatório Técnico do Instituto de Ciências Matemáticas e de Computação - ICMC. ISSN - 0103-2569. São Carlos – SP.

Francisquini, A. A. (2006). Estimação de Curvas de Carga em Pontos de Consumo e em Transformadores de Distribuição, Dissertação de Mestrado, Universidade Estadual Paulista "Júlio de Mesquita Filho" - UNESP, Faculdade de Engenharia de Ilha Solteira - FEIS, Ilha Solteira – SP.

Gonçalves, E. C. (2011). Data Mining com a Ferramenta Weka. III Fórum de Software Livre de Duque de Caxias - FSLDC, Duque de Caxias – RJ.

Guirelli, C. R. (2006). Previsão da Carga de Curto Prazo de Áreas Elétricas Através de Técnicas de Inteligência Artificial. Tese apresentada a Escola Politécnica da Universidade de São Paulo - USP. São Paulo - SP, 2006, 127p. Disponível em: < www.teses.usp.br/teses/.../TESECLEBERROBERTOGUIRELLI.pdf >. Acesso em 08 de Dezembro de 2012.

Jain, A. K.; Murty, M. N.; Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv*., 31(3):264–323.

Kadowaki, M; Ohishi, T; Soares Filho, S; Lima, W. S. (2004). Modelo de Previsão de Demanda de Carga de Curtíssimo Prazo para o Período da Ponta. *XXXVI Simpósio Brasileiro de Pesquisa Operacional - SBPO* - O Impacto da Pesquisa Operacional nas Novas Tendências Multidisciplinares, pp. 2160 - 2171. São João del-Rei – MG. 2004.

Kagan, N; Oliveira C. C. B; Robba, E. J. (2010). *Introdução aos Sistemas de Distribuição de Energia Elétrica*. São Paulo – SP: Blucher, Ed. 02.

Kaufman, L. e Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Keogh, E.; Wei, L.; Xi, X; Vlachos, M.; Lee, S.; Protopapas, P. (2006). LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. *VLDB '06 Proceedings of the 32nd international conference on Very large data bases*. Páginas 882-893. Seoul, Korea.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, Hoboken.

Leone, K. M. A. (2006). Previsão de carga de curto prazo usando ensembles de previsores selecionados e evoluídos por algoritmos genéticos. Tese, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas. Disponível em: < http://www.bibliotecadigital.unicamp.br/document/?view=vtls000410708 >. Acesso em 12 de Novembro de 2012.

Liu, L. (2009). *Time Series Analysis and Forecasting*. 2° Ed. v. 2.2., Scientific Computing Associates® Corp. Illinois, Chicago – USA.

Liu, Y.; Li, Z.; Xiong, H.; Gao, H.; Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *2010 IEEE International Conference on Data Mining*. ISBN: 978-0-7695-4256-0, Sydney, Australia.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, ISBN 3-540-26239-3, New York.

Mandal, P.; Srivastava, A. K.; Negnevitsky, M.; Jung-Wook Park. (2008). An Effort to Optimize Similar Days Parameters for ANN Based Electricity Price Forecasting. *Industry Applications Society Annual Meeting IAS '08*, pp 1 – 9, Edmonton - Alta.

Méffe, A. (2001). Metodologia para Cálculo de perdas Técnicas por Segmento do Sistema de Distribuição. Dissertação de Mestrado. Escola Politécnica da Universidade de São Paulo. São Paulo - SP.

Miller, J.M. (2009). Energy storage system technology challenges facing strong hybrid, plug-in and battery electric vehicles. *Vehicle Power and Propulsion Conference*, 2009. IEEE, pages 4-10. Dearborn, MI

Moghram, I; Rahman, S. (1989). Analysis and evaluation of five short-term load forecasting techniques. *Power Systems, IEEE Transactions on*. Volume 4, 2° Ed., 1989. Páginas 1484-1491.

Mörchen, F. (2006). Time series knowledger mining. Master's thesis, Philipps-Universität Marburg, Marburg, Germany.

Moretin, P. A.; Toloi, C. M C. (1981). *Modelos para Previsão de Séries Temporais*. Instituto de Matemática Pura e Aplicada, Volume II. Rio de Janeiro - RJ.

Morettin, P. A. e Toloi, C. M. C. (2006). *Análise de Séries Temporais*. 2° ed., Edgard Blücher LTDA, São Paulo - SP.

Mota, L. T. M; Mota A. A; França, A. L. M. (2004). Modelagem e simulação de cargas residenciais termostáticas para a recomposição do sistema elétrico a partir de uma abordagem orientada de objetos. *SBA Controle & Automação* vol.15, no.2, pp. 202 – 214, Campinas - SP.

Mu, Q.; Wu, Y.; Pan, X.; Huang, L.; Li, X. (2010). Short-term Load Forecasting Using Improved Similar Days Method. *Power and Energy Engineering Conference APPEEC*, pp. 1 - 4, Asia-Pacific.

Müller, M. (2007). *Information Retrieval for Music and Motion*. ISBN 978-3-540-74047-6 Springer Berlin Heidelberg, New York.

Neves, R. C. D. e Alvares, L. O. C. (2003). Pré-processamento no processo de descoberta de conhecimento em banco de dados. Dissertação. Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-Graduação em Computação. Rio Grande do Sul - RS.

Pelleg, D. e Moore, A. (2000). X-means: extending k-means with efficient estimation of the number of clusters. In Proceedings of the *Seventeenth International Conference on Machine Learning*, pages 727-734, São Francisco - CA.

Pellegrini, F. R. e Fogliatto, F. S. (2001). Passos para Implantação de Sistemas de Previsão de Demanda - Técnicas e Estudo de Caso. *Revista PRODUÇÃO*, v. 11, n. 1, pp. 43 - 64, São Paulo - SP. Disponível em < http://www.scielo.br/pdf/prod/v11n1/v11n1a04.pdf >. Acessado em 10 de Julho de 2013.

Pimenteli, E. P.; França, V. F.; Omar, N. (2003). A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: Anais do *Simpósio Brasileiro de Informática na Educação*, Rio de Janeiro, RJ. SBC.

Procel; Eletrobrás. (2007). Pesquisa de Possse de Equipamentos e Hábitos de Uso - Ano Base 2005. Relatório Brasil - Classe Residencial.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, Inc. San Francisco - CA.

Pyle, D. (2006). *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, Inc. San Francisco, California - USA.

Rahman, S.; Hazim, O. (1993). A generalized knowledge-based short-term load-forecasting technique. *IEEE Transactions on Power Systems*, Volume 8, 2° Ed., pp. 508 - 514, 1993.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Volume 20, Pages 53–65.

Scottá, F. C.; Fonseca, E. L. (2013). Análise de tendências em séries temporais de dados meteorológicos e dados de sensoriamento remoto orbital em áreas de vegetação campestre natural do bioma Pampa, localizadas na Depressão Central do RS. Anais *XVI Simpósio Brasileiro de Sensoriamento Remoto - SBSR*, INPE, pp. 3003 - 3009, Foz do Iguaçu - PR.

Senjyu, T.; Mandal, P.; Uezato, K.; Funabashi, T. (2005). Next day load curve forecasting using hybrid correction method. *Power Systems, IEEE Transactions on*, Volume 20, Ed. 1, pgs 102-109.

Senjyu, T; Higa, S; Uezato, K. (1998). Future load curve shaping based on similarity using fuzzy logic approach. *Generation, Transmission and Distribution, IEE Proceedings*, 1998, 375-380.

Souza, R. C. (1989). Modelos Estruturais para Previsão de Séries Temporais : Abordagens Clássica e Bayesiana. *17° Colóquio Brasileiro de Matemática*, Rio de Janeiro - RJ.

Theodoridis, S.; Koutroumbas, K. (2006). *Pattern Recognition*. Academic Press, Inc. ISBN:0123695317, Third Edition. Orlando - FL.

Tseng, V. S.; Lee, C. (2009). Effective Temporal Data classification by Integrating Sequential Pattern Mining and Probabilistic Induction. Department of Computer Science and Information Engineering, National Chen-Kung University. *Expert Systems with Applications*. Volume 36, Issue 5, July 2009, Pages 9524–9532.

Weiss, G. M. (2004). Mining with Rarity: a Unifying Framework, *ACM SIGKDD Explorations Newsletter* 6(1): 7–19, Nova York - NY.

# Apêndice A

# Artigo Publicado

Este apêndice contém o artigo apresentado e publicado no Simpósio Brasileiro de Sistemas Elétricos - SBSE, 2014, Foz do Iguaçu – PR. As formatações e numeração de páginas são mantidas de acordo com a publicação no evento.

# Clusterização de Curvas de Carga para o Método de Dias Similares

M. R. Müller e E. M. C. Franco

***Resumo*—O método de dias similares permite realizar previsão de carga de curtíssimo prazo a partir de dados históricos de consumo de energia elétrica, além de dados correlatos, que permitem traçar analogias com um dia futuro. Este trabalho apresenta a utilização de clusterização de curvas de carga do nível mais desagregado para o método de dias similares, permitindo a obtenção de conjuntos reduzidos de dados com desempenho similar ou superior aos utilizados originalmente. Implementações convencionais do mesmo método são utilizadas para comparação de resultados. O cenário que fornece os dados para os estudos, assim como os equipamentos empregados e a etapa de pré-processamento de dados são apresentadas. A análise de silhuetas de cluster foi empregada com o objetivo de validar os agrupamentos. Por meio do cálculo do MAPE foi possível verificar a assertividade das previsões, indicando superioridade daquela baseada nas curvas de carga clusterizadas.**

***Palavras-chave*—Método de Dias Similares; Previsão de Carga; Clusterização de Dados; Medidores Eletrônicos.**

## I. Introdução

A perspectiva de aumentar lucros e evitar prejuízos estimula empresas de comercialização de energia elétrica - EE a investirem em ferramentas de previsão de demanda.

Ao passo que dados estejam disponíveis, diversos estudos podem ser conduzidos com o intuito de criar métodos que permitam produzir estimativas para um determinado horizonte futuro.

Dispor de adequadas estimativas para a demanda de EE pode significar mais lucros para as empresas, uma vez que o ganho está atrelado ao cumprimento dos limites contratados.

Evitar compras emergenciais de energia de outras distribuidoras, ou o pagamento de pesadas multas devido ao descumprimento de obrigações inerentes ao serviço de fornecimento de EE, também constitui o objetivo das empresas de energia.

Comumente dividem-se as previsões de acordo com a faixa de tempo futuro que abrangem, inicialmente têm-se as de curtíssimo/curto prazo, que abrangem horizontes de algumas horas a poucos dias, as de médio prazo que envolve períodos de meses a um ano, e por fim as de longo prazo, projetando a demanda para até o limiar de uma década [01].

As previsões de curtíssimo prazo, objeto do presente trabalho, são tidas como fundamentais para orientar o planejamento da operação do sistema, transferências de energia e o gerenciamento da demanda [02].

Obter dados de consumo de EE é fundamental para suprir metodologias de previsão de carga. A aquisição desses dados atualmente é facilitada graças ao uso de modernos medidores digitais de consumo de energia elétrica.

Os medidores digitais permitem capturar dados de consumo em intervalos pré-definidos, possibilitando ainda que esses sejam transmitidos por uma rede de dados até um servidor de armazenamento.

No passado não se verificava a atual disponibilidade de dados de consumo de energia, fazendo com que estudos fossem conduzidos com menores amostras de dados.

Com o surgimento das redes inteligentes e a intensificação dos estudos nessa área, espera-se um aumento significativo da disponibilidade de dados, permitindo novas abordagens para a previsão de carga.

Apesar da menor disponibilidade de dados de consumo de EE no passado, os métodos de previsão de carga não constituem algo novo, sendo aplicados desde a década de 80. Tais métodos podem ser divididos em dois grupos, o primeiro formado pelos métodos estatísticos, e o segundo, constituído por métodos que têm suas bases em princípios da inteligência artificial. Existem ainda os métodos híbridos, baseados na combinação de dois ou mais métodos de previsão [03].

Entre os métodos clássicos, pode-se destacar: Regressão Linear Múltipla, Modelo Auto-regressivo Integrado de Média Móvel - ARIMA, Alisamento Exponencial e Análise Espectral. Entre os métodos Baseados em Inteligência Artificial, podem ser citados os Sistemas Especialistas, as Redes Neurais, os que empregam Lógica Fuzzy e os Algoritmos Genéticos [03].

O método de dias similares - MDS se baseia em dados que alcançam horizontes de um ou mais anos, de onde se procura determinar similaridades entre os dias catalogados com as características conhecidas para um dia futuro (dia da semana, estação do ano e temperatura média prevista). Sendo assim, a previsão se dá graças a capacidade do algoritmo em encontrar analogias entre dias passados e futuros [04, 05].

A abordagem de dias similares se destaca por não lidar apenas com a parte não linear da curva de carga, mas também com dias especiais, como os fins de semana e feriados. Esta abordagem também se mostra útil em situações em que os modelos de previsão de carga precisos são difíceis de projetar [05].

De maneira simples o método de dias similares pode ser

Agradecimentos a Fundação Parque Tecnológico Itaipu - FPTI, pela bolsa[1].

M. R. Müller[1], UNIOESTE/PGESDE (marcos_ricardo@live.com).
E. M. C. Franco, UNIOESTE/PGESDE (emfra.unioeste@gmail.com).

representado com uma tabela atributo-valor, onde são catalogados os dados referentes as curvas diárias de carga e dados externos, como os meteorológicos.

O MDS tem como uma de suas fragilidades o tempo computacional, que se eleva em sinergia com a quantidade de dados históricos considerados para a previsão. Com base nisso o trabalho propõe uma abordagem que permite realizar uma redução dos dados históricos, acarretando em tempos computacionais menores.

O presente trabalho utiliza tarefas de clusterização de dados para reduzir o número de curvas de cargas necessárias para realizar previsões com o método de dias similares (reduzindo em aproximadamente 95% o conjunto inicial), obtendo índices de assertividade semelhantes ou superiores aos obtidos com o conjunto convencional de curvas de cargas.

Os resultados encontrados a partir das curvas de carga clusterizadas são comparados com outros, obtidos pelo método de dias similares baseado no conjunto total de curvas de carga e suas médias.

Os diferentes usos baseados no conjunto inicial de curvas de carga permitiu suprir resultados que foram mutuamente confrontados. O cálculo do MAPE possibilitou avaliar a assertividade das previsões, apontando superioridade no uso das curvas de carga clusterizadas.

## II. MÉTODOS DE PREVISÃO

Uma abordagem do MDS baseia-se na divisão de todas as curvas de carga entre os dias da semana correspondentes, agrupados de acordo com as quatro estações do ano. Obter a médias desses dias é uma maneira mais ou menos adequada de se caracterizar os dias da semana nas estações do ano.

A previsão se dá ao selecionar a curva média corresponde ao dia da semana na estação do ano em que se quer prever. Os passos realizados pelo MDS - 1 são apresentados em *A*.

O algoritmo MDS – 2 faz uso do conjunto completo de curvas de carga, e é apresentado em *B*, já o MDS – 3 faz uso das curvas de carga clusterizadas, e é descrito em *C*.

### *A. MDS - 1*

Com base em [5] o MDS - 1 difere ao representar os dias da semana (divididos nas quatro estações do ano) a partir de suas médias. Devido a redução do número de curvas ocorrer de maneira prévia (*off-line*), exige pouco em tempo de execução.

O algoritmo inicia com a entrada de dados referentes ao dia que se quer prever (estação do ano e dia da semana), passa para a busca da curva média correspondente, e termina com a apresentação da curva de previsão.

### *B. MDS - 2*

O MDS - 2 é o que computacionalmente exige mais, devido ao processo de busca na base de dados, a comparação entre as curvas selecionadas e a extração da média ocorrer em tempo de execução (*on-line*).

O algoritmo inicia com a entrada de dados correspondentes ao dia que se quer prever (estação do ano, dia da semana e temperatura média prevista[1]), além da informação referente ao crescimento médio anual do consumo de EE. Com base nos parâmetros de entrada e a similaridade entre as curvas selecionadas, o MDS retorna uma curva média como previsão [04, 05].

### *C. MDS - 3*

O MDS – 3 de maneira similar ao MDS - 1 realiza tarefas de pré-processamento das curvas de carga, fazendo com que a execução do algoritmo corresponda a uma tarefa de busca em uma base reduzida de dados, atendendo determinados parâmetros de entrada.

O MDS - 3 apresentou o melhor tempo médio de execução, bastante próximo do verificado no MDS - 1, e aproximadamente 14 vezes superior ao do MDS - 2.

Como entradas o algoritmo espera receber a estação do ano, o dia da semana e a temperatura média prevista para o dia de previsão.

O índice de crescimento anual do consumo de EE é utilizado na tarefa de pré-processamento das curvas de carga, para que seja aplicado aos dados do ano anterior ao último empregado.

Na tarefa de pré-processamento, as curvas de carga de uma determinada estação do ano são divididas por dia da semana para serem submetidas a tarefas de clusterização, utilizando o algoritmo *Simple K-means* com a distância euclidiana.

O algoritmo de clusterização *K-means* também é conhecido como K-médias. Esse algoritmo utiliza o conceito de centróides como protótipos representativos dos grupos, sendo calculados pela média de todos os objetos do grupo que representa.

O objetivo do algoritmo K-means "[...] é encontrar a melhor divisão de $P$ dados em $K$ grupos $Ci, i = 1, \dots, K$, de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada" [06].

O cálculo da distância euclidiana consiste em verificar a distância entre pares de pontos de duas sequencias [07]. Dadas duas séries temporais $P = (p_1, \dots, p_n)$ e $Q = (q_1, \dots, q_n)$ de mesmo tamanho $n$, a distância Euclidiana entre essas duas séries é definida de acordo com a Equação 1:

$$D_{istânciaEuclidiana}(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (1)$$

Para as tarefas de clusterização foram utilizados dois cenários, no primeiro (MDS – 3A) utilizou-se apenas a temperatura media para realizar a clusterização, no segundo

[1] Todas as informações de temperatura prevista foram retiradas do sítio da web do Instituto Tecnológico SIMEPAR <www.simepar.br>, agência que tem como uma de suas finalidades prover a sociedade com informações de natureza meteorológica.

cenário (MDS – 3B), além da temperatura considerou-se o mínimo, o máximo e a média das curvas de carga. Foram realizados testes utilizando a mediana no lugar da média, isso contribuiu para que os clusters formados fossem menos significativos, de acordo com a análise da silhueta de cluster.

Inicialmente são gerados três clusters para cada dia da semana. A **Fig. 1** apresenta as curvas de carga clusterizadas a partir das quarta feiras do inverno de 2012 e 2013, tendo como base a temperatura média o mínimo, o máximo e a média das curvas de carga. Os clusters um a três apresentam temperaturas médias de 37, 32 e 30 graus célsius, respectivamente.

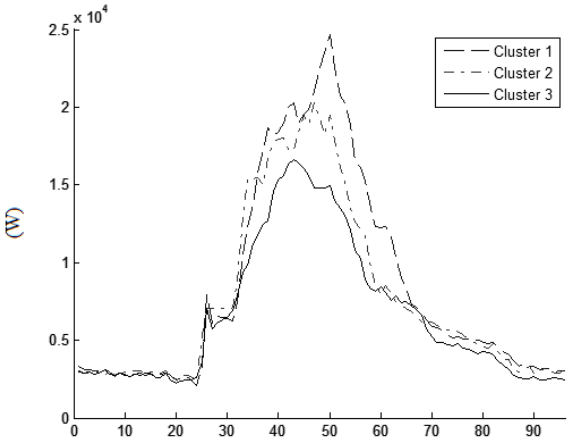



**Fig. 1 - Curvas Clusterizadas**

Após criados os clusters que representam as curvas de carga, executam-se tarefas de comparação entre eles, com o objetivo de encontrar clusters semelhantes, que possam ser representados por apenas uma curva de carga, reduzindo o conjunto de curvas necessárias.

Para que as curvas de carga possam ser submetidas a tarefas de comparação, calcula-se o mínimo, a média e máximo. De maneira a permitir que apenas a forma (*shape*) das curvas sejam comparadas, seus dados são normalizados.

Todas as curvas normalizadas são comparadas entre si, com o objetivo de se verificar a distância absoluta entre elas. As que possuem menores distâncias absolutas (dentro de uma faixa estabelecida) são verificadas quanto a seus mínimos, médias e máximos, sendo que se forem semelhantes dentro de uma faixa de **Y** (5%) por cento (considerando todos os pontos pares entre duas séries temporais – curvas de carga) são tidas como bastante semelhantes, e a média simples entre elas passa a representa-las no algoritmo. O valor de cinco para **Y** foi estabelecido com base em testes para o conjunto de dados considerado, devendo ser adequado para outras situações.

O algoritmo inicia com a entrada de dados correspondentes ao dia que se quer prever, informando a estação do ano, o dia da semana e temperatura média prevista.

Com base nas entradas do usuário, o algoritmo busca em uma base de curvas pré-processadas a que tem temperatura média mais próxima a informada pelo usuário (em caso de empate, escolhe randomicamente entre elas) e a retorna como previsão.

## III. Cenário de Estudos

As curvas de carga utilizadas para o presente trabalho são oriundas do nível mais desagregado, correspondentes a um consumidor comercial do ramo alimentício.

A obtenção dos dados de consumo de energia elétrica iniciou em abril de 2012, entretanto problemas técnicos foram responsáveis pela introduziram grande quantidade de *outliers* e dados faltantes. A partir da segunda metade do mesmo ano a medição se tornou contínua e estável.

Os dados obtidos por meio dos medidores eletrônicos se caracterizam como séries temporais. São amostrados de uma base de dados de aferições realizadas em intervalos iguais de tempo e de maneira continua.

As aferições são realizadas com intervalos de quinze minutos, de tal modo que noventa e seis aferições compõem uma curva de carga diária.

As curvas diárias de consumo foram dividas de acordo com as estações do ano e dias da semana, sendo apresentadas as que compreendem o inverno do ano de 2012 e 2013.

O trabalho considera os registros pertinentes ao inverno, pois estes são os que apresentam menor quantidade de dados faltantes e *outliers* (aproximadamente 0,13 %) para o conjunto total.

## IV. Pré-processamento dos Dados

Durante o processo de obtenção dos dados, que inicia com a medição das grandezas elétricas junto aos transformadores, e passa pelo transporte pela rede dedicada, até a interceptação dos dados pelo servidor e seu subsequente armazenamento, diversos processos são efetuados, e protocolos de comunicação de dados são observados.

Devido a grande quantidade de processos envolvidos, a sensibilidade dos medidores, interferências externas entre outros, erros podem ocorrer.

À etapa de pré-processamento cabe preparar os dados para que estes possam ser adequadamente analisados [08].

Nesta etapa deve-se observar a existência de valores faltantes, *outliers*, amostragem irregular e tendência. Uma vez que as séries utilizadas no estudo possuem suas mensurações realizadas em espaços idênticos de tempo, não há necessidade de observações quanto amostragem irregular.

Extraindo o máximo e o mínimo se tornou possível identificar valores faltantes e aberrantes. Para que os valores fossem apontados como possivelmente aberrantes, adotou-se o limite de variação de três desvios padrão. Os casos identificados foram analisados individualmente.

Dados faltantes e *outliers* foram substituídos pela média do vizinho anterior com o vizinho posterior. Séries que continham sequencias de valores faltantes ou aberrantes, ou um número superior a quatro desses valores, não foram consideradas nos estudos.

## V. Validação de Clusters

Com o objetivo de validar a qualidade dos agrupamentos, e

verificar a melhor abordagem para efetuar a clusterização, foi utilizada a medida não supervisionada silhueta de cluster - SC, que deve retornar valores positivo e próximos de 1 para indicar alto nível de coesão na estrutura formada.

Mais precisamente, segundo Rousseeuw [9], valores menores que 0.25 indicam que nenhuma estrutura significativa foi encontrada, valores entre 0.265 e 0.50 indicam estruturas fracas e potencialmente artificiais, enquanto valores entre 0.51 e 0.70 correspondem a uma estrutura razoável, por fim, valores entre 0.71 e 1 indicam a ocorrência de estruturas fortes.

Para esta versão do trabalho a quantidade de clusters gerados para cada tipo de dia foi fixada em três, uma vez que os resultados obtidos com números maiores de clusters não foram expressivos, resultando em MAPEs similares aos encontrados utilizando o conjunto formado por três clusters por tipo de dia.

Para permitir a avaliação dos agrupamentos utilizados nas tarefas de previsão, foram calculadas as médias das SC para os três clusters de cada agrupamento baseado no dia da semana, conforme **Tabela 1**. Onde a clusterização baseada em múltiplos parâmetros é representada por CM e a clusterização com base na temperatura é representada por CT.

**Tabela 1 - Silhueta de Cluster dos Agrupamentos**

| | Silhueta de Cluster - SC | | | | | | |
|---|---|---|---|---|---|---|---|
| Dia | Seg. | Ter. | Qua. | Qui. | Sex. | Sáb. | Dom. |
| CM | 0.69 | 0.92 | 0.72 | 0.84 | 0.85 | 0.76 | 0.65 |
| CT | 0.02 | 0.07 | 0.10 | -0.04 | 0.27 | -0.12 | -0.09 |

A clusterização CM apresentou os valores de SC mais altos, apontando a ocorrência de cinco estruturas fortes, correspondentes à terça, quarta, quinta sexta e sábado, e duas razoáveis, domingo e segunda.

A clusterização baseada apenas na temperatura obteve os piores resultados, onde seis das sete clusterizações se mostram sem nenhuma estrutura significativa (domingo, segunda, terça, quarta, quinta e sábado), apenas a clusterização correspondentes as curvas de carda de sextas feiras apresentaram um resultado diferente, mesmo assim inexpressivo, indicando uma estrutura fraca e potencialmente artificial.

Conforme verificado na **Tabela 1**, a abordagem CM obteve os melhores resultados, indicando que tal estratégia é potencialmente boa para clusterizar este tipo de dados, em contrapartida, a abordagem CT apresentou resultados bastante ruins, onde nenhuma estrutura foi classificada como forte ou razoável, indicando assim, uma capacidade pobre de representar isoladamente este tipo de dados.

## VI. Avaliação das Previsões

A fim de avaliar o desempenho das previsões realizadas, o índice de Erro Percentual Absoluto Médio (MAPE) foi empregado.

Este índice é utilizado em diversos trabalhos e considerado por concessionárias de energia elétrica como uma forma padronizada para avaliar o desempenho de predições de carga. O MAPE é calculado de acordo com a Equação 2:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad (2)$$

onde $A_t$ é o valor real, $F_t$ é o valor de previsão e $n$ é o tamanho da série temporal. O MAPE retorna um valor que deve ser pequeno, indicando maior precisão do método de previsão.

## VII. Implementação Computacional

A codificação dos métodos apresentados foi realizada no ambiente interativo para computação numérica MATLAB® versão 7.10.0, nome que também designa a linguagem utilizada.

Para permitir a visualização, a análise computacional e estatística dos dados utilizados nos métodos de previsão, além do MATLAB® utilizou-se o pacote de software Weka (*Waikato Environment for Knowledge Analysis*), versão 3.6.

A implementação e os testes foram realizados num mesmo computador, configurado com Windows 7 Home Premium (spk 1), processador AMD Athlon (tm) II P340 Dual-Core 2.20GHZ e 4GB de memória RAM.

## VIII. Análise das Curvas de Carga

Recursos gráficos permitem que os dados sejam facilmente interpretados, possibilitando a identificação de possíveis padrões ou anomalias.

A **Fig. 2** apresenta as curvas de carga (em W) do inverno de 2013 divididas entre os dias da semana.

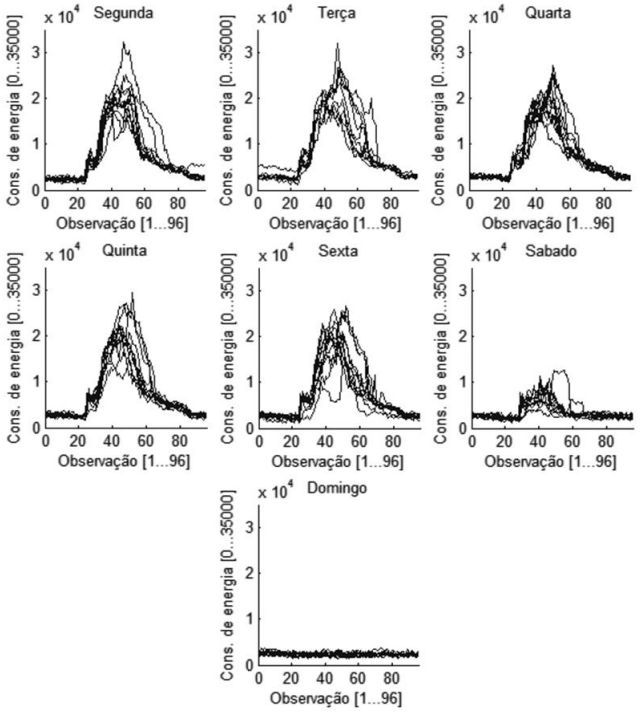



**Fig. 2 - Curvas de Carga do Inverno**

Antes que as curvas agrupadas possam ser representas por suas médias, é importante que elas tenham sido corretamente divididas, e que possíveis anomalias sejam detectadas e adequadamente tradas. A **Fig. 3** apresenta as médias (em W) das curvas de carga apresentadas no gráfico anterior.

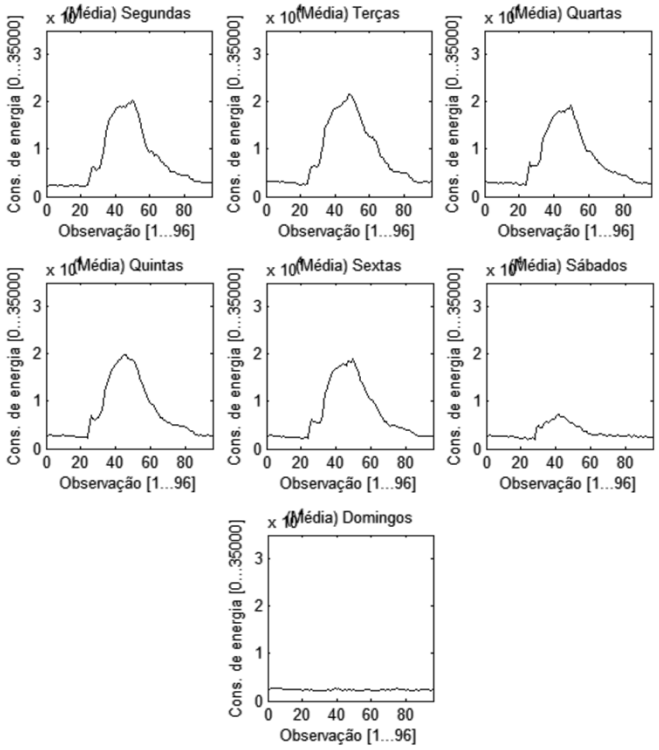



**Fig. 3 - Médias das Curvas de Carga do Inverno**

Ao observar a **Fig. 2** é possível identificar dias que se apresentam como atípicos, um gesto instintivo seria retira-los, entretanto antes de se tomar qualquer decisão, faz-se necessário verificar de maneira aprofundada tais dias.

A ocorrência de eventos festivos, feriados, paralisações, mudanças bruscas de temperatura entre outros episódios, podem ter ocasionado tal característica atípica. Atentar para essas questões permite evitar que dados consistentes sejam descartados.

Para permitir a comparação entre as curvas médias, realizou-se o cálculo do intervalo de confiança - IC de noventa e cinco por cento, possibilitando ainda, visualizar os trechos em que as curvas de carga apresentam maiores variações.

A **Fig. 4** apresenta o cálculo do intervalo de confiança (de 95%) de uma segunda feira. É possível observar que entre os instantes 40 e 50, e 59 e 71 ocorreram as maiores variações entre as curvas que compuseram a média.

A análise do IC permite perceber onde as curvas clusterizadas mais se diferenciam, possibilitando ainda traçar estratégia de correção para os pontos críticos (maiores distâncias entre o erro superior e inferior para um mesmo instante). Entre as estratégias que podem ser adotadas estão a modificação do número de clusters (ao aumentar ou diminuir é possível que se obtenham agrupamentos mais significativos), e a utilização da última curva similar registrada para a correção da curva de previsão.

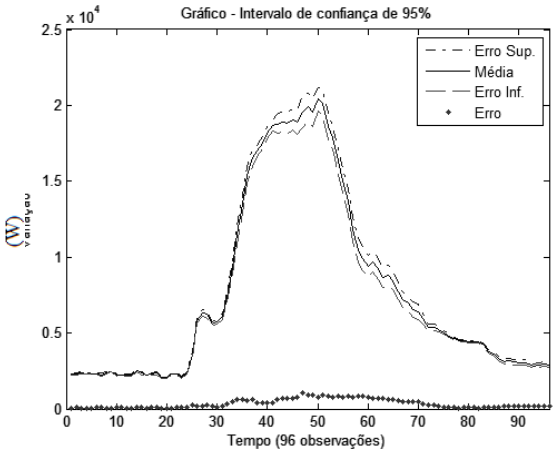



**Fig. 4 - Intervalo de Confiança de 95%**

## IX. Ensaios de Previsão

Por meio dos gráficos e do cálculo do MAPE foi possível verificar a assertividade dos três métodos codificados, e assim compara-los. A **Tabela 2** apresenta o cálculo do MAPE para os método de previsão MDS – 1, MDS – 2, MDS – 3A e MDS – 3B.

**Tabela 2 – Cálculo do MAPE**

| | MAPE (%) | | | |
|---|---|---|---|---|
| **Dias da Semana** | **Método** | | | |
| | MDS - 2 | MDS - 1 | MDS – 3B | MDS – 3A |
| Dom | 17,14 | 15,81 | 13,49 | **12,26** |
| Seg | **12,64** | 18,48 | 16,75 | 16,66 |
| Ter | **9,28** | 22,74 | 16,14 | 16 |
| Qua | 17,8 | 17,47 | **15,29** | 17,39 |
| Qui | 17,1 | 16,27 | **14,02** | 15,35 |
| Sex | **15,05** | 18,77 | 15,84 | 17,95 |
| Sáb | 29,99 | 35,15 | **26,11** | 31,56 |

O MDS – 3B representa a clusterização com base na temperatura média, no mínimo, no máximo e na média de cada uma das curvas de carga, já o MDS – 3A representa a clusterização com base na temperatura média.

Para facilitar a visualização dos resultados de previsão, determinados dados da **Tabela 1** foram destacados com negrito ou sublinhado. Os dados em negrito correspondem ao melhor MAPE entre os métodos, e os sublinhados ao segundo melhor resultado.

Em nenhuma das previsões realizadas o MDS - 1 apresentou o melhor ou o segundo melhor resultado.

O MDS – 3A apresentou resultados razoáveis, conseguindo apresentar a melhor previsão em um dos casos, e a segunda melhor em quatro casos.

O MDS - 2 se mostrou como o melhor em três das sete previsões, e como o segundo melhor em uma delas.

Foi o MDS – 3B que se destacou de maneira geral, mostrando-se superior aos outros. Obteve o melhor resultado em três dos sete casos, e o segundo melhor em dois dos sete casos.

A **Fig. 5** apresenta o desempenho do MDS - 2 e do MDS - 1 para previsão de uma quarta feira no inverno de 2013, com um

MAPE de 17,8 e 17,47 respectivamente.

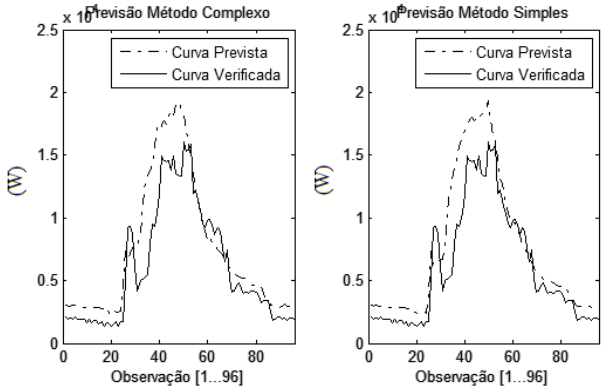



**Fig. 5 - Previsão: Métodos Convencionais**

A **Fig. 6** apresenta a previsão para mesma quarta feira abordada no gráfico anterior, demonstrando superioridade do método MDS – 3B, e o segundo melhor resultado ficando para o MDS – 3A, com MAPE's de 15,29 e 17,39 respectivamente.

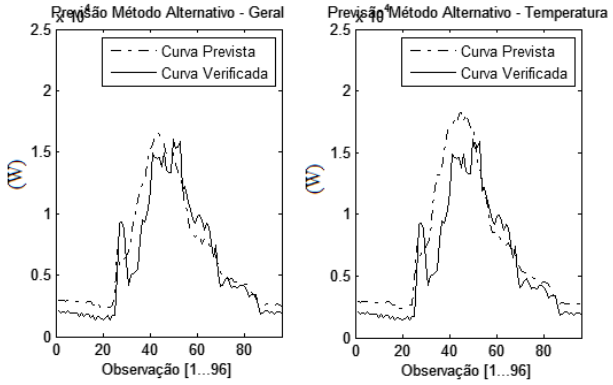



**Fig. 6 - Previsão: Método Alternativo**

Ao sobrepor as curvas verificadas com as curvas previstas juntamente com seus IC's, foi possível observar que uma parte considerável do erro da previsão era amenizado ao se considerar o erro superior e o inferior do IC, desta forma, os resultados do cálculo do MAPE poderiam ser diminuídos ao se considerar este intervalo.

## X. Considerações Finais

Em relação aos resultados de previsão apresentados, é importante notar a natureza diversa dos dados utilizados, originários do nível de consumo mais desagregado, onde os índices de incerteza são consideravelmente aumentados. De tal maneira, a tarefa de prever o comportamento de um consumidor específico não é trivial, e os índices de assertividade por vezes se fazem aumentados.

É importante destacar que o objetivo principal deste trabalho foi apresentar o uso de clusterização de dados, e não realizar previsões de carga, uma vez que a técnica pode ser utilizada com outros métodos de previsão.

A qualidade dos dados utilizados nos estudos exerce influência direta na capacidade de assertividade das previsões. Questões como a precisão dos medidores, a análise e o pré-processamento dos dados devem ser tratadas com atenção.

Os parâmetros de entrada também exercem influência direta nas previsões obtidas, desta maneira, a temperatura média prevista utilizada é capaz de alterar completamente os resultados de previsão, sendo necessário tratar com atenção esta questão, utilizando dados de fontes confiáveis.

Os resultados obtidos permitiram verificar que o MDS – 1 se mostrou como o menos assertivo nas previsões, isso até certo ponto era esperado, uma vez que seu nível de generalização é bastante alto.

O MDS – 3A ficou em terceiro entre os quatro, demonstrando que a clusterização baseada apenas na temperatura média não foi a melhor abordagem.

O MDS – 2 trouxe bons resultados, porém na média foi inferior que o MDS – 3B. O MDS – 3B apresentou o melhor conjunto de previsões, e ainda tem a vantagem de ter um tempo de execução significativamente inferior ao segundo colocado.

O local em que se insere a microrrede que forneceu os dados para este trabalho apresenta alta variação de temperaturas, registrando em uma mesma semana de inverno amplitudes térmicas próximas de 20°C. Acredita-se que as altas variações de temperatura, sobretudo no inverno, também tenham influenciado diretamente na capacidade de assertividade das previsões, causando MAPE's altos.

## XI. Referências


[1] Leone, K. M. A. 2006. **Previsão de carga de curto prazo usando ensembles de previsores selecionados e evoluídos por algoritmos genéticos**. Tese de Mestrado, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas. Disponível em: < http://www.bibliotecadigital.unicamp.br/document/?view=vtls000410708 >. Acesso em 12 de Novembro de 2012.

[2] Rahman, S.; Hazim, O. 1993. ***A generalized knowledge-based short-term load-forecasting technique***. IEEE Transactions on Power Systems, Volume 8, 2° Ed., 1993.

[3] Guirelli, C. R. 2006. **Previsão da Carga de Curto Prazo de Áreas Elétricas Através de Técnicas de Inteligência Artificial**. Tese apresentada a Escola Politécnica da Universidade de São Paulo - USP. São Paulo - SP, 2006, 127p. Disponível em: < www.teses.usp.br/teses/.../TESECLEBERROBERTOGUIRELLI.pdf >. Acesso em 08 de Dezembro de 2012.

[4] Kadowaki, M; Ohishi, T; Soares Filho, S; Lima, W. S. 2004. **Modelo de Previsão de Demanda de Carga de Curtíssimo Prazo para o Período da Ponta**. XXXVI Simpósio Brasileiro de Pesquisa Operacional - SBPO - O Impacto da Pesquisa Operacional nas Novas Tendências Multidisciplinares. São João del-Rei – MG. 2004.

[5] Senjyu, T; Higa, S; Uezato, K. 1998. ***Future load curve shaping based on similarity using fuzzy logic approach***. Generation, Transmission and Distribution, IEE Proceedings, 1998, 375-380.

[6] Pimenteli, E. P.; França, V. F.; Omar, N. **A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização**. In: Anais do Simpósio Brasileiro de Informática na Educação, Rio de Janeiro, RJ. SBC, 2003.

[7] Alencar, A. B. (2007). **Mineração e visualização de coleções de séries temporais**. Dissertação - Instituto de Ciências Matemáticas e de Computação - ICMC da Universidade de São Paulo - USP. São Carlos - SP.

[8] Neves, R. C. D.; Alvares, L. O. C. **Pré-processamento no processo de descoberta de conhecimento em banco de dados**. Dissertação. Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-Graduação em Computação. Rio Grande do Sul, 2003.

[9] Rousseeuw, P. J. (1987). **Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis**. Journal of Computational and Applied Mathematics, nº20.