UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉNICA
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA INDUSTRIAL – PEI

ADONIAS MAGDIEL SILVA FERREIRA

RECONHECIMENTO DE PADRÕES E TIPIFICAÇÃO DE PERFIS DE CONSUMO:
CONTRIBUIÇÕES PARA A MELHORIA DA GESTÃO NA DISTRIBUIÇÃO DA ENERGIA ELÉTRICA

Salvador
2015

ADONIAS MAGDIEL SILVA FERREIRA

RECONHECIMENTO DE PADRÕES E TIPIFICAÇÃO DE PERFIS DE CONSUMO:
CONTRIBUIÇÕES PARA A MELHORIA DA GESTÃO NA DISTRIBUIÇÃO DA ENERGIA ELÉTRICA

Tese apresentada ao Programa de Pós Graduação em Engenharia Industrial, Escola Politécnica, Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Doutor em Engenharia Industrial.

Orientadores:
Prof. Dr. Cristiano Hora de Oliveira Fontes
Prof. Dr. Carlos Arthur Mattos Teixeira Cavalcante

Salvador
2015

F383 Ferreira, Adonias Magdiel Silva

Reconhecimento de padrões e tipificação de perfis de consumo: contribuições para a melhoria da gestão na distribuição de energia elétrica/ Adonias Magdiel Silva Ferreira. – Salvador, 2015.

205 f.: il. color.

Orientadores: Prof. Cristiano Hora de Oliveira Fontes
Prof. Dr. Carlos Arthur Mattos Teixeira Cavalcante.

Tese (Doutorado) – Universidade Federal da Bahia. Escola Politécnica, Instituto de Matemática, 2015.

1. Energia elétrica - distribuição. 2. Análise por agrupamento. 3. Séries temporais multivariadas. 4. Séries temporais univariadas. I. Fontes, Cristiano Hora de Oliveira. II. Cavalcante, Carlos Arthur Mattos Teixeira. III. Universidade Federal da Bahia. IV. Título.

CDD: 621.31

TERMO DE APROVAÇÃO

ADONIAS MAGDIEL SILVA FERREIRA

RECONHECIMENTO DE PADRÕES E TIPIFICAÇÃO DE PERFIS DE CONSUMO: CONTRIBUIÇÕES PARA A MELHORIA DA GESTÃO NA DISTRIBUIÇÃO DA ENERGIA ELÉTRICA

Tese submetida ao Programa de Pós-Graduação em Engenharia Industrial da Universidade Federal da Bahia como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências em Engenharia Industrial

Aprovada em: 10 de abril de 2015.

Banca Examinadora:

Professor Cristiano Hora de Oliveira Fontes – Orientador_____
Doutor em Engenharia Química pela Universidade de Campinas – UNICAMP.
Universidade Federal da Bahia – UFBA.

Professor Carlos Arthur Mattos Teixeira Cavalcante – Orientador_____
Doutor em Engenharia de Produção pela Universidade de São Paulo – USP.
Universidade Federal da Bahia – UFBA.

Professor Caiuby Alves da Costa_____
Doutor em Electronique pela Université Paris – PARIS-SUD 11.
Universidade Federal da Bahia – UFBA.

Professor Daniel Barbosa_____
Doutor em Engenharia Elétrica pela Universidade de São Paulo – USP.
Universidade Federal da Bahia – UFBA e Universidade Salvador – UNIFACS.

Professor Kleber Freire da Silva_____
Doutor em Engenharia Elétrica pela Universidade de São Paulo – USP.
Universidade Federal da Bahia – UFBA e Universidade Salvador – UNIFACS.

Professor José Viriato Coelho Vargas_____
Doutor em Engenharia Mecânica pela Duke University – Duke, Estados Unidos.
Universidade Federal do Paraná – UFPR.

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA INDUSTRIAL

PEI

Ata da Sessão Pública de Defesa de Tese de Doutorado - Colegiado do Curso de Doutorado em Engenharia Industrial –UFBA

Ata da sessão especial do Colegiado do Curso de Doutorado em Engenharia Industrial da Universidade Federal da Bahia, realizada em dez de abril de dois mil e quinze, para instalação da Banca Examinadora de Tese de Doutorado em Engenharia Industrial do candidato **Adonias Magdiel Silva Ferreira** intitulada ***"Reconhecimento de Padrões e Tipificação de Perfis de Consumo: Contribuições para a melhoria da Gestão na Distribuição da Energia Elétrica"***. Às 14h do citado dia, na sala de vídeo conferência Hernani Sobral, 6º andar da Escola Politécnica, realizou-se a Sessão Pública do Colegiado do Curso de Doutorado em Engenharia Industrial para instalação da Banca Examinadora de Tese de Doutorado em Engenharia Industrial do referido candidato. Compareceram à sessão os seguintes **membros da Banca Examinadora: Prof. Dr. Carlos Arthur Mattos T. Cavalcante (PEI-UFBA – orientador); Prof. Dr. Cristiano Hora de Oliveira Fontes (PEI-UFBA – orientador); Prof. Dr. Caiuby Alves da Costa (DEE-UFBA); Prof. Dr. Daniel Barbosa (UNIFACS); Prof. Dr. José Viriato Coelho Vargas (UFPR); Prof. Dr. Kleber Freire da Silva (UNIFACS).** Instalada a Banca Examinadora, foram esclarecidos os procedimentos e foi passada a palavra ao examinador para apresentação do trabalho de tese. Ao final da apresentação, passou-se à arguição por parte da Banca, a qual, em seguida, reuniu-se para a elaboração do parecer. No retorno, foi lido o parecer a respeito do trabalho apresentado pelo candidato, tendo a Banca Examinadora conferido o resultado de:

☒**APROVADO COM REVISÕES E POSTERIOR DEFERIMENTO DOS ORIENTADORES.**

☐**APROVADO COM RESTRIÇÕES PARA POSTERIOR ANÁLISE DA BANCA.**

☐**REPROVADO.**

Em seguida, nada mais havendo a tratar, foi encerrada a sessão pela Profa. Karen Pontes, Coordenadora do Colegiado do Curso, tendo sido, logo a seguir, lavrada a presente ata que é assinada abaixo pelos membros da Banca Examinadora.

Salvador, 10 de abril de 2014

Prof. Dr. Carlos Arthur Mattos T. Cavalcante (PEI-UFBA – orientador)

Cristiano Hora de Oliveira Fontes – Orientador
Doutor em Engenharia Química pela Universidade de Campinas – UNICAMP.

Prof. Dr. Caiuby Alves da Costa (DEE-UFBA)

Prof. Dr. Daniel Barbosa (UNIFACS)

Prof. Dr. José Viriato Coelho Vargas (UFPR)

Prof. Dr. Kleber Freire da Silva (UNIFACS)

Espaço reservado à coordenação do curso:
Defesa homologada em reunião do Colegiado em 15/05/2015 **tendo como resultado final** APROVADO

Márcio L. F. Nascimento
Vice-Coordenador do PEI-UFBA

Dedico este trabalho aos meus pais ADONIAS FERREIRA (*in memorian*) e ZILDETH SILVA FEREIRA por terem se doado superando às vezes o impossível para permitir o meu acesso à educação, apresentando-me o caminho da cidadania, dignidade e honestidade.

Aos meus irmãos e parentes por terem vivenciados juntos momentos importantes da minha vida.

Aos meus sogros BENEDITO DA CONCEIÇÃO DOS ANJOS e ANALICE SANTOS DOS ANJOS que me ajudaram a ter a tranquilidade necessária para que eu pudesse concluir esta tese.

A minha esposa ALANA MARA SANTOS DOS ANJOS FERREIRA e meu filho ADONIAS MAGDIEL SILVA FERREIRA JR. que vivenciaram comigo momentos de incertezas, muito comuns para quem tenta trilhar novos caminhos, esta conquista também pertence a vocês.

AGRADECIMENTOS

São muitos, e tão singulares... mas peço compreensão se alguma falta cometer.

Ao Prof. Cristiano Hora de Oliveira Fontes, amigo e orientador, sua diligência, abnegação e motivação me impulsionou continuamente, transformando minhas dúvidas e angústias em pontos de partidas para as necessárias descobertas e avanços do conhecimento, suas colaborações, conselhos, palavras de incentivos foram determinantes para minha trajetória.

Ao Prof. Carlos Arthur Mattos Teixeira Cavalcante, amigo e orientador, pelo acolhimento no Laboratório de Sistemas de Integrados de Produção – LABSIP e no Grupo de Pesquisa MODELE, também suas colaborações, conselhos, palavras de incentivos foram muito úteis.

Aos professores da banca examinadora que aceitaram o convite para colaborar com suas experiências e competências.

Aos professores Karla Esquerre, Caiuby Alves e o Diretor da NORSUL Jorge Marâmbio, pelas colaborações dadas durante a qualificação do projeto de tese.

Ao Programa de Pós-Graduação em Engenharia Industrial (PEI) da UFBA onde tive o privilégio de ser doutorando e conviver com funcionários, colegas e professores simpáticos e comprometidos com a excelência do curso.

A Escola Politécnica minha unidade de trabalho na UFBA.

Ao Prof. Geraldo Queiroz pelo incentivo e apoio.

Ao Departamento de Engenharia Mecânica, ambiente de trabalho onde tenho a honra de conviver com colegas atenciosos e solidários.

Ao LABSIP, pelo apoio, infraestrutura, qualidade, motivação e companheirismo dos professores, pesquisadores, bolsistas de iniciação científica e pessoal de apoio que participam ou já participaram de projetos desenvolvidos neste espaço, sobretudo, àqueles que também contribuíram para a conclusão desta tese.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelos recursos aportados que aumentaram a qualidade da minha disponibilidade para a pesquisa.

A PETROBRÁS que possibilitou através de convênio a melhorar a infraestrutura do LABSIP.

A ELETROBRÁS que disponibilizou os dados para a pesquisa.

A NORSUL que possibilitou através da sua expertise dar o suporte competente para as medições dos dados que constituíram as amostras das pesquisas de campo desta tese.

A todos aqueles os quais tive o privilégio de ministrar aulas, orientar trabalhos acadêmicos, pesquisar juntos, conviver... ao longo desta trajetória.

Sou grato pelo amor e apoio da minha família.

Sou sempre grato a Deus pelos insucessos e sucessos, justiças e injustiças que vivi, porque juntos trouxeram-me maturidade e produziram em mim o fruto da perseverança.

Finalmente, esta conquista só foi possível graças à colaboração e solidariedade de todos que convivi ao longo desta jornada, incluindo-se também aqueles que por qualquer lapso não foram aqui citados, que de uma forma direta ou indireta contribuíram para a realização desta tese.

"Àquele que é poderoso para nos guardar de tropeços e para nos apresentar irrepreensíveis, com alegria diante da sua glória, ao único Deus, nosso Salvador, mediante Jesus Cristo, Senhor nosso, glória, majestade, império e soberania, antes de todas as eras, e agora, e por todos os séculos. Amém".

Judas 24-25.

RESUMO

Os consumidores residenciais apresentam uma diversidade de hábitos no uso da energia elétrica e um dos maiores desafios é o de prever a demanda a fim de equalizar a oferta com o consumo. Neste sentido, o desenvolvimento de métodos de agrupamento baseados no reconhecimento de padrões de consumo é de fundamental importância no gerenciamento da eficiência do setor elétrico. O objetivo deste trabalho é propor um método constituído de dois algoritmos: um aplicável a séries temporais univariadas e outro aplicável a séries temporais multivariadas, ambos desenvolvidos para o reconhecimento de padrões de séries temporais com o mesmo número de pontos amostrados e para o mesmo período de observação. Os dados estão relacionados aos programas de eficiência energética implementados por duas empresas distribuidoras de energia elétrica, e em ambos os casos, a substituição de refrigeradores se referiu às residências de consumidores de baixa renda das distribuidoras. Os dados foram coletados diretamente dos refrigeradores das unidades consumidoras antes (caso I) e depois (caso II) da substituição dos equipamentos. Refrigeradores novos foram doados às unidades consumidoras em substituição aos equipamentos antigos com menores recursos tecnológicos e prazos de vida útil esgotados. Dentre as opções indicadas pelo Protocolo Internacional para Medição e Verificação de Performance (PIMVP), este trabalho se enquadra na Opção B. A coleta de dados teve como alvo as medições das grandezas da potência elétrica e temperatura associadas aos refrigeradores, e respectivamente, os sistemas de medição SAGA 2000 e Termohigrômetro Extech RHT 10 foram utilizados nos períodos antes e após substituição. O algoritmo *FCM* (*Fuzzy C-Means*) foi utilizado como referência comparativa tanto para a versão univariada quanto para a versão multivariada, sendo que, para a versão multivariada, foi adotada uma versão modificada do *FCM* baseada em uma métrica de similaridade que utiliza componentes principais (*Similarity Principal Componente Analysis – SPCA*). Na versão univariada, antes das substituições (caso I), o método proposto e o *FCM* tiveram, respectivamente, um índice global de silhueta (medida da qualidade de agrupamento) de 0,28 e 0,25. Após as substituições (caso II) dos refrigeradores, ambos métodos reconheceram a existência de apenas um grupo e padrões semelhantes. Na versão multivariada, no caso I, o método proposto teve o índice global de silhueta de 0,19, e no caso II, o índice global de silhueta foi de 0,46. Os índices obtidos pelo método *FCM* no caso I e no caso II, respectivamente, foram de -0,12 e 0,21. O método proposto apresentou uma identificação de uma maior diversidade de padrões; o reconhecimento da sazonalidade através de uma abordagem multicritérios; o melhoramento da tomada de decisão através de uma melhor classificação dos perfis de consumidores heterogêneos; e a definição do número de clusters através de uma abordagem baseada em grupos semi-hierárquica, revelando-se assim como uma importante contribuição para o estado-da-arte. A partir dos desafios e resultados obtidos neste trabalho, são sugeridas possibilidades de trabalhos futuros que incorporem análises para a previsão de diferentes dinâmicas temporais, período de amostragem e incerteza de medição.

Palavras-chave: Agrupamento, Distribuição de Energia Elétrica, Séries Temporais Multivariadas, Séries Temporais Univariadas, Tipificação de Cargas.

ABSTRACT

The Residential customers have a diversity of habits in the use of electric power and one of the biggest challenges is to forecast demand in order to equalize the supply to consumption. In this sense, the development of clustering methods based on the recognition of patterns of consumption is of fundamental importance in managing the electricity sector efficiency. The objective of this work is to propose a method consists of two algorithms: one applicable to univariate time series and another applicable to multivariate time series, both developed for the recognition of time series patterns with the same number of sample points for the same period observation. The data are related to energy efficiency programs implemented by two distributors of electricity, and in both cases, the replacement of refrigerators referred to the homes of low-income consumers of the distributors. Data were collected directly from the refrigerators of consumer units before (case I) and after (case II) of the equipment replacement. New refrigerators were donated to consumer units to replace old equipment with lower technological resources and exhausted useful lives. Among the options listed by the International Protocol for Performance Measurement and Verification (IPMVP), this work fits in option B. The data collection had how target the measurements of quantities of electric power and temperature associated with refrigerators, and respectively, the measurement systems SAGA 2000 and Thermohygrometer Extech RHT 10 were used in the periods before and after replacement. The FCM (Fuzzy C-Means) algorithm was used as a comparative reference for both version (univariate and multivariate), being that to the multivariate was used a modified version of the *FCM*, based on a similarity metric that uses principal component (Similarity Principal Component Analysis - *SPCA*). In the univariate version before the replacement (case I), the proposed method and FCM have, respectively, a global silhouette index (clustering quality measurement) of 0.28 and 0.25. After replacements of refrigerators (case II), both methods have recognized the existence of only one group and similar patterns. In the multivariate version, in the case I, the proposed method had the global silhouette index of 0.19, and in the case II, the global silhouette index was 0.46. The index obtained by the *FCM* method in case I and case II, respectively, were -0.12 and 0.21. The proposed method showed an identification of a greater diversity of patterns; recognition of the seasonality by a multi-criteria approach; improvement of decision-making, through better classification of heterogeneous consumer profiles; and the definition of the number of clusters through an approach based on semi-hierarchical cluster, revealing itself as an important contribution to the state of the art. From the challenges and results of this study, possibilities for future work incorporating analysis for predicting different temporal dynamics, sampling period and measurement uncertainty are suggested.

Key Words: Clustering, Electricity Distribution, Multivariate Time Series, Pattern of loads, Univariate Time Series.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ACR	Ambiente de Contratação Regulada
ANEEL	Agência Nacional de Energia Elétrica
ANN	*Artificial Neural Network*
ANOVA	*Analysis of Variance*
ARIMA	*Autoregressive Integrated Moving Average*
ARX	*Autoregressive Exogenous*
ASHRAE	*American Society of Heating, Refrigerating and Air-Conditioning Engineers*
AVQ	*Adaptive Vector Quantization*
BIC	*Bayesian Information Criterion*
BSVs	*Bounded Support Vectors*
CCA	*Canonical Correlation Analysis*
CCEE	Câmara de Comercialização de Energia Elétrica
CDI	*Clustering Dispersion Indicator*
CEAL	Companhia Energética de Alagoas
CEMAR	Companhia Energética do Maranhão
CLP	*China Light and Power*
CMSE	Comitê de Monitoramento do Setor Elétrico
CNPE	Conselho Nacional de Políticas Energéticas
DTW	*Dynamic time Warping*
EANN	*Ensemble Artificial Neural Network*
ELM	*Extreme Learning Machine*
EM	*Expectation-Maximization*
EPE	Empresa de Pesquisa Energética
ESCO	*Energy Services Company*
FCM	*Fuzzy C-Means*
fMRI	*functional Magnetic Resonance Imaging*
GSI	*General Silhouette Index*
HISMOOTH	*hierarchical smoothing models*
HMM	*Hidden Markov Models*
ICL	*Integrated Completed Likelihood*
IPCL	*Incrementally Characterizes Patterns*
IRC	*Iterative Refinement Clustering*
ISPC	*Incremental Summarization and Pattern Characterization*
ISPC	*Incremental Summarization and Pattern Charaterization*
KDD	*Knowledge-Discovery in Databases*
MANOVA	*Multivariate Analysis of Variance*
MC	*Markov Chains*
MIA	*Mean Index Adequacy*
MMNF	*Min-Max Neuro-Fuzzy*
MNE	Ministério de Minas e Energia
NTL	*Non-Technical Loss*
ONS	Operador Nacional do Sistema Elétrico
OS	*Online Sequencial*
PCA	*Principal Component Analysis*
PDM	*Pattern decomposition method*
PEE	Programa de Eficiência Energética
PIMVP	Protocolo Internacional de Medição e Verificação de Desempenho

PLD	Preço de Liquidação de Diferenças
PNN	*Probabilistic Neural Network*
PROCEL	Programa Nacional de Conservação de Energia Elétrica
SCRA	*Sequence Cluster Refinement Algorithm*
SIN	Sistema interligado Nacional
SOM	*self-organizing maps*
SPCA	*Similarity Principal Componente Analysis*
STAC	Seleção, Tipificação e Agrupamento de Curvas de cargas
STM	Séries Temporais Multivariadas
STU	Séries Temporais Univariadas
SVC	*Support Vector Clustering*
SVM	*Support Vector Machine*
TNB	*Tenaga Nasional Berhad*
UC's	Unidades Consumidoras
WEACS	*Weighted Evidence Accumulation Clustering*

LISTA DE SÍMBOLOS

n	Objetos numéricos
X	Conjunto de objetos numéricos
x_i	Objeto ou ponto no espaço real multidimensional
$\Re^p$	Espaço real multidimensional
p	Dimensionalidade
x_{ik}	*k-ésimo* atributo associado a um objeto
X'	Vetor de variáveis aleatórias
$\underset{\sim}{\mu}$	Média um vetor de variáveis aleatórias
Σ	Matriz de variância/covariância
Y	Autovetores
$\underset{\sim}{a}'_j$	Vetor de constantes
A	Matriz dos autovetores
$\tilde{Y}$	Vetor das componentes principais
λ_j	Autovalores
Ψ	Matriz dos autovalores
F	Fatores comuns
μ_i	Fator único
ε_i	Fator de erro do modelo
l_{ik}	Cargas fatoriais
h_i^2	Cumunalidade
L	Matriz de cargas fatoriais
Z_i	Escores fatoriais
μ_i	Média
σ_i	Desvio padrão
$\widehat{\lambda}_i$	Autovalores com máxima proporção da variância
$\widehat{e}_i$	Autovetor normalizado
R_{pxp}	Matriz de correlação amostral
$\widehat{\Psi}$	Matriz dos autovalores com máxima proporção da variância
Δ	Matriz quadrada de dissimilaridade
δ_{ij}	Elementos de uma matriz quadrada de dissimilaridade
X	Matriz de elementos escalonados
D	Matriz quadrada de dissimilaridade dos elementos escalonados
d_{ij}	Medidas de distâncias
χ^2	Estatística da distribuição do qui-quadrado
t	Estatística da distribuição de *Student*
X	Variável aleatória
ν	Graus de liberdade
H_0	Hipótese nula
H_1	Hipótese alternativa
f_j	Frequência observada
$\hat{f}_j$	Frequência esperada
r	Coeficiente de correlação amostral
s	Desvio padrão amostral
$S_{y_1y_2}$	Variância combinada

V_i	Variáveis dependentes
W_j	Variáveis independentes
$Corr$	Correlação canônica
N	Tamanho da amostra
P	O número de variáveis dependentes
Q	O número de variáveis independentes
$MQEG$	Média de quadrados entre os grupos
$MQDG$	Média de quadrados dentro dos grupos
F	Estatística da distribuição de Fisher
K	Número de variáveis
N	Número de casos
λ	Estatística da distribuição de Wilks
W	Dispersão interna do grupo
T	Dispersão total
B	Dispersão entre grupos
U	Matriz de pertinência
V	Vetores com os centros
W	Vetores com as penalidades
m	Grau de *fuzzyficação*
Z_i	Matriz de objetos de séries temporais
$z_{ij}(t)$	Medição da variável de um objeto em um instante de tempo
t	Instante de tempo
W	Matriz de distância sob restrição de uma alinhamento temporal
P	Matriz de probabilidade de transição
$f_T(\lambda_s)$	Estimadores de matrizes espectrais de séries estacionárias
λ_s	Espectro de séries
T	Períodos das séries
$\rho^2_{i,j}(\tau)$	Função de correlação cruzadas para defasagens no tempo
τ	Defasagens no tempo
$\rho(\tau)$	Função de autocorrelação
$s_{i,j}$	Quantificação da similaridade entre duas séries temporais
N_K	Objetos de um grupo
L	Grupo de objetos
G	Total de grupos
s_i^L	Índice de silhueta do objeto de um grupo
I_s	Índice de silhueta de um objeto
a_i^L	Distância média entre objetos de um mesmo grupo
b_i^L	Mínima distância média entre objetos de diferentes grupos
I_s	Índice de silhueta de um objeto
IGS	Índice Global de Silhueta ou média aritmética dos índices de silhueta entre todos os objetos
α	Coeficiente de ponderação
$a'(q)$	Primeiro maior elemento de uma coluna da matriz de pertinência
$b'(q)$	Segundo maior elemento de uma coluna da matriz de pertinência
q	Um coluna da matriz de pertinência
min	Unidade da grandeza de base do tempo e corresponde a 60 segundos no Sistema Internacional de Unidades (SI)
pu	Unidade adimensional e corresponde a 1 no SI.

h	Unidade da grandeza de base do tempo (t) e corresponde a 3600 segundos no SI
P_i	Potencial do objeto em relação a um centro de grupo
r	Constante positiva
ε	Tolerância
kWh	Energia elétrica em quilowatt hora e corresponde a 3600000 J ou 3600000 $m^2.kg.s^{-2}$ no SI
m	Comprimento em metro unidade da grandeza de base no SI
kg	Massa em quilograma unidade da grandeza de base no SI
s	Tempo em segundo unidade de grandeza de base no SI
W	Potência elétrica em watt ou joule por segundo unidade de grandeza derivada no SI e sua representação na grandeza de base é $m^2.kg.s^{-3}$
ºC	Temperatura em Celsius sua unidade de grandeza de base no SI é o K e zero grau Celsius equivale a 273,15 kelvins
K	Temperatura termodinâmica em kelvin unidade de grandeza de base no SI
Y	Matriz de dados
$\Re^{v \times w}$	Espaço real com dimensionalidade $v \times w$
v	Objetos
w	Rótulos dos objetos
t_i	Vetor de escores das componentes principais
p_i	Componente principal
E	Matriz de resíduos
$SPCA_{pq}$	Medida de similaridade entre os objetos de componentes principais
θ_{ji}	Ângulo formado entre duas componentes
D	Matriz de dissimilaridade
K	Matriz de dissimilaridade transformada
V	Matriz de autovetores da matriz K
Λ	Matriz de autovalores da matriz K
kW	Potência em quilowatt
COP^{-1}	Inverso do coeficiente de performance sua representação na unidade de grandeza de base é K/K
T_o	Temperatura externa em K
T_f	Temperatuara interna em K
$a\,(k)$	Função de defasagem do sinal de entrada
$b\,(k)$	Função de defasagem do sinal de saída
$v\,(k)$	Ruído branco
k	Instante de tempo
q^{-1}	Operador de defasagem
n	Ordem de defasagem
PC	Consumo de energia elétrica

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

No Brasil, a gestão da oferta e da demanda do setor elétrico está sob a égide de marcos regulatórios que visam melhorar as relações entre o Estado, as empresas e os consumidores, presentes nas atividades de geração, transmissão, distribuição e comercialização da energia elétrica. Particularmente, no cenário pós-racionamento ocorrido no ano 2001, a demanda desacelerou significativamente e a iniciativa privada imediatamente pressionou o Estado para estabelecer regras visando o equilíbrio econômico-financeiro das empresas atuantes no setor. Por exemplo, o marco regulatório implementado pela Lei 10.848, de 15/03/2004, e regulamentado pelo Decreto 5.163, de 30/07/2004, passou a considerar as especificidades do setor elétrico nacional, e contribuiu para a consolidação do papel regulador do Estado e para a privatização das empresas do setor, através do estabelecimento de uma rede de agentes responsáveis pelo planejamento e operação do sistema elétrico brasileiro.

Tanto em termos estratégicos quanto em termos operacionais, a finalidade precípua desta rede de agentes é a gestão integrada da produção e da demanda, atual e futura, de energia elétrica. Seus agentes integrantes são: o Conselho Nacional de Políticas Energéticas – CNPE, o Ministério de Minas e Energia – MNE, a Empresa de Pesquisa Energética – EPE, o Comitê de Monitoramento do Setor Elétrico – CMSE, a Agência Nacional de Energia Elétrica – ANEEL, a Câmara de Comercialização de Energia Elétrica – CCEE e o Operador Nacional do Sistema Elétrico – ONS (ANEEL, 2000; ANEEL, 2008; ANEEL, 2010; FERREIRA et al., 2010).

O CNPE é o órgão interministerial de assessoramento da Presidência da República que tem como principais atribuições formular a política energética e assegurar a disponibilidade de infraestruturas necessárias ao atendimento das demandas por energia elétrica do país, inclusive nas áreas mais remotas ou de difícil acesso.

O MME é o ministério do Governo Federal responsável pela condução da política energética do país, de acordo com as diretrizes definidas pelo CNPE. A EPE, instituída pela Lei nº 10.847/04 e criada pelo Decreto nº 5.184/04, é uma empresa vinculada ao MME cuja finalidade é realizar estudos e pesquisas destinados a subsidiar o planejamento do setor

energético. O CMSE é um órgão sob a coordenação direta do MME com a função de acompanhar e avaliar a continuidade e a segurança do suprimento elétrico em todo o território nacional.

A ANEEL, instituída pela Lei nº 9.427/96 e constituída pelo Decreto nº 2.335/97, tem as atribuições de regular e fiscalizar a produção, transmissão, distribuição e comercialização de energia elétrica, estabelecendo os valores das tarifas para os consumidores finais, zelando pela qualidade dos serviços prestados e pela universalização do atendimento, sempre preservando a viabilidade econômica e financeira dos agentes e da indústria.

A CCEE foi instituída pela Lei nº 10.848/04 e criada pelo Decreto nº 5.177/04, e entre suas principais obrigações estão a apuração do Preço de Liquidação de Diferenças (PLD), utilizado para valorar as transações realizadas no mercado de curto prazo; a contabilização dos montantes de energia elétrica comercializados; a liquidação financeira dos valores decorrentes das operações de compra e venda de energia elétrica realizadas no mercado de curto prazo e a realização de leilões de compra e venda de energia no Ambiente de Contratação Regulada – ACR, por delegação da ANEEL.

O ONS, criado pela Lei nº 9.648, de 27/05/1998 e regulamentado pelo Decreto nº 2.655 de 02/07/1998, com as alterações do Decreto nº 5.081 de 14/05/2004, tem por atribuições operar, controlar e supervisionar a geração de energia elétrica no Sistema interligado Nacional – SIN, e administrar a rede básica de transmissão de energia elétrica no Brasil. Tem por objetivos principais atender as demandas de carga, otimizar custos operacionais e garantir a confiabilidade operacional do SIN.

Esta estrutura organizacional foi concebida para equalizar os problemas e as dificuldades inerentes ao setor recentemente privatizado, no qual é comum a existência de conflitos de interesses entre o Estado, os investidores e os consumidores. O Estado, ao mesmo tempo em que incentiva a eficiência do sistema, busca também ampliar os benefícios para o consumidor, principalmente na forma de políticas sociais. Os consumidores, por sua vez, têm um claro interesse pela contínua disponibilidade a preços satisfatórios. Por outro lado, os investidores têm interesse em maximizar os ganhos de seus investimentos com o menor risco possível na operação do sistema elétrico, conflitando com o interesse do Estado de promover de políticas sociais.

A ANEEL, entre outras responsabilidades tem a incumbência pela equalização dos interesses do Estado, dos investidores e dos consumidores no tocante à regularização e

fiscalização do sistema elétrico brasileiro, buscando a conciliação da oferta com a demanda real e potencial e a confiabilidade e qualidade dos serviços prestados, bem como a universalização do atendimento.

O sistema elétrico brasileiro é formado pelos sistemas de geração, transmissão e distribuição de energia, sendo esta última planejada de acordo com as previsões de demanda, previsões estas que são dificultadas pela presença de uma alta diversidade de modos e intensidades de utilização da energia elétrica, e também pela diversidade de agentes envolvidos e de variáveis de decisão consideradas.

Na geração, grandes usinas produzem a energia elétrica que é então transmitida para subestações de energia onde a tensão é elevada e então enviada para o sistema de transmissão de alta tensão. Na transmissão a energia gerada em alta tensão é recebida pela subestação mais próxima ao centro de consumo, onde sofre uma redução de tensão antes de seguir para as linhas de distribuição e ser entregue ao consumidor por meio das linhas de serviço, nas tensões nominais de uso final. De acordo com Stevenson (1986), a geração de energia elétrica ocorre geralmente próxima aos recursos naturais, como é o caso da hidroelétrica, que representa a matriz elétrica mais utilizada no Brasil. Dentre as outras formas de geração utilizadas, destacam-se as usinas termelétricas que são prioritariamente instaladas próximas às reservas de combustíveis fósseis como o carvão ou gás natural. Esta tendência de localizar as usinas de energia elétrica próximas aos recursos energéticos naturais justifica-se pela redução nos custos de transporte dos insumos de geração.

A energia transmitida em alta tensão é recebida pela subestação mais próxima ao centro de consumo (cidades e parques industriais), onde sofre uma redução da tensão para atender a várias classes de tensão, e antes de seguir para as linhas de distribuição e entregue ao consumidor por meio das linhas de serviço, nas tensões nominais de uso final. Segundo Júnior (2006), dessa última tarefa encarregam-se as empresas distribuidoras (as companhias energéticas) que fornecem energia aos consumidores finais, geralmente classificados nas categorias residencial, comercial ou industrial.

O consumo diário de energia elétrica nas unidades residenciais apresenta diferentes padrões de uso devido à presença de diferentes classes de consumidores, à existência de diferentes níveis tarifários, o que resulta em diferentes comportamentos de demanda. Segundo Gerbec et al (2002), os consumidores residenciais apresentam uma diversidade de hábitos no

uso da energia elétrica e um dos maiores desafios é o de prever a demanda a fim de equalizar a oferta com o consumo.

1.2 JUSTIFICATIVA

O desenvolvimento de métodos de agrupamento baseados no reconhecimento de padrões de consumo é de fundamental importância no gerenciamento da eficiência do setor elétrico. De uma forma geral, os métodos de agrupamento, ou de reconhecimento de padrões aplicados ao consumo de energia elétrica, propõem-se a identificar os padrões de consumo intrínsecos aos grupos de uma amostra de dados a partir da extração de características desta última (Gan et al., 2007). Isto resulta na obtenção de grupos semelhantes entre si (homogêneos) e com altas dissimilaridades (heterogeneidade) entre eles, permitindo assim um melhor gerenciamento da eficiência do sistema elétrico.

O gerenciamento da eficiência do sistema elétrico pelo lado da demanda é de fundamental importância na tomada de decisões estratégias visando o uso mais eficiente da energia elétrica, a maximização da capacidade de fornecimento e o aumento da lucratividade e a rentabilidade das companhias energéticas. No Brasil, conforme dispõe a Lei no 9.991, de 24 de julho de 2000, as companhias energéticas devem aplicar um percentual mínimo da receita operacional líquida em Programas de Eficiência Energética – PEE regulados pela Agência Nacional de Energia Elétrica – ANEEL.

Segundo Hordeski (2005) a eficiência energética é um termo genérico e existem várias medidas de eficiência energética com diferentes finalidades e aplicações. De forma geral, a eficiência energética refere-se à utilização de menos energia para produzir a mesma quantidade de serviços ou bens. Conforme a *International Energy Agency* (*IEA*, 2007), a eficiência energética é uma função que associa o montante de energia final consumida (gás natural, carvão ou eletricidade) com o serviço energético ofertado (produção, transporte, calor e frio). Em particular, Patterson (1996) define que a eficiência energética térmica mede a relação de consumo de energia real com o consumo de energia ideal ou teórico.

As companhias energéticas brasileiras têm buscado a eficiência do sistema de distribuição através, entre outros, de projetos que incentivam a troca de um grande número de equipamentos por outros comprovadamente mais eficientes, principalmente em regiões com baixos índices de eficiência no consumo de energia elétrica.

Especificamente, o já mencionado Programa de Eficiência Energética (PEE), regulado pela ANEEL, estabelece que as companhias energéticas devem aplicar 0,5% da receita operacional líquida em projetos de fomento à eficiência e pesquisa & desenvolvimento. De acordo com os dados da ANEEL, aproximadamente 25% da quantidade dos projetos de eficiência energética são da tipologia Baixa Renda.

Estes projetos são dirigidos a unidades consumidoras de baixo poder aquisitivo visando: fomentar a substituição de equipamentos ineficientes; incentivar a realização de palestras educativas e de atividades voltadas para o combate ao furto de energia e ao estímulo do seu uso eficiente e seguro; regularizar consumidores clandestinos, mediante a instalação de ramal de ligação até o ponto de entrega ao consumidor; e promover reformas/instalações nos equipamentos de entrada e instalações internas dessas unidades consumidoras.

Além de consumidores residenciais, atendem também unidades consumidoras de cunho filantrópico/assistenciais, associações de bairro, creches, escolas, hospitais públicos e afins, desde que não exerçam atividades com fins lucrativos e estejam localizadas geograficamente nas comunidades atendidas, caracterizando atendimento predominante aos consumidores ali residentes. As unidades comerciais podem ser contempladas também, mas a concessionária ou permissionária deve descrever e justificar no relatório final os critérios utilizados na seleção e caracterização das unidades comerciais beneficiadas.

Ações coordenadas de troca de refrigeradores merecem o devido destaque pelo fato deste equipamento ser responsável por cerca de 25% a 30% do consumo mensal de energia elétrica dos projetos de eficiência energética da tipologia Baixa Renda. Segundo a ANEEL (2012) o refrigerador afeta o comportamento da demanda de energia elétrica residencial e qualquer prejuízo na sua eficiência contribui para a diminuição da eficiência do sistema elétrico brasileiro. Tal como em outros equipamentos, o rendimento de um refrigerador se reduz com o tempo. Em geral, a troca desses equipamentos em condições subnormais por um novo não é um tipo de ação que conta com adesão expressiva dos consumidores de baixa renda por causa de fatores socioeconômicos justificáveis. Isto tem acarretado, para o sistema como todo, um aumento no consumo de energia que pode ser evitado.

Diante disso, a análise dos dados de consumo de energia elétrica medido em unidades residenciais é um importante recurso de monitoramento das demandas específicas de clientes (TSEKOURAS et al, 2011), fazendo com que o reconhecimento de padrões de consumo viabiliza a extração de informações úteis para o desenvolvimento de ferramentas de apoio à

tomada de decisão, e bem assim para a melhoria dos sistemas de produção e das tecnologias de gestão do setor elétrico (LIN et al., 2006; PIATETSKY, 2007). Em particular, destaca-se o processo de tomada de decisão associado à avaliação de programas de eficiência energética implementados pelas Companhias Energéticas. Neste cenário, ressalta-se o processo de tomada de decisão associado à implementação ou não de programas de eficiência energética pelas Companhias Energéticas e à avaliação dos resultados alcançados por estes programas. Esta avaliação pode ser realizada através de um procedimento de identificação de modelos, com base no reconhecimento de padrões de consumo antes e após à implementação dos programas, a fim de estimar o ganho em eficiência efetivamente alcançado com a substituição de equipamentos. Embora este procedimento de identificação de modelos não seja o objetivo central deste trabalho, adicionalmente, no final do quinto capítulo (Análise dos dados e Discussões), discute-se as principais questões relacionadas a esta situação.

1.3 ESCOPO DA TESE

Este texto está estruturado em sete capítulos. O primeiro capítulo contextualiza o estudo e discute a justificativa. O segundo capítulo apresenta o embasamento teórico referente aos métodos de reconhecimento de padrões que nortearam a elaboração dos capítulos restantes, bem como, balizaram o desenvolvimento do novo método apresentado. Sem ter a pretensão de se fazer uma abordagem ampla dos fundamentos utilizados, e sem perder de vista o rigor científico, o terceiro capítulo faz uma revisão bibliográfica com foco no setor elétrico, as discussões estão concentradas nos modelos matemáticos e estatísticos adotados como base para o método proposto. O quarto capítulo apresenta a metodologia utilizada, destacando: hipóteses, objetivos e pressupostos que nortearam a pesquisa, bem como, o processo de amostragem e materiais e métodos para a coleta dos dados. O quinto capítulo descreve o novo método de reconhecimento de padrões em curvas de consumo do setor elétrico para o caso univariado com os principais resultados da análise dos dados e discussões pertinentes. O sexto capítulo apresenta a extensão do método para o caso multivariado também com os principais resultados da análise dos dados e discussões pertinentes. O último capítulo apresenta a conclusão sobre os principais aspectos verificados na aplicação do método proposto e trabalhos futuros são sugeridos.

2 EMBASAMENTO TEÓRICO DA ANÁLISE DE AGRUPAMENTOS

2.1 VISÃO GERAL

A análise de agrupamentos (*Cluster Analysis*) compreende a partição de uma amostra ou conjunto de dados (que aqui serão referenciados como **objetos**) em grupos (*clusters*) através de características extraídas dos próprios dados/objetos. Métodos de agrupamento constituem-se também em métodos de aprendizado não supervisionado na medida em que os objetos não são preliminarmente rotulados (JAIN et al., 1999; KAVITHA & PUNITHAVALLI, 2000; MITSA, 2012).

Jain et al. (1999) apresentam uma extensa revisão sobre o tema, e sob o ponto de vista do aprendizado não supervisionado de padrões (observações, itens de dados, ou objetos) em grupos, abordam o problema de agrupamento em muitos contextos, mencionando vários pesquisadores de muitas áreas do conhecimento, o que reflete o seu amplo apelo e utilidade como um dos passos fundamentais na análise exploratória de dados.

Dentro deste contexto, os autores apresentam uma visão geral de análise de agrupamento a partir de uma perspectiva de reconhecimento de padrões, sobretudo, utilizando métricas estatísticas de distâncias, com o objetivo de prover suporte ao processo de formação de grupos. Existem duas abordagens de agrupamentos, quais sejam, abordagens hierárquica e não-hierárquica. A abordagem hierárquica compreende a partição da amostra inicial em grupos que, por sua vez, dão origem a outros grupos e assim sucessivamente até o alcance da convergência. Desta forma, o número de grupos final é um resultado do método hierárquico. Métodos não hierárquicos podem ser utilizados após a aplicação de um método hierárquico e são baseados em problemas de otimização nos quais as variáveis de decisão são os centros (padrões) de cada grupo e as pertinências de cada objeto a cada um dos grupos. O número de grupos deve ser definido previamente nos métodos não hierárquicos. Trebuna & Halcinová (2013) apresentam uma revisão sobre a análise de agrupamentos destacando também as abordagens hierárquicas (métodos aglomerativos e divisivos) e não-hierárquicas (método *k-means* e *FCM*).

O objetivo de uma análise de agrupamento é agrupar objetos baseando-se nas seguintes premissas (HOPNNER et al, 2000; BEZDEK et al, 2005,):

- Homogeneidade – objetos pertencentes a um mesmo grupo devem ser similares tanto quanto possível;
- Heterogeneidade – Objetos pertencentes a grupos diferentes devem ser diferentes tanto quanto possível.

O emprego do termo "objeto" é intencional e permite que se destaque no contexto da análise de agrupamentos uma inerente diversidade de aplicações e cenários. Os objetos se classificam em **dados de objetos** e **dados relacionais** (BEZDEK et al, 2005). No primeiro caso, têm-se objetos numéricos e não numéricos (dados categóricos). Objetos representados por dados categóricos possuem atributos qualitativos sem que haja um ordenamento natural entre estes. O caráter qualitativo dos atributos sugere, não exclusivamente para este caso, o emprego da lógica *fuzzy* como estratégia de agrupamento (HOPNNER et al, 2000) devido à incerteza de informação presente neste contexto. Dados relacionais, por sua vez, referem-se a índices que quantificam o nível de correlação entre os objetos.

A análise de agrupamento é constituída das seguintes etapas, quais sejam, caracterização, extração de características, aplicação de um algoritmo de agrupamento e validação de grupos. Os principais aspectos enfatizados na etapa de caracterização são a transformação de objetos e sua visualização. A extração de características permite o reconhecimento de aspectos comuns e divergentes entre os objetos. Na etapa da aplicação de algoritmo de agrupamento de objetos, tem-se o reconhecimento ou obtenção dos grupos e dos respectivos centros (padrões) a partir das características extraídas (GAN et al, 2007).

De forma geral, uma amostra de *n* objetos numéricos pode ser representada genericamente pelo conjunto $X = \{x_1, x_2, \ldots, x_n\}$. Cada objeto $x_i \quad (i = 1, \ldots, n)$ é um vetor no espaço $\Re^p$ (*p* é a **dimensionalidade** do objeto) e, neste caso, tem-se o chamado agrupamento pontual (*point-prototype clustering*, x_i é um ponto em $\Re^p$). Por sua vez, $x_{ki} \quad (k = 1, \ldots, p)$ é o *k-ésimo* **atributo** associado ao objeto x_i. Por outro lado, se cada objeto $x_i \quad (i = 1, \ldots, n)$ está associado a múltiplos vetores ou simplesmente não pode ser referendado como um ponto no espaço $\Re^p$ tem-se então outro tipo de problema denominado de agrupamento não pontual (*non point-prototype clustering*).

2.2 CARACTERIZAÇÃO DOS OBJETOS

A caracterização dos objetos refere-se ao pré-processamento, padronização, transformação e visualização. No primeiro, as ações são dirigidas para a detecção de objetos atípicos. A padronização consiste na normalização de valores, ou seja, a conversão de valores originais dos objetos para valores de escalas com amplitudes padronizadas. A transformação consiste na tentativa de redução de dimensionalidade de cada objeto. A visualização compreende a verificação de tendência dos objetos para a formação de grupos ou de indícios da qualidade final dos agrupamentos.

2.2.1 Transformação ou redução de dimensionalidade

Destacam-se aqui as técnicas de análise de componentes principais e a análise fatorial cuja principal finalidade compreende a redução da dimensionalidade (número de atributos) associados aos objetos e a retenção do máximo possível da variabilidade presente na amostra original. Considerando que os atributos são genericamente variáveis que definem ou caracterizam um determinado objeto, as componentes principais representam novas variáveis não correlacionadas e ordenadas de modo que as primeiras retenham a maior parte da variabilidade presente na amostra original (GAN et al (2007). A análise fatorial, por sua vez, é um procedimento para a definição de novas variáveis que retém um grau de explicação aceitável para a variabilidade da amostra final.

- **Análise de componentes principais**

A Análise de Componentes Principais (*Principal Component Analysis - PCA*) é uma técnica descritiva que possibilita a redução da dimensionalidade dos objetos, hierarquizando a importância das combinações lineares obtidas entre as variáveis (HAIR et al., 2005). Assim, através desta técnica é possível eliminar as combinações lineares de baixo poder explicativo da variabilidade e reter somente aquelas com alto poder de explicação.

Algebricamente, para fazer a redução da dimensionalidade através da *PCA*, calcula-se os autovalores e os autovetores da matriz de correlação das variáveis, e projeta os dados ortogonalmente no subespaço definido pelos autovetores pertencentes aos maiores autovalores. Diferentemente da análise fatorial, a *PCA* não possui um suporte estatístico para testar a significância dos autovetores a serem selecionados (HAIR et al., 2005). A sua ênfase é dada na explicação das correlações entre as variáveis e os autovetores, encontrando funções

matemáticas que maximizam a explicação da variação existente nos dados. Fazendo uma explicação mais detalhada deste processo, de acordo com Reis (2001), considere inicialmente $X' = [\ X_1\ X_2\ \ldots\ X_p]$ um vetor de variáveis aleatórias com média $\underset{\sim}{\mu}$ e matriz de variância/covariância Σ. Pretende-se encontrar um novo conjunto de variáveis $Y_1, Y_2, \ldots Y_p$, não correlacionadas entre si e cujas variâncias decresçam da primeira para a última, isto é:

$$Var[Y_1] \geq Var[Y_2] \geq \ldots \geq Var[Y_p] \quad (1)$$

Cada nova variável Y_j pode então ser tomada como uma combinação linear de X:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p = \underset{\sim}{a}_j' \underset{\sim}{X} \quad (2)$$

sendo $\underset{\sim}{a}_j' = [a_{1j} a_{2j} \ldots a_{pj}]$ um vetor de constantes tal que

$$\underset{\sim}{a}_j' \underset{\sim}{a}_j = \sum_{i=1}^{p} a_{ij}^2 = 1 \quad (3)$$

e

$$\underset{\sim}{a}_j' \underset{\sim}{a}_r = 0 \quad (4)$$

Para $j \pm r, j, r = 1, 2, \ldots, p$ e a_j com $j = 1, \ldots, p$ são denominados de autovetores.

A primeira componente principal Y_1 é obtida escolhendo-se o vetor de constantes $\underset{\sim}{a}_1$ de modo Y_1 tenha a máxima variância possível. Ou seja, escolhe-se $\underset{\sim}{a}_1$ de maneira a maximizar a variância de Y_1 ($\underset{\sim}{a}_1' \underset{\sim}{X}$) atendendo às restrições estabelecidas.

A segunda componente é derivada de modo similar, ou seja, escolhendo $\underset{\sim}{a}_2$ tal que Y_2 tenha variância máxima e seja ortogonal à primeira componente Y_1. Seguindo o mesmo processo, encontram-se $Y_3, Y_4, \ldots, Y_p$, todas não correlacionadas entre si e com variância decrescente.

Seja A a matriz dos autovetores:

$$A = [a_1 a_2 \ldots a_p] \quad (5)$$

e $\underset{\sim}{Y}$ o vetor das componentes principais. Então:

$$\underset{\sim}{Y} = A' X \tag{6}$$

A matriz de variância /covariância de *Y* será:

$$Var[\underset{\sim}{Y}] = A' \sum A = \psi \tag{7}$$

sendo,

$$\psi = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \tag{8}$$

ou ainda

$$\sum = A' . \psi . A \tag{9}$$

no qual, A é uma matriz ortogonal. Os autovalores (λ_j) representam as variâncias das respectivas componentes principais. A soma destas variâncias é:

$$\sum_{j=1}^{p} Var[Y_j] = \sum_{j=1}^{p} \lambda_j = tr(\psi) \tag{10}$$

mas,

$$tr(\psi) = tr(A' . \sum . A) = tr(\sum . A . A') = tr(\sum) = \sum_{i=1}^{p} Var[X_i] \tag{11}$$

Ou seja, a soma das variâncias das variáveis originais é igual a somas das variâncias das componentes obtidas.

Vale destacar que, a redução da dimensionalidade do vetor *Y* pode se feita através da seleção das suas primeiras componentes principais, obtidas escolhendo-se os autovetores que tenham as maiores explicações da variância de *X.*

- **Análise fatorial**

De acordo com Hair et al. (2005), a análise fatorial usualmente é utilizada como análise confirmatória para a seleção das componentes principais que retêm um maior grau de explicação da variabilidade dos dados. Dessa forma, a *PCA* pode ser usada na análise fatorial para obter as componentes principais candidatas a receberem o *status* de fator ou

componentes principais mais significativas. O método de análise fatorial baseia-se na combinação linear entre as variáveis presentes na amostra original (X_i) e k fatores comuns (F):

$$X_i = l_{i1}F_1 + l_{i2}F_2 + ... + l_{ik}F_k + \mu_i + \varepsilon_i \quad (12)$$

l_{ik} são as cargas fatoriais ($i=1,...,k$), F_k são os fatores comuns, μ_i é o fator único e ε_i é o fator de erro do modelo.

As cargas fatoriais indicam a intensidade das relações entre os escores reduzidos das variáveis X_i e os fatores comuns. Dessa forma, um primeiro fator é escolhido para maximizar a soma dos quadrados das cargas fatoriais em relação a ele. Em seguida, obtém-se um segundo fator para que seja maximizada a soma de quadrados das cargas fatoriais, e assim por diante para os demais fatores. Para verificar o grau de explicação da variabilidade presente na amostra conferido pelos fatores, a comunalidade (h_i^2,) é definida da seguinte forma:

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + ... + l_{ij}^2 \quad (13)$$

Para estimar os escores fatoriais, uma rotação de fatores ortogonais (ou fatores não correlacionados) pelo método *Varimax* pode ser utilizada (HAIR et al., 2005). Após a rotação ortogonal dos fatores, determinam-se os escores fatoriais associados. Por definição, os escores fatoriais são calculados para cada fator em cada observação, com o objetivo de situá-las no espaço dos fatores comuns. Para isso, obtém-se primeiramente a matriz de escores fatoriais, resultado da multiplicação da matriz de cargas fatoriais pela inversa da matriz de correlação das observações.

Para operacionalizar a estimação dos parâmetros da análise fatorial, algumas suposições são necessárias:

i. $E[F_{mx1}] = 0$, o que implica que todos os fatores tem esperança igual a zero;

ii. $Var[F_{mx1}] = I_{mxm} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$

Ou seja, todos os fatores F_j são não correlacionados e têm variâncias iguais a 1;

iii. $E[\varepsilon_{px1}] = 0$, o que implica que $E[\varepsilon_j] = 0$, $j = 1, 2, ..., p,$ ou seja, todos os erros tem média iguais a zero;

iv. $Var[\varepsilon_{pxp}] = \psi_{pxp} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \lambda_p \end{bmatrix}$

ou seja, $Var[\varepsilon_{pxp}] = \lambda_j$ e $Cov(\varepsilon_i,\varepsilon_j) = 0$, $\forall i \neq j$, o que significa dizer que os erros são não correlacionados entre si e não possuem necessariamente a mesma variância;

v. Os vetores ε_{px1} e F_{mx1} são independentes. Portanto, $Cov(\varepsilon_{px1},F_{mx1}) = 0$.

Por conseguinte, quando o modelo ortogonal é assumido, tem-se a matriz P_{pxp}:

$$P_{pxp} = LL' + \psi \tag{14}$$

Isso vem do fato de que:

$$P_{pxp} = Var(Z) = Var(LF + \varepsilon) = Var(LF) + Var(\varepsilon) = LIL' + \psi = LL' + \psi \tag{15}$$

onde I é a matriz identidade de dimensão pxp.

Sejam

$$Z_i = \left[\left(\frac{X_i - \mu_i}{\sigma_i}\right)\right] \tag{16}$$

onde Z_i são os escores reduzidos das variáveis originais, onde X_{px1} é um vetor aleatório com μ_i e σ_i representando, respectivamente, a média e o desvio padrão da variável X_i, $i = 1, 2, ..., p$.

$$\begin{array}{rcccccccccc} Z_1 & = & l_{11}F_1 & + & l_{12}F_2 & + & \dots & + & l_{1m}F_m & + & \varepsilon_1 \\ Z_2 & = & l_{11}F_1 & + & l_{12}F_2 & + & \dots & + & l_{1m}F_m & + & \varepsilon_1 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ Z_2 & = & l_{11}F_1 & + & l_{12}F_2 & + & \dots & + & l_{1m}F_m & + & \varepsilon_1 \end{array} \tag{17}$$

$$\varepsilon_{px1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \tag{18}$$

$$L_{pxm} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{bmatrix} \tag{19}$$

Um dos principais métodos de estimação das matrizes L_{pxm} e ψ_{pxp} é o método das componentes principais:

$$\hat{L}_{pxm} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \quad \sqrt{\hat{\lambda}_2}\hat{e}_2 \quad \cdots \quad \sqrt{\hat{\lambda}_m}\hat{e}_m\right] \tag{20}$$

Os $\hat{\lambda}_i$, $i = 1,2, ..., m$ autovalores que retêm maior proporção da variância total. $\hat{e}_i$ é o autovetor normalizado.

$$\hat{\psi}_{pxp} = diag(R_{pxp} - \hat{L}_{pxm}\hat{L}'_{mxp}) \tag{21}$$

diag(.) denota a matriz diagonal.

Ainda pode-se dizer também que análise fatorial consiste na aplicação do teorema da decomposição espectral à matriz R_{pxp} (MINGOTI, 2005). Por este teorema, a matriz de correlação pode ser decomposta como uma soma de p matrizes, cada uma relacionada com um autovalor da matriz R_{pxp}. Para um dado m:

$$R_{pxp} = \sum_{i=1}^{p}\hat{\lambda}_i\hat{e}_i\hat{e}_i' = \sum_{i=1}^{m}\hat{\lambda}_i\hat{e}_i\hat{e}_i' + \sum_{i=m+1}^{p}\hat{\lambda}_i\hat{e}_i\hat{e}_i' \tag{22}$$

Assim, uma aproximação para a matriz LL' será dada por:

$$\hat{L}\hat{L} = \sum_{i=1}^{m}\hat{\lambda}_i\hat{e}_i\hat{e}_i' \tag{23}$$

Nas situações em que o método das componentes principais é usado para estimação das matrizes L_{pxm} e ψ_{pxp}, a proporção de variância explicada por um determinado fator se reduz a $\dfrac{\hat{\lambda}_i}{\sum_{i=1}^{p}\hat{\lambda}_i}$. Este valor representa o quanto cada fator consegue captar da variabilidade dos escores reduzidos Z_i ($i = 1,2, ...,p$).

2.2.2 Visualização dos dados

No que tange a visualização de dados, o escalonamento multidimensional é uma técnica matemática de interdependência comumente utilizada para mapear distâncias entre pontos em uma representação gráfica espacial. Esta técnica possibilita uma avaliação da tendência dos dados para formação de grupos e indícios da qualidade dos grupos a serem reconhecidos através de um determinado algoritmo de agrupamento. O escalonamento

multidimensional é uma técnica que engloba um grupo bastante amplo de procedimentos de análise multivariada voltados ao tratamento de dados dispostos em tabelas de contingências ou matrizes de dissimilaridades. Sua utilização é apropriada para representar graficamente *j* objetos em um espaço de dimensão menor do que o original (portanto é também uma técnica de redução de dimensionalidade), levando-se em consideração a distância ou similaridade entre os objetos.

Por exemplo, considere Δ, de tamanho *n* x *n*, uma matriz quadrada de dissimilaridades. Cada elemento δ_{ij} de Δ representa a distância entre os objetos *i* e *j*. Para *n* = 4, temos que:

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix} \quad (24)$$

A solução fornecida através da análise do escalonamento multidimensional é uma matriz de objetos *n* x *m* (*n* é o número de objetos e *m* é o número de atributos) (FÁVERO, 2009). Na presente situação, uma solução com duas dimensões para os quatros objetos anteriores é a seguinte:

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} \quad (25)$$

Cada ponto (x_{i1}, x_{i2}) representa uma coordenada do objeto *i* nos eixos *X* e *Y* do espaço bidimensional. O cálculo das distâncias correspondentes a todos os objetos da matriz *X* proporciona uma nova matriz *D*:

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{31} & d_{32} & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{bmatrix} \quad (26)$$

Se a dimensão da matriz de *X* for a mesma de Δ, as matrizes *D* e Δ são equivalentes quanto a variabilidade dos dados relacionais. Por outro lado, pode-se buscar reduzir, o máximo possível, a dimensão da matriz de *X* (normalmente reduz-se para a dimensão igual a 2), de tal maneira que a matriz *D* tenha um grau de variabilidade muito próximo do grau de variabilidade da matriz Δ.

2.3 MEDIÇÃO RELACIONAL

A medição relacional entre os objetos é realizada através de métricas geométricas e/ou métricas estatísticas de acordo com critério de proximidade estabelecido. Nas métricas geométricas as medidas de ângulo e distância são usadas. As métricas estatísticas são baseadas em distribuições de probabilidade para mensurar o grau de relação entre os objetos.

De acordo com Mitsa (2012), a métrica de proximidade estabelece o critério de escolha de objetos para compor um mesmo grupo. Tem-se duas categorias, quais sejam, medidas de similaridade e de dissimilaridade. Na medida de similaridade, o aumento dos seus valores implica que mais parecidos serão os objetos. Por outro lado, na medida de dissimilaridade quanto maior forem os seus valores menos similares serão os objetos. O coeficiente de correlação e a distância euclidiana são, respectivamente, exemplos de medida de similaridade e dissimilaridade.

2.3.1 Medidas geométricas

A medição da distância e do ângulo entre dois objetos são exemplos de medidas geométricas comumente utilizadas na mensuração da similaridade e da dissimilaridade. O coeficiente de correlação pode ser considerado também como um tipo de medida angular. Considerando cada objeto com dimensionalidade p (vetor no espaço $\Re^p$), sendo x_i e x_k dois objetos, tem-se:

- Distância Euclidiana:

$$d_{ij} = \|x_i - x_j\| = \sqrt{(x_i - x_j)^T . (x_i - x_j)} \rightarrow \Re^p \rightarrow \Re^+ \quad (27)$$

- Distância de Mahalanobis:

$$d_{ij} = \|x_i - x_j\|_M = \sqrt{(x_i - x_j)^T . M^{-1} (x_i - x_j)} \rightarrow \Re^p \rightarrow \Re^+ \quad (28)$$

onde M é a estimativa amostral da matriz variância-covariância dentro do agrupamento de objetos com p rótulos.

- Distância de Minkowski:

A distância Minkowski é a generalização da distância Euclidiana entre dois objetos.

$$d_{ij} = \left(\sum_{p=1}^{m} \left(x_{ip} - x_{jp} \right)^{q} \right)^{1/q} \quad (29)$$

2.3.2 Medidas estatísticas

Segundo Khalil (2011), medições estatísticas são aquelas obtidas com base em modelos de probabilidades. Particularmente para objetos representados por protótipos pontuais, as medidas estatísticas baseiam-se nos modelos de qui-quadrado (χ^2) (teste de aderência entre distribuições) e *t* de *Student* (correlações entre variáveis). No caso dos objetos representados por protótipos não pontuais são considerados os modelos qui-quadrado (χ^2) (correlação canônica) e distribuição *F* (*Multivariate Analysis of Variance* - *MANOVA*). Adicionalmente aos modelos de probabilidade, regras de decisões são consideradas através de testes de hipóteses que avaliam a consistência das medidas estatísticas como critérios para agrupar objetos.

- **Estatística *χ2* - teste de aderência entre distribuições**

O modelo de probabilidade qui-quadrado pode ser aplicado na versão univariável (SEPPÃLÃ, 1995; JANES, 2001) e na versão multivariável (KHALIL, 2011). Seppãlã (1995) chamou atenção que as empresas de energia elétrica utilizam de dados de cargas de clientes para a gestão e planejamento das operações do sistema elétrico, por isso, é necessário mais informação dos perfis de carga do cliente. Nesta situação, seu estudo apresentou um novo método para estimar intervalos de confiança da carga do cliente, discutindo os principais pressupostos sustentadores da aplicação dos modelos estatísticos inferenciais. O estudo de caso baseou-se nos dados de medição da carga horária de uma empresa distribuidora de energia elétrica da Finlândia.

Janes (2001) discutiu o teste do qui-quadrado para relação entre variáveis categóricas resultantes de um processo de discretização de séries temporais continua. O estudo de caso compreendeu na análise de dados para a biblioteca e ciências da informação.

Khalil et al. (2011) desenvolveram três modelos para a estimação da qualidade da água de sítios. O primeiro modelo foi baseado na rede neural artificial (*Artificial Neural Network – ANN*), o segundo modelo foi baseado no conjunto de redes neurais artificais (*Ensemble Artificial Neural Network – EANN,* e o terceiro modelo foi baseado na análise de correlação

canônica (*Canonical Correlation Analysis – CCA*) e no *EANN*. Os modelos *ANN* e *EANN* foram desenvolvidos para o estabelecimento da relação funcional entre valores médios da qualidade da água e atributos da bacia d'agua. Nos modelos baseado sem *EANN* e *CCA* foram usados para as formas canônicas atributos espaciais utilizando dados medidos de sítios. Então, uma *EANN* foi aplicada para identificar o relacionamento funcional entre valores médios da qualidade da água e os atributos na *CCA* espacial. A estatística qui-quadrado (χ^2) foi utilizada para avaliar as significâncias das correlações canônicas. Quatro variáveis da qualidade da água foram selecionadas como resultado deste modelo. A seleção das variáveis foi feita baseada na análise das componentes principais. As variáveis de qualidade da água que mostraram altas cargas fatoriais nos quatro primeiros componentes foram selecionadas. Os três modelos foram aplicados na mesma amostra. Uma medida de validação (Jackkinife) foi usada para avaliar o desempenho dos três modelos. Os resultados mostraram que o modelo *EANN* forneceu uma generalização melhor do que o modelo *ANN*. Contudo, o modelo *CCA* baseado no modelo *EANN* teve um melhor desempenho em relação aos outros dois modelos quanto a acurácia da predição.

A regra de decisão para agrupar objetos utiliza a estatística qui-quadrado (χ^2) para realizar os testes estatísticos de aceitação ou rejeição da hipótese de nulidade estabelecida. Se X é uma variável aleatória com modelo Gama de probabilidade a estatística χ^2 é dita ter um modelo qui-quadrado com ν graus de liberdade:

$$\chi^2 = 2.X \tag{30}$$

A variável ou estatística χ^2 (BUSSAB & MORETIN, 2006) é uma medida que pode ser utilizada para comparar uma distribuição de frequência observada com uma distribuição de frequência esperada (ou de referência). O valor de uma estimativa da estatística χ^2 está associado a uma probabilidade de aceitação da hipótese nula (H_0) ou rejeição da hipótese alternativa H_1. O teste compreende duas hipóteses, quais sejam:

H_0: as distribuições de frequências observadas não diferem das distribuições de frequências esperadas;

H_1: as distribuições de frequências observadas diferem das distribuições de frequências esperadas.

A hipótese H_0 será aceita se o valor da estatística χ_c^2 (Eq. (31)) estiver associado a uma probabilidade de aceitação satisfatória. Caso contrário, a hipótese H_1 será aceita.

$$\chi_c^2 = \sum_{j=1}^{n} \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j}; \tag{31}$$

n é o número de classes da distribuição esperada, p é o número de parâmetros estimados, f_j é a frequência observada na j-ésima classe; $\hat{f}_j$ é frequência esperada na j-ésima classe. $v = n - p - 1$ é o número de graus de liberdade.

- **Estatística *t* de *Student* para correlação entre variáveis**

O teste de significância do coeficiente de correlação amostral pode ser feito com base na distribuição t de *Student* (BUSSAB & MORETIN, 2006), a qual fornece os valores críticos para regra de decisão de aceitação ou rejeição da hipótese de nulidade estabelecida. Se X é uma variável aleatória com distribuição Beta a estatística t é dita ter uma distribuição de *Student* com v graus de liberdade:

$$t = \sqrt{vX(1-X)} \tag{32}$$

De outra forma, tomando-se Z, uma variável aleatória com distribuição normal possuindo média 0 e variância 1, e χ^2 uma variável aleatória com distribuição qui-quadrado possuindo v graus de liberdade (Z e χ^2 sejam independentes):

$$t = \frac{Z}{\sqrt{\chi^2 / v}} \tag{33}$$

Uma aplicação do modelo de distribuição t de *Student* é o teste de significância do coeficiente de correlação (r) (BUSSAB & MORETIN, 2006):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}; \tag{34}$$

A estatística t é submetida a uma regra de decisão para verificar se o valor encontrado de r é estatisticamente válido, ou seja, a estatística t é associada a uma probabilidade de aceitação ou rejeição, respectivamente, das hipóteses nula (H_0) e alternativa (H_1):

H_0: As variáveis não são correlacionadas;

H_1: As variáveis são correlacionadas.

A hipótese H_0 é aceita quando *t* estiver associado a uma probabilidade de aceitação satisfatória. Caso contrário, a hipótese H_1 será aceita.

- **Estatística *t* de *Student* para o teste de diferenças de médias**

De acordo com O'Gorman (1997), o teste de significância da diferença entre médias amostrais pode ser feito também com base na distribuição *t* de *Student*. O'Gorman (1997) propôs um teste de duas amostras com maior poder de teste para várias alternativas de testes estatísticos. O teste apresentado é um teste adaptativo que usa as estatísticas de ordem nas duas amostras para determinar uma função de pontuação adequada para um teste de classificação. Este teste é mostrado, através de estudos de Monte Carlo, para manter o seu nível de significância para muitas distribuições e configurações de tamanho de amostra. O teste proposto teve o maior poder ou estava dentro de 5% do maior poder entre os exames considerados por 90,4% das simulações. Dentre os testes estatísticos utilizados como referência, destaca-se o teste *t* de *Student*. Para tamanhos iguais e variâncias iguais a estatística *t* é calculada da seguinte forma:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S_{y_1 y_2} \cdot \sqrt{\frac{2}{n}}} \tag{35}$$

no qual $\bar{y}$ é a média amostral, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ é o desvio padrão amostral e $S_{y_1 y_2} = \sqrt{\frac{s_{y1}^2 + s_{y2}^2}{2}}$ é a variância combinada. O número de graus de liberdade é *2n – 2.*

Esta estatística *t* também é associada a uma probabilidade de aceitação ou rejeição, respectivamente, das hipóteses nula (H_0) e alternativa (H_1):

H_0: A diferença entre as médias é estatisticamente desprezível (hipótese nula).

H_1: A diferença entre as médias não é estatisticamente desprezível (hipótese alternativa).

A hipótese H_0 é aceita quando *t* estiver associado a uma probabilidade de aceitação satisfatória. Caso contrário, a hipótese H_1 será aceita.

- **Estatística *χ2* – teste do coeficiente de correlação canônica**

O conceito de correlação canônica foi primeiramente proposto por Hotelling (MINGOTI, 2005), sendo uma técnica estatística com o objetivo de quantificar o grau de associação existente entre um conjunto de variáveis independentes com um conjunto de variáveis dependentes.

Segundo Hair et al. (2005), não há necessidade da comprovação da normalidade para aplicar a correlação canônica. Johnson & Wichern (2007) e Fávero (2009) apresentam a correlação canônica da seguinte forma:

$$W_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{iq}X_q \tag{36}$$

$$V_i = b_{i1}Y_1 + b_{i2}Y_2 + \cdots + b_{ip}Y_p \tag{37}$$

$$(V_1, \cdots, V_q) = f(W_1, \cdots, W_p) \tag{38}$$

V_i(*i=1*,.., *p*) são variáveis dependentes e W_j(*j=1*,...,*q*) são as variáveis independentes.

$$W = a'X \tag{39}$$

$$V = b'Y \tag{40}$$

Então

$$Var(W) = a'Cov(X)a = a'\Sigma_{11}a \tag{41}$$

$$Var(V) = b'Cov(Y)b = b'\Sigma_{22}b \tag{42}$$

$$Cov(W,V) = a'Cov(X,Y)b = a'\Sigma_{12}b \tag{43}$$

A correlação canônica ($Corr(W,V)$) pode ser expressa por:

$$Corr(W,V) = \frac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}a}\sqrt{b'\Sigma_{22}b}} \tag{44}$$

Σ são sub-matrizes de covariâncias e ξ pares de variáveis canônicas (ξ=min [(*p*,*q*)-*1*]) irão expressar a variância total dos dois conjuntos de variáveis. Dentre os pares de variáveis canônicas, estima-se $a_{11}, a_{12}, \cdots, a_{1q}$ e $b_{11}, b_{12}, \ldots, b_{1p}$, de modo que a correlação canônica $(Corr(W_1, V_1))$ seja máxima.

A Estatística $\chi 2$ também pode ser aplicada na análise da correlação canônica (KHALIL et al., 2011) para testar a sua significância:

$$\chi^2 = -\left[N - 1 - 0{,}5(p+q+1)\right] \cdot \ln\left[\prod_{i=1}^{q}\left(1 - Corr(W_1, V_1)^2\right)\right] \tag{45}$$

sendo N é o tamanho da amostra, P é o número de variáveis dependentes, Q é o número de variáveis independentes. A distribuição de qui-quadrado terá $p\ x\ q$ graus de liberdade.

A estatística χ^2 é associada a uma probabilidade de aceitação ou rejeição, respectivamente, das hipóteses nula (H_0) e alternativa (H_1):

H_0: A correlação entre as variáveis canônicas é estatisticamente desprezível (hipótese nula);

H_1: A correlação entre as variáveis canônicas é não estatisticamente desprezível (hipótese alternativa).

A hipótese H_0 é aceita quando t estiver associado a uma probabilidade de aceitação satisfatória. Caso contrário, a hipótese H_1 será aceita.

- **Estatística *F* - análise multivariada da variância (*MANOVA*)**

Uma aplicação de interesse particular da estatística F é a *MANOVA* (ZIMBA et al., 2003). De acordo com Zimba et al. (2003), a *MANOVA* é uma técnica de dependência que compara as diferenças de médias para duas ou mais variáveis dependentes quantitativas, com base em um conjunto de variáveis independentes qualitativas. É uma generalização da análise univariada de variância (*Analysis of Variance* - *ANOVA*). Zimba et al. (2003) analisaram as relações temporais de variáveis pertencentes a área de biologia. Este estudo analisou as relações de idade de longo prazo com nutrientes, zooplâncton e fitoplâncton e incidência de ocorrência de *off-flavor*. Os níveis de nutrientes (ferro, silício, total de fósforo e total de nitrogênio), composição da comunidade fitoplanctônica, compostos *off-flavor* e composição da comunidade de zooplâncton foram determinados para 10 réplicas. Os dados foram analisados por meio da análise função canônica discriminante, análise de variância multivariada (*MANOVA*), e abordagens de regressão múltipla.

No teste de igualdade de k vetores de médias as hipóteses a testar são:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k \Rightarrow$ Todos os grupos tem vetores de médias iguais;

H_1:* $\mu_i \neq \mu_j$ com $i \neq j \Rightarrow$Pelo menos dois grupos tem vetores de médias diferentes.

Na análise de variância simples, a estatística utilizada para aceitação ou rejeição da hipótese nula consiste no quociente entre a média de quadrados entre os grupos (*MQEG*) e a média de quadrados dentro dos grupos (*MQDG*) (REIS, 2001):

$$T = \frac{QEG/(K-1)}{QDG/(N-K)} = \frac{MQEG}{MQDG} \tag{46}$$

Esta razão resulta numa distribuição *F* com (*K–1, N – K*) graus de liberdade. Sendo *K* é o número de variáveis e *N* é o número de casos. Generalizando para análise multivariada, considera-se λ de Wilks (REIS, 2001):

$$\lambda = \frac{|W|}{|T|} = \frac{|W|}{|B+W|} \tag{47}$$

O determinante de *W* mede a dispersão interna dos grupos e o determinante de *T* mede a dispersão total. Assim, de acordo com Reis (2001), quanto menor o valor da estatística λ mais separados estão os grupos. Considera-se que as matrizes de variância/covariância, embora desconhecidas, são iguais para os *k* grupos:

$$\sum\nolimits_1 = \sum\nolimits_2 = ... = \sum\nolimits_k \tag{48}$$

Reis (2001) afirma que, sob certas condições, o teste do quociente de verossimilhanças λ pode ser reduzido à seguinte expressão:

$$\lambda = \left[\frac{|W|}{|T|}\right] \tag{49}$$

É possível obter a estatística λ da seguinte forma (REIS, 2001):

$$\lambda = \left|T^{-1}.W\right| \tag{50}$$

Mas

$$T = B + W \tag{51}$$

Logo:

$$\lambda = \left|(B+W)^{-1}.W\right| = \left|\left(B^{-1}+W^{-1}\right)W\right| = \left|\left(B^{-1}.W + W^{-1}.W\right)\right| = \left|\left(B^{-1}.W + I\right)\right| = \left|\left(W^{-1}B + I\right)\right|^{-1} \tag{52}$$

Se $n \geq p + k$, então:

$$\lambda = \frac{|W|}{|B+W|} = \left|W^{-1}.B + I\right|^{-1} \cap \lambda_{(p,n-k,k-1)} \tag{53}$$

onde p é o número de variáveis, n é o número de indivíduos e k é o número de grupos. O Sinal $\cap$ indica que λ segue a distribuição lambda de Wilks com graus de liberdade (p), ($n - k$) e ($k - 1$).

À medida que o valor de λ diminui ele avança para a rejeição de H_0. Outra forma de encontrar o λ é a seguinte (REIS, 2001):

$$\lambda = \prod_{j=1}^{s}\left(1+\lambda_j\right)^{-1} \tag{54}$$

em que os λ_j são os valores próprios de $W^{-1}.B$, e s o número de valores próprios não nulos dessa matriz.

Possíveis aproximações para a distribuição de λ de Wilks são (REIS, 2001):

$$\frac{(n-k)-p+1}{p}.\frac{1-\lambda}{\lambda} \overset{o}{\cap} F_{(p,n-k-p+1)} \tag{55}$$

$$-\left[(n-1)-\frac{1}{2}(p+k)\right]\ln\lambda \overset{o}{\cap} \chi^2_{(px(k-1))} \tag{56}$$

À medida que o valor de λ diminui ele avança para a rejeição de H_0. Esta conclusão pode ser feita porque λ está inversamente relacionado a estatística F, ou seja, quanto menor o valor de λ, maior é o valor de F e menor é a probabilidade de aceitação de H_0.

2.4 ALGORITMOS DE AGRUPAMENTOS

O avanço dos recursos computacionais fomentou a aplicação das técnicas de algoritmos de agrupamentos em diversas áreas do conhecimento (WITTEN 2005; JOHNSON, 2007), sobretudo, na mineração de dados (HAN et al., 2000) que consiste em obter conhecimentos que podem apoiar processos decisórios de modo a melhorar os sistemas de produção e tecnologia de gestão. Piatetsky (2007) examinou os principais avanços no campo de conhecimento da mineração de dados ocorridos nos últimos dez anos. A seguir é apresentada uma discussão sobre as principais abordagens dos algoritmos de agrupamentos com o intuito de contribuir para uma melhor compreensão do método proposto.

2.4.1 Abordagens dos algoritmos de agrupamentos

Os algoritmos de agrupamento de objetos efetuam o reconhecimento dos grupos e dos padrões utilizando duas abordagens de agrupamentos: a hierárquica e a não hierárquica. Vale reiterar que a abordagem de agrupamento hierárquica realiza a partição dos objetos de forma gradativa e sequencial, enquanto que na abordagem não-hierárquica, há uma predefinição do número de grupos a serem obtidos (HAIR et al, 2005), e que em geral, o método hierárquico é usado para estimar o número de grupos ótimos (validado por meio de um índice de qualidade) e, depois com esta informação, aplica-se o método não-hierárquico para obter os grupos finais. O nível de coesão/homogeneidade dentro dos grupos e de heterogeneidade entre eles é quantificado através de índices específicos de qualidade do agrupamento. Mais adiante (seção 2.4.2) será apresentada uma discussão mais detalhada sobre caracterização, medição relacional e aplicação de algoritmos de agrupamento em séries temporais.

- **Algoritmos hieráquicos**

Os algoritmos hierárquicos podem ser divididos em dois tipos (JAIN et al., 1997; TREBUNA & HALCINOVÁ, 2013): aglomerativos e divisivos. Ambos são realizados em etapas onde novos agrupamentos são formados utilizando-se um critério de limite de distância entre os grupos. No caso aglomerativo cada objeto representa inicialmente um grupo e, a partir daí, novos agrupamentos são formados por similaridade. Nas etapas posteriores os grupos vão se unindo por similaridade de tal forma que, na etapa final, os objetos constituam um único grupo. Ao contrário do aglomerativo, no divisivo todos os objetos começam em um só grupo sendo separados, primeiramente, os grupos mais dissimilares, até que cada grupo se torna um grupo isolado (com um objeto). Os algoritmos hierárquicos possibilitam obter os possíveis agrupamentos que podem ser formados em cada etapa do procedimento de aglomeração dos objetos ou divisão dos grupos. A qualidade dos agrupamentos formados por estes algoritmos dependem da maneira como as distâncias são calculadas entre os objetos. Segundo Johnson e Wichern (2007), na formação dos agrupamentos aglomerativos, os métodos para a formação dos agrupamentos mais freqüentes são: menor distância, maior distância, distância média, centróide e Ward.

O método da menor distância se baseia na formação de grupos separados pela menor distância entre os mesmos. Neste método, o primeiro grupo da etapa subsequente é formado com base na menor distância entre dois grupos candidatos. Dados dois grupos,

respectivamente, compostos por dois objetos (*i* e *j*) e um objeto *k*, a distância entre eles é representada pela distância mínima entre quaisquer objetos pertencentes aos respectivos grupos:

$$d_{(ij)k} = \min\{d_{ik}, d_{jk}\} \quad (57)$$

O método da maior distância baseia-se na distância máxima, ao contrário do método da menor distância. Neste método, a distância entre dois grupos é definida como a distância máxima entre todos os pares de possibilidades de observações nos dois grupos. O método busca obter agrupamento de objetos cujas distâncias entre os mais afastados seja a menor. Dados dois grupos, respectivamente, formados por dois objetos (*i* e *j*) e um objeto *k*, a distância entre eles é representada pela distância máxima entre quaisquer objetos pertencentes aos respectivos grupos:

$$d_{(ij)k} = \max\{d_{ik}, d_{jk}\} \quad (58)$$

O método de distância média trata a distância entre dois grupos como sendo a distância média entre todos os pares de indivíduos dos dois grupos, buscando agrupar os agregados cuja distância média é a menor. Como esta técnica se utiliza do valor médio, ao contrário dos métodos da menor distância e da maior distância, há a vantagem de não se precisar de valores extremos e de se utilizarem todos os elementos do grupo, ao invés de um único par de extremos. Portanto, dados dois grupos, respectivamente, formados por dois objetos (*i* e *j*) e um objeto *k*, a distância média entre eles é dada por:

$$d_{(ij)k} = média\{d_{ik}, d_{jk}\} \quad (59)$$

O método centróide, por sua vez, baseia-se na distância entre os centróides dos grupos candidatos a formarem um novo grupo, priorizando a menor distância entre eles. Este método identifica os dois grupos separados pela menor distância entre os objetos mais próximos e os coloca no mesmo agrupamento. Segundo Hair et al. (2005), centróide são valores médios dos respectivos atributos de todos os objetos de um grupo. Neste método, toda vez que um grupo recebe um novo objeto, um novo centróide é computado.

O método de Ward busca a atingir sempre o menor erro interno entre os objetos que compõe cada grupo e o centróide. Isto equivale a buscar a variância mínima entre os objetos de cada grupo. Trata-se de um método que tende a proporcionar grupos com aproximadamente o mesmo número de observações (HAIR et al., 2005). Segundo Reis (2001), o método Ward pode ser resumido nas seguintes etapas: primeiramente são calculadas

as médias das variáveis para cada grupo; em seguida é calculado o quadrado da distância euclidiana entre estas médias e os valores das variáveis para cada indivíduo; somam-se as distâncias para todos os indivíduos; por último, minimiza-se a variância dentro dos grupos.

- **Algoritmos não hieráquicos**

A abordagem não hierárquica é um processo dinâmico e interativo de formação de grupos. Uma vez especificado o número de grupos, o algoritmo tem como objetivo identificar ou reconhecer a melhor distribuição dos objetos conforme a premissas de homogeneidade intra e heterogeneidade inter grupos. Os procedimentos não hierárquicos compreendem métodos que têm como objetivo a obtenção de uma partição de n elementos em c($c \geq 2$) grupos (Mingoti, 2005) gerando também, como resultado, uma matriz de partição $U(c \times n)$ tal que u_{ik} refere-se à pertinência do objeto k ($k=1,...n$) ao grupo i ($j=1,...,c$). Cada coluna da matriz de partição fornece os graus de aderência ou pertinência de um dado objeto a todos os grupos reconhecidos. Os métodos não hierárquicos não requerem o cálculo e armazenamento de uma nova matriz de distância a cada iteração o que reduz o tempo computacional e possibilita a sua aplicação em grandes bases de dados.

Dentre os métodos de partição não hierárquica, os métodos baseados nos modelos *c-means* (tendência central da média aritmética) são os mais utilizados e validados quanto à eficiência e aplicabilidade em problemas de agrupamento de protótipo pontual (BEZDEK et al., 2005, HOPNNER, 2000).

Os métodos baseados em modelos *c-means* compreendem o seguinte problema de otimização:

Dada uma amostra de n objetos $X=\{x_1,\ldots,x_n\}$, de dimensionalidade p, e c grupos:

$$\min_{U,V}\left\{ J_m(U,V,W) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \left\| x_k - v_i \right\|_A^2 + \sum_{i=1}^{c} w_i \sum_{k=1}^{n} (1-u_{ik})^m \right\} \tag{60}$$

Sujeito a:

$u_{ii} \in [0,1]$, $\sum_{k=1}^{n} u_{ik} > 0 \ \forall_i \rightarrow$ partição possibilística;

$u_{ii} \in [0,1],\ \sum_{k=1}^{n} u_{ik} > 0\ \forall_i$ e $\sum_{i=1}^{c} u_{ik} = 1\ \forall_k \rightarrow$ partição probabilística

$u_{ii} \in \{0,1\},\ \sum_{k=1}^{n} u_{ik} > 0\ \forall_i$ e $\sum_{i=1}^{c} u_{ik} = 1\ \forall_k \rightarrow$ partição rígida.

Sendo $V = \{v_1,\ldots,v_c\}$, $v_i \in \Re^p$ é o vetor com os centros (protótipos ou padrões) de cada grupo, $W = \{w_1,\ldots,w_c\}$, $w_i \in \Re^+$ é uma penalidade associada ao *i-ésimo* grupo e m $(m \in \Re \quad \text{e} \quad m \geq 1)$ é denominado de *grau de fuzzyficação* ou "fuzzyficador".

As variáveis de decisão deste problema de otimização com restrições são os centros dos clusters $V = \{v_1,\ldots,v_c\}$ (padrões propriamente ditos) e os graus de pertinência $(u_{ik} | i = 1,\ldots,c\ ;\ k = 1,\ldots,n)$. As penalidades associadas a cada grupo (w_i, i=1,..., c) são parâmetros não nulos apenas no caso da partição possibilística. A participação possibilística se diferencia da probabilística na medida em que nesta última é imposto ao problema de otimização a restrição severa de que a soma dos graus de pertinência de um mesmo objeto a todos os grupos deve ser igual à unidade. Sendo assim, a restrição imposta na função objetivo através dos parâmetros de penalidade serve para privilegiar a obtenção de graus de pertinência elevados na partição possibilística, uma vez que a partição probabilística não se aplica neste caso.

Quanto maior o valor de m maior será o nível de *fuzzyificação* (ou incerteza) associado ao problema de agrupamento. Nível elevado de incerteza no agrupamento implica em dificuldade (ou falta de certeza) na definição dos grupos. Ou seja, se $m \rightarrow 1$ o agrupamento tenderá à participação rígida com graus de pertinência próximos de 0 ou 1 simplesmente (BEZDEK et al., 2005, HOPPNER, 2000). Valores elevados de m são sugeridos quando não há previamente uma clara distinção entre os grupos. Em geral, m=2 é o valor adotado nas aplicações envolvendo ambas as partições possibilística e probabilística (HOPNNER, 2000).

O método *Fuzzy-C-Means* (*FCM*) é um algoritmo apropriado para resolver problemas de agrupamento pontual, inclusive em aplicações no setor elétrico (ZALEWSKI, 2006; NIZAR et al., 2006). O *FCM* consiste em uma partição probabilística envolvendo o seguinte problema de otimização:

$$\min_{U,V}\left\{J_m(U,V,W)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\left\|x_k-v_i\right\|_A^2\right\}\ ,\ m>1 \tag{61}$$

Sujeito a:

$$u_{ii}\in[0,1],\ \sum_{k=1}^{n}u_{ik}>0\ \forall_i$$

$$\sum_{i=1}^{c}u_{ik}=1\ \forall_k .$$

A métrica de similaridade comumente utilizada compreende a distância Euclidiana (DUNN, 1974 ; BEZDEK et al., 2005).

A aplicação das condições de otimalidade (condições necessárias de primeira ordem) para o problema de otimização proposto gera a seguinte solução analítica para o ponto ótimo (geralmente mínimo local):

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_i-v_j\right\|}{\left\|x_i-v_k\right\|}\right)^{\frac{2}{m-1}}} \tag{62}$$

$$v_j=\frac{\sum_{i=1}^{N}u_{ij}^{m}.x_i}{\sum_{i=1}^{N}u_{ij}^{m}} \tag{63}$$

Com base nas equações 62 e 63 o algoritmo clássico *FCM* consiste no seguinte procedimento iterativo (LIAO, 2005):

i. Selecione $c(2\le c\le n)$, $m(1\le m\le\infty)$, e ε (um pequeno valor para finalizar o procedimento iterativo). Coloque l=0 e inicia a matriz de pertinência, $U^{(l)}$.

ii. Calcule o centro do cluster, $v_i^{(l)}$ pela Eq. (63).

iii. Atualize a matriz de pertinência $U^{(l+1)}$ pela Eq. (62) se $x_k \neq v_i^{(l)}$ caso contrário, μ_{jk} = 1(0) se $j = (\neq)i$.

iv. Compute $\Delta=\left\|U^{(l+1)}-U^{(l)}\right\|$. Se $\Delta>\varepsilon$, o l é incrementado em uma unidade e segue-se para etapa ii. Finalize o processo, se $\Delta\le\varepsilon$.

A partição rígida, de acordo com Bezdek et al. (2005), é adotada pelo método *k-means* que é um método clássico de agrupamentos de objetos. Este algoritmo forma *k* grupos a partir de uma amostra onde cada objeto pertença a apenas um dos grupos. As etapas do procedimento do algoritmo são: determinação aleatória de *k* objetos para serem os primeiros centróides dos agrupamentos; associação de cada um dos (*n*-*k*) objetos ao centróide mais próximo através da distância euclidiana para cada centróide; e recalculo de um novo centro com base nos elementos associados a cada centróide. Estes passos são repetidos até que ocorra a convergência do agrupamento. Esse algoritmo sofre influência da diferença de escala entre os valores dos atributos, e sendo assim, a normalização das variáveis é uma ação preliminar recomendável.

Bensaid et al. (1996) ressalta que todos os algoritmos de análise de agrupamentos estão sujeitos a, pelo menos, os seguintes problemas: escolher e validar o número correto de grupos e assegurar que a interpretação dos resultados correspondam à realidade estudada. Algoritmos de agrupamento baseados em otimização, como *k-means* ou *FCM*, tendem a produzir uma distribuição ou partição equitativa de objetos nos diferentes grupos.

2.4.2 Agrupamentos em séries temporais

Uma série temporal é uma série de observações (medições) feitas sequencialmente ao longo do tempo e associadas a uma variável de um determinado processo (MITSA, 2012). Dois tipos de objetos representam diferentes problemas de agrupamento e reconhecimento de padrões em séries temporais, quais sejam, Séries Temporais Univariadas (STU) e Séries Temporais Multivariadas (STM) (D'URSO et al, 2012; YANG e SHAHABI, 2004; LI e WEN, 2014).

Considerando-se uma série de observações ao longo do tempo $z_i(t) \quad (i=1,\ldots,k;\ t=1,\ldots,m)$, onde *k* é o número de variáveis medidas, *m* é o número de observações e *i* se refere à medição efetuada em cada instante de tempo, um objeto de STU compreende o caso no qual k=1. Caso contrário ($k \geq 2$) há o objeto compreende uma STM.

O agrupamento de séries temporais univariadas é amplamente explorado na literatura (LIAO, 2005) e frequentemente considerado como um problema de agrupamento de protótipo pontual no espaço multidimensional ($\Re^p$). Métricas tradicionais de similaridade

(distância euclidiana) e métodos de agrupamento tradicionais aplicados para agrupamento de dados estáticos (tais como *Fuzzy C-Means* e *k-means*), podem ser empregados neste caso (LIAO, 2005).

Por outro lado, agrupamentos envolvendo séries temporais multivariadas devem ser considerados quando há a necessidade de um tratamento integrado de mais de uma variável para o reconhecimento dos grupos e dos padrões, ou seja, para a extração do conhecimento a partir dos dados. Uma série temporal multivariada deve ser considerada como um todo e não pode ser transformada em uma concatenação de várias séries univariadas (YANG e SHAHABI, 2004). O agrupamento de STM compreende um problema de agrupamento de protótipo não-pontual e métricas tradicionais de similaridade, adequadas para o caso univariado, não se aplicam.

Desafios adicionais devem ser considerados no agrupamento de objetos representados por STM, tais como, a extração características, as métricas de similaridade e a seleção de variáveis apropriadas (redução da dimensão do problema) (COPPI et al, 2010; D'URSO et al, 2012; FONTES et al, 2012; KAVITHA e PUNITHAVALLI, 2010; RANI e SIKKA, 2012; SINGHAL e SEBORG, 2005). Alternativas específicas com base na *PCA* e Transformadas *Wavelets* são apresentadas em alguns trabalhos (D'URSO et al, 2012, CHAOVALIT et al, 2011). A seleção de variáveis adequadas pode ser realizada através do uso de técnicas específicas (tais como *PCA*).

Um objeto que consiste em uma STM pode ser representado pela seguinte matriz $m \times k$:

$$Z_i = \begin{bmatrix} z_{i1}(1) & \cdots & z_{ik}(1) \\ \vdots & \ddots & \vdots \\ z_{i1}(m) & \cdots & z_{ik}(m) \end{bmatrix}$$

em que Z_i é o objeto, z_{ij} (t) é a medição da variável j (j = 1, ..., k) no instante de tempo t (t = 1, ..., m) no objeto Z_i (i = 1, ..., n objetos). A coluna j contém a série histórica relacionada à variável j.

Uma outra forma de apresentação de uma STM é apresentada por D'URSO e MAHARAJ (2012):

$$X \equiv \{x_{ipt} : i = 1, I; p = 1, P; t = 1, T\}, \quad (64)$$

sendo i (i=1, I) representa o objeto, p (p=*1, P)* a variável e t (t=1, T) o instante de tempo; x_{ipt} representa a *p-ésima* variável referente ao *i-ésimo* objeto no instante t (Figura 1).

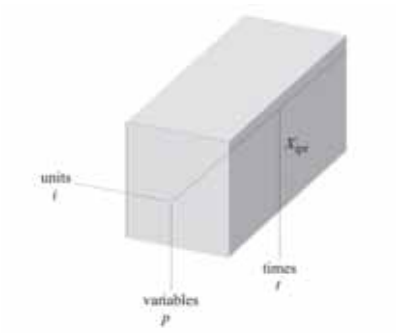


Figura 1 - Arranjo de objetos de séries temporais multivariadas ao longo do tempo.
Fonte: D'Urso e Maharaj (2012).

De um modo geral o agrupamento de séries temporais está associado ao reconhecimento de padrões (protótipos) de comportamento, dentro de uma janela de tempo, sendo importante para identificar tendências ou perfis específicos que podem ser úteis na caracterização de um determinado fenômeno (GAN et al., 2007). Devido a influência do fator tempo na formação de grupos, alguns algoritmos de agrupamentos de objetos atemporais são inapropriados para objetos caracterizados por séries temporais. Na engenharia o reconhecimento de padrões em séries temporais tem sido utilizado para a detecção e diagnóstico de falhas, levantamento de trajetórias ótimas e caracterização de perfis de consumo, sobretudo, de energia elétrica (ZALEWSKI, 2006; NIZAR et al., 2006; HAN et al., 2000).

De acordo com Liao (2005), um importante aspecto no agrupamento de séries temporais é a métrica adotada para verificar a proximidade entre elas. As métricas de proximidades são funções matemáticas que, sobretudo, incorporam nas suas restrições a natureza temporal dos dados. Particularmente, segundo Liao (2005), a medição de distância entre séries temporais contemplam algumas métricas específicas, tais como, distância do alinhamento temporal dinâmico (*Dynamic time Warping – DTW*), distância *Kullback-Lieber*, *J*-divergência e divergência de *Chernoff* (*J divergence and symmetric Chernoff information divergence*) e índice de dissimilaridade baseado na função de correlação cruzada entre duas séries. Existem métricas de distâncias que foram desenvolvidas especialmente para séries temporais multivariadas (QUADRO 1) como, por exemplo, a distância *SPCA* (SINGHAL e SEBORG, 2005).

De acordo com Liao (2005), para duas séries temporais, $Q = q_1, q_2, ..., q_i, ..., q_n$ *e* $R = r_1, r_2, ..., r_j, ..., r_m$, a distância *DTW* alinha estas séries temporais de tal modo que a diferença entre elas seja a mínima possível. Diante disso, considera-se uma matriz $n \times m$, onde o elemento (i, j) desta matriz contém a distância $d(q_i, r_j)$ entre dois pontos q_i *e* r_j. A distância euclidiana é a métrica de similaridade adotada. Um alinhamento temporal, $W = w_1, w_2,..., w_k,..., w_K$ onde max $(m,n) \leq K \leq m + n - 1$, é um conjunto dos elementos da matriz de distância que satisfaz três restrições, quais sejam, contorno, continuidade e monotonicidade. A restrição de contorno requer a fixação do primeiro e último ponto do alinhamento temporal, isto é, $w_1 = (1,1)$ e $w_K = (m,n)$. A restrição de continuidade limita a admissão de passos para as células adjacentes. A restrição de monotonicidade impõe que os pontos do alinhamento temporal sejam monotonicamente espaçados no tempo. Dessa forma, busca-se um alinhamento temporal que tem a distância mínima entre as duas séries. Matematicamente:

$$d_{DTW} = \min \frac{\sum_{k=1}^{K} w_k}{K}, \tag{65}$$

A métrica de similaridade de *Kullback-Liebler* considera que P_1 *e* P_2 são matrizes de probabilidade de transição de duas cadeias de Markov com distribuição s de probabilidade para cada uma, e $p_{1_{ij}}$ e $p_{2_{ij}}$ as probabilidade de transição em P_1 e P_2, respectivamente, ou seja, a probabilidade de uma série temporal i evoluir para um comportamento parecido com uma série temporal j ou vice-versa. A distância assimétrica de *Kullback-Liebler* de duas distribuições de probabilidade é:

$$d(p_{1i}, p_{2i}) = \sum_{j=1}^{s} p_{1_{ij}} \log(p_{1_{ij}} / p_{2_{ij}}) \tag{66}$$

A versão simétrica da distância *Kullback-Liebler* de duas distribuições é:

$$d(p_{1_i}, p_{2_i}) = \left[d(p_{1_i}, p_{2_i}) + d(p_{2_i}, p_{1_i}) \right] / 2 \tag{67}$$

A distância média entre P_1 e P_2 é então:

$$d(P_1, P_2) = \sum_{i=1,s} d(p_{1i}, p_{2i}) / s \tag{68}$$

Para as distâncias *J*-divergência e divergência de *Chernoff*, considera-se que $f_T(\lambda_s)$ e $g_T(\lambda_s)$ são estimadores de duas matrizes espectrais, para dois diferentes objetos de séries estacionárias de T períodos, onde $\lambda_s = 2\pi s/T$, $s = 1, 2, ..., T$, a *J*-divergência e divergência de *Chernoff* são, respectivamente, computadas como:

$$J(f_T;g_T)=\frac{1}{2}T^{-1}\sum_s\left(tr\left\{f_T g_T^{-1}\right\}+tr\left\{g_T f_T^{-1}\right\}-2p\right) \quad (69)$$

e

$$JB_\alpha(f_T;g_T)=\frac{1}{2}T^{-1}\sum_s\left(\log\frac{\left|\alpha f_T+(1-\alpha)g_T\right|}{\left|g_T\right|}+\log\frac{\left|\alpha g_T+(1-\alpha)f_T\right|}{\left|f_T\right|}\right) \quad (70)$$

sendo $0<\alpha<1$ e p é o tamanho da matriz espectral.

Por fim, para a função de correlação cruzada, considera que $\rho_{i,j}^2(\tau)$ denota a correlação cruzada entre duas séries temporais x_i e v_j com τ defasagem no tempo. Um índice de dissimilaridade baseado na função de correlação cruzada é definido como:

$$d_{i,j}=\sqrt{(1-\rho_{i,j}^2(0))/\sum_{\tau=1}^{\max}\rho_{i,j}^2(\tau)} \quad (71)$$

no qual *max* é a defasagem máxima entre duas séries temporais. A quantificação da similaridade deste índice pode ser definida como:

$$s_{i,j}=\exp(-d_{i,j}) \quad (72)$$

Uma taxonomia para os diferentes métodos de agrupamento de séries temporais univariaveis e multivariáveis (LIAO, 2005) é apresentada a seguir contemplando-se três categorias:

- Métodos baseados em dados brutos.

São métodos que trabalham com dados na mesma forma em que foram coletados ou medidos, quer no domínio do tempo ou da frequência (LIAO, 2005). As métricas de similaridade utilizadas manipulam os dados na forma original. O Quadro 1 apresenta uma síntese das aplicações (LIAO, 2005) de algoritmos de agrupamentos em séries temporais baseados em dados brutos. O *FCM* (apresentado na seção 2.4.1.2) pode ser considerado um método de agrupamento baseado em dados brutos e, na sua formulação padrão, se aplica ao agrupamento de séries temporais univariadas adotando a distância euclidiana como métrica de similaridade.

Quadro 1 - Algoritmos de agrupamentos de séries temporais que se baseiam em dados brutos.

TIPO DE SÉRIE TEMPORAL	**AUTORES**	**APLICAÇÃO**
Univariada	Golay et al. (1998).	Aplicaram o algoritmo *FCM* em dados de funções de ressonância magnética (*functional Magnetic Resonance Imaging – fMRI*) que são séries temporais univariadas de igual comprimento, onde foram utilizadas diferentes tipos de medidas de distância, tais como, distância euclidiana e correlação cruzada. Várias abordagens de pré-processamento de dados foram feitas e o efeito do número de grupos também foi discutido. O número de grupos é selecionado adotando-se como índice de validação a variância interna de cada grupo.
	Wijk e Selow (1999).	Realizaram um agrupamento hierárquico de dados diários de consumo de energia baseado na raiz quadrada da distância média.
	Kumar et al. (2002).	Propuseram uma função de distância com base nos modelos gaussianos assumindo a independência de erros e usou um método de agrupamento hierárquico para definir um número desejável de grupos. O método proposto foi aplicado em dados relacionados a padrões de sazonalidade em varejo.
	Wismüller et al. (2002).	Mostraram que o método de quantização vetorial (CHERKASSKY e MULIER, 1998) pode ser aplicado no diagnóstico de anomalias cerebrais através de exames de imagens de ressonância magnética. Foi realizado um procedimento não supervisionado de agrupamento hierárquico para revelar a estrutura

		subjacente na amostra de dados com incremento gradual na qualidade de agrupamento.
	Möller – Levet et al. (2003).	Propuseram *Short Time Series* (STS), uma medida que verifica a similaridade no formato provocado pela relativa mudança de amplitude e da correspondente informação temporal de uma amostra de séries com intervalos discretos e desiguais. Todas as séries são consideradas amostradas nos mesmos instantes de tempo. O algoritmo *FCM* foi adotado com adaptação para o uso da métrica *STS* na função objetivo.
	Policker e Geva (2000).	Modelaram uma série temporal não estacionária com desempenho semelhante a modelagem através de um modelo oculto de Markov. Um procedimento de agrupamento *fuzzy* foi desenvolvido por Gath e Geva (1989) e aplicado em análise de exames eletroencefalogramas.
	Liao et al. (2002).	Aplicaram vários algoritmos de agrupamento, incluindo *k-means*, *FCM* e agrupamento de dados genéticos para simulação de dados de batalhas dispostos em séries temporais multivariadas de comprimento desigual.
Multivariada	Košmelj e Batagelj (1990).	Conceberam um algoritmo de agrupamento de séries temporais multivariadas de igual comprimento, originalmente desenvolvido para agrupamentos de objetos atemporais. O método combina técnicas de inteligência computacional com o método do corte transversal (*cross-sectional*) para reconhecer padrões de séries temporais de consumo de energia elétrica
	Kakizawa	Baseando-se nas abordagens do agrupamento

	et al. (1998).	hierárquico e não-hierárquico (*k-means*) buscaram a identificação de grupos de vetor de dados multivariados de terremotos e explosões de minas. As medidas de distâncias de *J*-divergência e divergência de *Chernoff* (RUEDA e HERRERA, 2006) foram usadas para computar as dissimilaridades entre as funções de autocovariância correlatas.
	Shumway (2003)	Investigou o agrupamento de séries temporais não-estacionárias através da aplicação de versões localmente estacionárias baseadas em um medida linear discriminante, a divergência de Kullback-Leibler (RUEDA e HERRERA, 2006).
	Liao (2005).	Desenvolveu um processo constituído de duas etapas para agrupamento de séries temporais multivariadas de comprimento igual ou desigual. Na primeira etapa através do algoritmo de agrupamento *k-means* ou *FCM* converteu séries temporais multivariáveis em séries temporais univariáveis. A segunda etapa utiliza o *k-means* ou *FCM* novamente para formar grupos de séries temporais univariadas, expressa como matrizes de probabilidade de transição. A distância Euclidiana é utilizada na primeira etapa, ao passo que várias medidas de distância, incluindo a distância simétrica de Kullback-Liebler, são empregadas na segunda etapa.

- Métodos baseados em extração de características.

Estes métodos não utilizam atributos originais dos objetos e sim características extraídas dos próprios objetos que assumem a função dos atributos. Isto implica que a dimensionalidade original dos objetos, normalmente, tende a se reduzir. O Quadro 2 apresenta

uma síntese sobre alguns algoritmos de agrupamentos de séries temporais univariadas que se baseiam na extração de características.

Quadro 2 – Algoritmos de agrupamentos de séries temporais que se baseiam em extração de características.

TIPO DE SÉRIE TEMPORAL	**AUTORES**	**DESCRIÇÃO**
Univariada	Fu et al. (2001).	Descreveram o uso de mapas auto organizáveis (*self-organizing maps – SOM*) para o agrupamento ou reconhecimento de padrões de representações espaciais, obtidas em decorrência da conversão das séries temporais de variáveis de desempenho de programas computacionais.
	Goutte et al. (1999).	Agrupou séries temporais de imagens de ressonância magnética funcional (*Functional Magnetic Resonance Imaging*–fMRI) de atividades cerebrais em grupos de voxels (menor ponto tridimensional de uma imagem digital) derivados dos pixels (menor ponto bidimensional de uma imagem) usando dois algoritmos: *k-means* e agrupamento hierárquico de Ward.
	Owsley et al. (1997).	Um algoritmo de refinamento da sequência de grupos (*Sequence Cluster Refinement Algorithm–SCRA*) foi desenvolvido por Owsley et al. (1997) para o monitoramento de dados de máquinas-ferramenta, (tornos, fresas e plainas, por exemplo) , através de Modelo Oculto de Markov (*Hidden Markov Models–HMM*), para acompanhar suas condições de operações.
	Shaw e King	Agruparam séries temporais com base na *PCA* através da aplicação de dois algoritmos hierárquicos, o algoritmo de Ward de variância mínima e o algoritmo

	(1992).	de distância mínima para resolver problemas de biotecnologia.
	Vlachos et al . (2003).	Propõem um método que suporta várias medidas de distância, apresentando uma abordagem para realizar o agrupamento incremental de séries temporais em várias resoluções usando a transformada *Haar wavelet*, a fim de extrair informações de grandes bancos de dados.

Especificamente para as séries temporais multivariadas, D'Urso e Maharaj (2012) propuseram uma abordagem de agrupamento baseado no algoritmo *FCM* com base nas funções de autocorrelação entre as séries temporais. O método foi aplicado numa base de dados oriundos da medição do nível de poluição do ar feita por doze estações de monitoramento na cidade de Roma. Outros trabalhos voltados para métodos de agrupamentos baseados em extração de características e aplicados a séries temporais multivariadas compreendem Kakizawaetal (1998), Singhal e Seborg (2005), Abonyietal. (2005), Wu e Li (2005). Em particular, o método proposto por Singhal e Seborg (2005) baseia-se na extração de características através do cálculo do grau de similaridade entre as séries temporais multivariadas usando dois fatores de similaridade. Um fator de similaridade é baseado na *PCA* e os ângulos que geram a orientação espacial dos subespaços (*SPCA*), enquanto o outro é baseado na distância Mahalanobis entre as séries temporais multivariadas.

- Métodos baseados em modelos.

Consistem em métodos baseados em modelos identificados a partir dos dados brutos. Nesta abordagem cada série temporal é representada por uma estrutura de modelo empírico ou por uma distribuição de probabilidade. Séries temporais são consideradas semelhantes quando os modelos que as descrevem são semelhantes. O Quadro 3 apresenta uma síntese de algoritmos de agrupamentos de séries temporais baseados em modelos.

Quadro 3 – Algoritmos de agrupamentos de séries temporais baseados em modelos.

TIPO DE SÉRIE TEMPORAL	**AUTORES**	**DESCRIÇÃO**
Univariada	Baragona (2001)	Avaliou três métodos meta-heurísticos para particionar um conjunto de séries temporais em grupos de tal forma que o valor absoluto máximo de correlação cruzada entre cada par de séries temporais que pertencem ao mesmo grupo fosse maior do que um limiar pré-definido. As correlações cruzadas são obtidas a partir dos resíduos dos modelos da série temporal original.
	Beran e Mazzola (2001).	Baseado em modelos hierárquicos de suavização (*hierarchical smoothing models –HISMOOTH*) para descrever a relação entre a estrutura simbólica de uma partitura e seu desempenho ao longo do tempo. Os modelos utilizam larguras de banda, faixas de freqüências necessárias para constituir o sinal, como parâmetro chave na caracterização destas estruturas simbólicas.
	Kalpakis et al. (2001).	Basearam-se no agrupamento de modelos *Autoregressive Integrated Moving Average - ARIMA* (*Autoregressive - AR*) obtidos por séries temporais usando a distância euclidiana entre o *Linear Predictive Coding Cepstra* de duas séries temporais como sua medida de dissimilaridade. Os coeficientes *Cepstrais* para uma de séries temporais *AR* são derivados a partir dos coeficientes auto-regressivos. A partição foi feito usando o método *medoids* (KAUFMAN e ROUSSEEUW, 1990) e a qualidade de agrupamento foi avaliada pelo índice de silhueta.

	Maharaj (2000).	Desenvolveu um processo de agrupamento hierárquico aglomerativo que é baseado no *p*-valor para testar a hipótese nula de que duas séries temporais tem a mesma representação de um modelo linear *AR* de ordem *k*, assumindo-se para cada série temporal a sua estacionariedade.
	Piccolo (1990).	Introduziu a distância euclidiana no agrupamento de estruturas dinâmicas (especificamente a classe de modelos *ARIMA*). A matriz de distância entre pares de modelos de séries temporais foi então processada pelo método da maior distância para a construção do dendrograma, árvore gerada pelo agrupamento hierárquico.
	Wang et al. (2002).	A estrutura foi apresentada por Wang et al. (2002) para o monitoramento do desgaste da máquina-ferramenta em um processo de usinagem por meio de modelos *HMM*. Os vetores de características são extraídos a partir dos sinais de vibração medidos durante operações de torneamento através da análise *wavelet*.
	Xiong e Yeung (2002).	Propuseram um método baseado em agrupamento de modelos *ARIMA* obtidos por séries temporais univariadas. Eles assumiram que as séries temporais são geradas por *k* diferentes modelos *ARIMA*, com cada modelo correspondendo a um grupo de interesse. Um algoritmo *Expectation-Maximization* (*EM*) foi usado para melhorar o desempenho dos coeficientes e parâmetros dos modelos.
Multivariada	Bienacki et al (2000)	Propuseram o *Integrated Completed Likelihood* (*ICL*), um critério para a escolha de um modelo de mistura gaussiana e um número ótimo de grupos. O

		critério *ICL* é uma adaptação do Critério de informação Bayesiano (*Bayesian Information Criterion - BIC*) para superar a tendência a superestimar o número de grupos.
	Li e Biswas (1999).	Apresentam um método de agrupamento de séries temporais utilizando a representação do Modelo Oculto de Markov (*Hidden Markov Models–HMM*).
	Li et al. (2001).	Apresentam um algoritmo de agrupamento *HMM* que usa *BIC* como critério de seleção de modelos em diferentes níveis e explora as características da função monotônica para melhorar a sua performance.
	Oates et al. (1999).	Considerando-se que um conjunto de séries temporais multivariada fosse gerada de acordo com os modelos de Markov, Oates et al. (1999) apresentaram um método de agrupamento híbrido para determinar automaticamente o número *k* de gerar *HMMs*, e para aprender os parâmetros dessas *HMMs*. Um algoritmo padrão de agrupamento hierárquico aglomerativo foi aplicado para obter uma estimativa inicial de k e para formar os grupos iniciais usando o *DTW* para avaliar a similaridade.
	Ramoni et al. (2002).	Apresentam o algoritmo *Bayesian algorithm for clustering by dynamics* (*BCD*). Dado um conjunto *S* de *n* membros de uma série temporal univariada, *BCD* transforma cada série em uma cadeia de Markov (*Markov Chains - MC*) e depois efetua os agrupamentos conforme suas similaridades mediante um método de agrupamento aglomerativo sem supervisão, resultando assim em grupos de objetos que determinam a matriz de transição. A semelhança

		entre duas matrizes de transição estimada é medida como uma média da distância de Kullback – Liebler simetrizada entre linhas correspondentes nas matrizes.

Exemplos mais recentes de estudos das três abordagens supracitadas são apresentados por Rani et al. (2012). Os algoritmos concebidos para o agrupamento de séries temporais multivariadas são resultantes de adaptações ou mutações feitas das métricas de distâncias dos seus algoritmos correlatos de agrupamento de séries temporais univariadas. Vale mencionar que os algoritmos de agrupamento em séries temporais univariadas (KEOGH e KASETTY, 2003; LIAO, 2005; RANI et al. 2012) utilizam abordagens padrões que usam modelos de agrupamentos protótipo pontual (LIAO, 2005; BEZDEZ et al., 2005). No entanto, o reconhecimento de padrões em séries temporais multivariadas representa um problema mais complexo (problema protótipo não pontual) com características intrínsecas (SINGHAL e SEBORG, 2005). Esse tipo de problema, conforme já foi dito, não pode ser resolvido diretamente com algoritmos clássicos de agrupamento tais como o *FCM*, pelo menos em sua forma tradicional. São necessários algoritmos especiais ou adaptação de algoritmos já existentes (COPPI et al., 2010).

2.4.3 Validação de grupos

A fim de comparar os resultados de diferentes algoritmos de agrupamentos, bem como avaliar se o número e a qualidade dos grupos obtidos são satisfatórios, torna-se necessário estabelecer critérios para a validação final dos resultados de agrupamento.

A validação de grupos é uma etapa final do processo de agrupamento. Sua importância aumenta à medida que se pretende alcançar um resultado de agrupamento consistente com a realidade sob análise. As medidas de validação de grupos avaliam as características de coesão de cada grupo e o quanto cada grupo está distante dos demais. A validação pode ser usada para avaliar o desempenho de um determinado método de agrupamento ou comparar desempenhos de vários métodos de agrupamentos aplicados numa mesma base de dados. Existem três tipos de critérios para investigar a qualidade do agrupamento, quais sejam, externo, interno e relativo (GAN ET AL., 2007). Sendo que o foco da discussão será dirigido para o critério relativo pelas razões que serão esclarecidas posteriormente.

No critério externo, a avaliação de uma partição de grupos ou agrupamento (P), produzido por um método, é baseada na comparação com um agrupamento de referência ou agrupamento preliminarmente particionado por um critério subjetivo de interesse (Q), imposto ao conjunto de dados para garantir um melhor controle de qualidade na formação de grupos. No caso presente, não houve uso de um agrupamento externo de referência, e este tipo de critério não pôde ser usado neste trabalho.

O critério interno avalia a partição de grupos ou de agrupamentos a partir das próprias características dos dados sob análise para estabelecer uma regra de parada ou um critério de otimização. O critério interno é utilizado para os dois tipos de agrupamentos: hierárquico e não hierárquico. No agrupamento hierárquico a idéia é validar os sucessivos arranjos de agrupamentos (dendrograma) gerados por meio de um método de formação de grupos. No agrupamento não hierárquico os parâmetros são adequadamente selecionados por um processo de otimização a partir da predefinição do número de grupos que se quer obter. A avaliação da qualidade do agrupamento pode ser obtida através da própria função objetivo adotada no algoritmo (considerando-se, por exemplo, os modelos genéricos *c-means*). Neste trabalho, o critério interno foi adotado para encontrar as partições para os agrupamentos não hierárquicos *FCM* e *FCM* modificado, tendo como predefinições os números de grupos sugeridos pelo critério relativo.

O critério relativo permite avaliar a qualidade de grupos gerados por um algoritmo ou entre diferentes algoritmos e a contribuição individual de cada objeto para o desempenho da qualidade dos grupos. O índice de silhueta (ROUSSEEUW, 1987) possui estas características e foi originalmente desenvolvido para avaliar a qualidade de agrupamentos com partições mutuamente excludentes (partição *crisp)*. Vale ainda dizer que existe um índice de silhueta modificado para a validar agrupamentos *fuzzy* (pertinências dos objetos no intervalo [-1;1]).

Considerando N_L objetos pertencentes ao grupo L e um total de G grupos (G ≥ 2), o índice de silhueta para cada objeto é:

$$S_i^L = \frac{b_i^L - a_i^L}{\max\left\{a_i^L, b_i^L\right\}} \qquad i = 1,\ldots,\sum_{k=1}^{G} N_k \tag{73}$$

sendo s_i^L é o índice de silhueta do objeto *i* pertencente ao grupo *L* (*1*≤*L*≤*G*). a_i^L (Eq. (74)) é a distância média entre objeto *i* e todos os outros objetos pertencentes ao mesmo grupo. b_i^L (Eq. (75)) é a mínima distância média entre o objeto *i* e os objetos pertencentes aos outros grupos.

$$a_i^L = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{N_L-1} d_{ij}^{L,L}}{N_L - 1} \qquad i = 1,\ldots,N_L \tag{74}$$

$$b_i^L = \min_{\substack{Z=1 \\ Z\neq L}}^{W} \left(\frac{\sum_{j=1}^{N_L} d_{ij}^{L,Z}}{W} \right) \qquad \text{onde } W = \sum_{\substack{K=1 \\ K\neq L}}^{G} N_K \tag{75}$$

nos quais $d_{ij}^{L,K}$ é distância entre o objeto *i*, pertencente ao grupo *L*, e o objeto *j* pertencente ao grupo *K*.

A distribuição dos índices de silhueta entre todos os objetos permite avaliar o nível de sobreposição entre os grupos e, portanto, a qualidade do agrupamento. Conforme a Eq. (73), quando a pertinência de um objeto em relação a um dado grupo está bem definida, o seu respectivo índice de silhueta se aproxima da unidade. Por outro lado, valores negativos significam que há mais homogeneidade entre os grupos e menos coesão interna, o que representa o pior cenário. Nos gráficos dos resultados que serão apresentados pelo presente trabalho o índice de silhueta de um objeto será representado por I_s e a média dos índices de silhueta entre todos os objetos será denominada de Índice Global de Silhueta (*IGS*).

O índice de silhueta pode ser também aplicado na validação de agrupamentos de objetos de séries temporais univariadas. Em relação à validação de agrupamentos de séries temporais multivariadas, deve-se adotar uma métrica de similaridade (distância entre os objetos) apropriada para tal fim. A fórmula a seguir esclarece melhor esta questão:

$$S_i^L = (b_i^L - a_i^L) / \max\left\{a_i^L, b_i^L\right\} \qquad i = 1,\ldots,\sum_{k=1}^{G} N_k \tag{76}$$

sendo S_i^L $\left(-1 \le S_i^L \le 1\right)$ é o índice de silhueta de i^{th} objeto pertencente ao grupo *L* (*1*≤*L*≤*G*). a_i^L (Eq. (77)) é a distância média entre o i^{th} objeto e todos outros objetos pertencentes à mesma

amostra do grupo. b_i^L (Eq. (78)) é a distância média mínima entre o i^{th} objeto e os objetos pertencentes aos outros grupos.

$$a_i^L = \sum_{\substack{j=1 \\ j\neq 1}}^{N_L-1} d_{ij}^{L,L} / (N_L - 1) \tag{77}$$

$$b_i^L = \min_{\substack{Z=1 \\ Z\neq L}}^{W} \left(\sum_{j=1}^{N_L} d_{ij}^{L,Z} / W \right) \tag{78}$$

onde $W = \sum_{\substack{K=1 \\ K\neq L}}^{G} N_K$.

Considerando que cada objeto pode ser representado somente por um vetor (componente principal com autovalor alto), $d_{ij}^{L,K}$ é a distância Euclidiana entre o componente principal do i^{th} objeto, pertencente ao grupo L, e o componente principal do j^{th} objeto pertencente ao grupo.

Existe também uma versão *fuzzy* para o índice de silhueta (CAMPELLO e HRUSCHKA, 2006):

$$S_{fuzzy} = \frac{1}{N} \frac{\sum_{q=1}^{N} \left(b'(q) - a'(q)\right)^{\alpha} s(q)}{\sum_{q=1}^{N} \left(b'(q) - a'(q)\right)^{\alpha}} \tag{79}$$

no qual $\alpha \geq 0$ é um coeficiente de ponderação e *s(q)* = S_i^L (silhueta crisp). $a'(q)$ e $b'(q)$ são, respectivamente, o primeiro e o segundo maiores elementos da *q*-ésima coluna da matriz de pertinências. O índice de silhueta *fuzzy* incorpora nas suas operações matemáticas as medidas da matriz de pertinências que revelam regiões de alta densidade aumentando a importância dos objetos próximos aos protótipos enquanto reduz a importância de objetos em áreas de sobreposição de grupos. Com isso o índice de silhueta *fuzzy* torna-se mais apropriado para a validação de agrupamentos fuzzy.

Em relação aos resultados dos algoritmos de agrupamentos *fuzzy*, quando os objetos são designados aos grupos que conferem as maiores pertinências, o agrupamento obtido pode também ser avaliado pelos índices *crisp* de validação de grupos. Este procedimento foi

adotado na análise de dados desta tese visando a comparação entre os algoritmos *fuzzy* e os algoritmos propostos, através de um critério relativo de validação de grupos.

Arbelaitz et al. (2013), comparou a qualidade de partição de trinta índices relativos de validação de grupos, considerando diferentes fatores, tais como, número de grupos, dimensionalidade dos dados, sobreposição de grupos, densidade dos grupo e ruídos dos dados. Os desempenhos dos índices foram avaliados utilizando critérios internos de validação. As partições dos grupos foram obtidas utilizando mais de um método de formação de grupos (*k-means*, *Ward* e distância média) para possibilitar um melhor controle dos resultados. Os resultados demonstraram que não houve um índice que fosse superior aos demais em todas as condições impostas experimentalmente. Todavia, o índice de silhueta, em relação à aplicação de diferentes métodos de formação de grupos, teve um desempenho superior aos demais, demonstrando ser apropriado para validar a coesão e separação de grupos.

3 REVISÃO BIBLIOGRÁFICA

3.1 CONSIDERAÇÕES INICIAIS

Há vários estudos voltados para o reconhecimento de padrões de séries temporais do consumo. Alguns autores podem ser citados, tais como Gerbec et al.(2002), que contemplam o reconhecimento de padrões de consumo usando um método de agrupamento hierárquico com a determinação do número adequado de padrões; Gemignani et al. (2009), que combinam os métodos hierárquicos e não-hierárquicos para melhorar o desempenho do reconhecimento de diferentes padrões de consumo no mesmo nível de tensão e Zalewski (2006), que usou a lógica "fuzzy" para estabelecer múltiplas pertinências de uma trajetória de consumo a diferentes grupos identificados pelo processo de reconhecimento de padrões e realizou o agrupamento de perfis de carga, a fim de classificar as subestações em grupos homogêneos de acordo com o pico de consumo.

Cita-se também Nizar et al. (2006) que propõem a combinação dos métodos de seleção de características (*Feature Selection*) e de extração de conhecimento de base de dados (*Knowledge-Discovery in Databases – KDD*) (HAN et al., 2000), utilizados para melhorar o desempenho do reconhecimento de padrões.

Um estudo recente sobre a extração de características para reconhecimento de padrões de hábito de usuários do sistema elétrico (DASWIN; XINGHUO, 2011) adotou o método de caracterização de padrões e sumarização incremental (*Incremental Summarization and Pattern Characterization – ISPC*) como procedimento de agrupamento.

Outros autores têm adotado o método *Fuzzy C-Means* (*FCM*) nas aplicações envolvendo reconhecimento de padrões de consumo de energia elétrica para proporcionar um melhor nível de coesão e de discriminação entre os grupos identificados (GERBEC et al., 2005; ZAKARIA et al., 2006; ANUAR e ZAKARIA, 2010) comparando o seu desempenho com diferentes algoritmos concorrentes. Chicco et al. (2012) fez uma revisão bibliográfica sobre os principais procedimentos ou algoritmos de agrupamentos concorrentes do método *FCM.*

Tradicionalmente, o reconhecimento de padrões de curvas de cargas é utilizado como um dos procedimentos para consubstanciar um planejamento tarifário mais eficiente. Dentre outras aplicações importantes, pode-se citar o gerenciamento de cargas pelo lado da demanda

(FERREIRA et al., 2003), que explora diversas possibilidades de se obter a contribuição dos consumidores para a racionalização do consumo da energia elétrica.

A curva de consumo de energia elétrica (também denominada de curva de carga) pode ser analisada em conjunto com outras séries temporais para elevar o nível de conhecimento sobre o comportamento da demanda do sistema elétrico (TSEKOURAS et al., 2011), a exemplo das variáveis meteorológicas (LIN et al. 2006, CHICCO et al., 2003). Este é o caso do consumo de energia de refrigeradores (GELLER et al., 1998; GOLDMAN et al., 2005). A temperatura externa tem um forte efeito sobre a eficiência térmica desses equipamentos, especialmente no que diz respeito ao modo de uso (frequência de abertura da porta do refrigerador), o que contribui para desvios na temperatura interna e nas condições de acondicionamento (CLINCH e HEALY, 2001).

O reconhecimento de padrões envolvendo apenas curvas de carga é tipicamente um problema de protótipo pontual (*point-prototype*), ou série temporal univariada (Bezdek et al., 2005). Conforme Keogh e Kasetty (2003), Liao (2005) e Rani et al. (2012) o reconhecimento padrões em séries temporais univariadas, em geral, é um problema de média complexidade. No entanto, o reconhecimento de padrões em séries temporais multivariadas (curvas de carga associadas a outras variáveis) representa um problema mais complexo de protótipo não pontual (*non point-prototype*), (BEZDEK et al., 2005), com características e dificuldades intrínsecas (SINGHAL e SEBORG, 2005).

Esse tipo de problema não pode ser resolvido diretamente com modelos clássicos de agrupamento como o *FCM* convencional. No caso de séries multivariadas, são necessários métodos mais elaborados para a extração de características (COPPI et al., 2010; KAVITHA e PUNITHAVALLI, 2010; D'URSO et al., 2012; FONTES et al., 2012) e desafios adicionais necessitam ser devidamente tratados, tais como a métrica de similaridade e o projeto do classificador.

3.2 MARCOS TEÓRICOS PARA ANÁLISE DE DADOS UNIVARIADOS

No atual modelo de regulação do setor elétrico brasileiro, programas de eficiência energética, estimulados por meio de benefícios fiscais, são fomentados pelo governo para melhorar o desempenho do setor de distribuição de energia elétrica. As tipologias destes

programas variam e, em particular, destaca-se o programa que compreende a troca de refrigeradores velhos por novos em comunidades de baixa renda.

Geller et al. (1998) revisaram o estado da arte sobre conservação de energia elétrica e programas de gestão da demanda no Brasil, afirmando que, embora algum progresso tenha sido feito, ainda há enorme potencial de melhoria da eficiência de energia. Geller et al. (1998) destacam medidas para a atração de investimentos, tais como, o Programa Nacional de Conservação de Energia Elétrica (PROCEL), através de projetos com a participação da iniciativa privada para incentivar a melhoria de serviços públicos locais e em parceria com bancos fomentadores de linhas de financiamentos específicos para patrocinar projetos de eficiência energética. Os autores destacaram que as concessionárias de energia elétrica podem desempenhar um papel muito maior na promoção do uso eficiente de energia elétrica. Além disso, foram apresentadas algumas recomendações tais como a obrigatoriedade de que todos os serviços públicos de distribuição de energia elétrica definam metas e operem programas de gestão da demanda para os seus clientes trabalhando em conjunto com o PROCEL, utilização de recursos da iniciativa privada nos programas do governo (por exemplo, 1% das receitas operacionais) e autorização para que projetos de economia de energia participem em processos licitatórios.

O Programa Nacional de Conservação de Energia Elétrica (PROCEL) foi criado pelo governo federal em 1985, e é executado pela Eletrobrás, com recursos da empresa, da Reserva Global de Reversão (RGR) e de entidades internacionais. De acordo com o site oficial, em 2013, a Eletrobrás PROCEL alcançou uma economia de 9,74 milhões de megawatts-hora (MWh), o equivalente a 2,1% de todo o consumo nacional de energia elétrica naquele ano. Em termos práticos, evitou-se o consumo anual de energia elétrica de aproximadamente 5 milhões de residências brasileiras. Os impactos ambientais foram também promissores: em 2013, evitou-se o equivalente a 935 mil toneladas de CO2 em decorrência do controle das emissões de gases de efeito estufa, o que corresponde às emissões de 321 mil veículos em um ano.

Goldman et al. (2005) apresentaram uma análise abrangente das Empresas de Serviços de Conservação de Energia (*Energy Services Company - ESCO*) dos EUA sobre as tendências da indústria e de seu desempenho. Foi feito o levantamento das empresas para estimar o tamanho total da indústria e um banco de dados foi desenvolvido contendo mercados-alvo e características típicas de projeto, energia poupada e economia do cliente. Conclusões

importantes foram obtidas acerca da formulação de políticas para o setor elétrico em outros países, tais como, a importância do apoio adequado para estimular indústrias *ESCO* do setor privado e a importância da liderança do governo na forma de políticas que promovam a eficiência energética, sinalizando a prioridade destas ações como incentivo ao desenvolvimento de uma indústria de serviços de eficiência energética.

Neste contexto, a análise dos resultados da implementação dos programas de eficiência energética tem servido de norteamento para melhorias de ações e políticas públicas voltadas para a racionalização do consumo de energia elétrica pelo lado da demanda. Um aspecto importante nesta análise é recorrer a técnicas de reconhecimento de padrões de consumo, para monitorar alterações de regime de comportamento temporal, capaz de sinalizar os benefícios esperados em decorrência das modulações das cargas e ganhos energéticos. A análise multivariada de dados e mais especificamente a análise de agrupamentos é uma ferramenta que atende a estas expectativas, especialmente por meio de reconhecimento de padrões em séries temporais representadas por curvas de cargas (NAZARKO e STYCZYNSKI, 1999; GERBEC et al., 2002; ZALEWSKI, 2006; GEMIGNANI et al., 2009; TSEKOURAS et al., 2011, CHICCO et al., 2012).

Nazarko e Styczynski (1999) apresentaram uma discussão sobre a modelagem de curva de carga, destacando a sua importância para a análise econômica e planejamento da operação de sistemas de distribuição. Particularmente, os autores chamaram atenção para o gerenciamento pelo lado da demanda em mercados desregulamentados e cuja previsão da demanda diária está baseada em dados de pesquisa de clientes com alto grau de diversidade no modo de uso da energia elétrica. O trabalho dos autores apresentou métodos de modelagem de previsões e reconhecimento de padrões de cargas para o planejamento de longo prazo dos sistemas de distribuição de energia. Os métodos de modelagem basearam-se em procedimentos estatísticos, lógica *fuzzy* e redes neurais artificiais.

Gerbec et al. (2002) apresentaram uma tipificação da curva de carga utilizando um algoritmo de agrupamento hierárquico e destacaram a vantagem deste algoritmo na seleção do número adequado de grupos. Os autores partiram do pressuposto que as informações sobre o padrão de consumo dos consumidores é de primordial importância para a gestão da demanda e planejamento do sistema elétrico em ambiente desregulamentado. Uma metodologia para determinação do perfil de consumo diário foi proposta baseando-se na análise de agrupamento através do método hierárquico. O método foi testado em 120 perfis medidos.

Zakaria et al. (2006) destacaram que no ambiente de mercado desregulamentado da Malásia os consumidores podem comprar energia elétrica de qualquer empresa, independentemente do tamanho e localização. Por conseguinte, a determinação de perfil de carga do cliente pode facilitar a análise de informações para a previsão, planejamento das operações e planejamento da demanda, favorecendo as empresas de serviços públicos a obter melhores estratégias de marketing e melhoria de eficiência.

Levando-se em consideração as mudanças de política energética em vigor na Polônia, Zalewski et al. (2006) aplicaram lógica *fuzzy* para o agrupamento e tipificação de curvas de carga e demonstraram a importância deste método para a classificação e modelagem na reorganização dos sistemas tarifários, formulação de estratégias de preços e planejamento de capacidade. O objetivo foi modelar a variação do consumo de energia elétrica a partir da estrutura Takagi-Sugeno de inferência *fuzzy*. O autor realizou o agrupamento de perfis de carga a fim de classificar as subestações em grupos homogêneos de acordo com o pico de consumo. Foram feitas regressões, com base na defuzzificação do sistema de inferência *fuzzy*, expressando a relação entre os picos de cargas das subestações e um conjunto de características de clientes.

Gemignani et al. (2009) combinaram algoritmos hierárquicos e não-hierárquicos para melhorar a eficiência do reconhecimento de diferentes padrões de consumo no mesmo nível de tensão e destacaram a importância do estudo das curvas de cargas para traçar o perfil de consumo de energia elétrica dos consumidores. Dessa forma, destacaram que as revisões tarifárias do sistema elétrico brasileiro empregam teoria de amostragem juntamente com técnicas de análise de reconhecimento de padrões (ou tipificação) em curvas de carga. Neste contexto, propuseram metodologias alternativas para a seleção de curvas típicas e a classificação de dados, ambas adequadas às necessidades das revisões tarifárias. Com isso, mostraram que o método de seleção interfere pouco na quantidade de agrupamentos gerados, porém, é significativo na qualidade dos resultados. Nos métodos de classificação, o método "ponto a ponto" apresentou, na maioria dos casos, a menor quantidade de agrupamentos gerados, enquanto, as aproximações das curvas foram ligeiramente melhores para os processos "ponto a ponto" e "ponto a ponto ponderado". O "fator de carga por período" apresentou-se, quase sempre, no máximo, com o mesmo desempenho em relação aos outros métodos. Em linhas gerais, apesar da variação da metodologia do processo de classificação

não causar impacto significativo no resultado final, os métodos “ponto a ponto” e “ponto a ponto ponderado” foram os mais adequados aos objetivos.

Tsekouras et al. (2011) apresentaram a concepção e desenvolvimento de um banco de dados para o registro do comportamento de clientes de serviços públicos de energia elétrica. O banco de dados armazena informações sobre suas curvas de demanda, dados tarifários, dados das redes de energia e gás, bem como dados sobre a gestão da demanda e da racionalização do consumo de energia elétrica. O objetivo principal deste banco de dados é apoiar modelos de uso final e subsidiar os algoritmos de previsão de demanda no estudo do impacto de programas de eficiência energética e no apoio a decisões de natureza financeira.

Chicco et al. (2012) apresentaram uma revisão de métodos de reconhecimento de padrões utilizando a análise de agrupamentos de curvas de cargas. Uma síntese de métodos de reconhecimento de tipologias em curvas de carga (problema univariado, ou seja, sem a inclusão de variáveis adicionais tais como temperatura e umidade relativa do ar) é apresentada no Quadro 4.

Quadro 4 – Métodos de agrupamentos de séries temporais do setor elétrico.

MÉTODO DE AGRUPAMENTO	**AUTORES**
Quantificação vetorial adaptativa (*Adaptive Vector Quantization – AVQ*).	Tsekouras et al. (2007).
Entropia relativa de Renyi.	Chiccoet al. (2010).
Siga o líder (*Follow-the-Leader*).	Chicco et al. (2003), Chicco et al. (2004), Chicco et al.(2005), Yu et al. (2005), Carpaneto et al. (2005) e Chiccoet al. (2006).
Sistema híbrido de inferência *fuzzy* e autoregressivo *ARIMA*.	Nazarkoet al. (2005).
Método FCM.	Marques et al. (2004), Chicco et al. (2005), Tsekouras et al. (2007).
Agrupamento Hierárquico.	Chicco et al. (2005), Chicco et al. (2006), Tsekouras

	et al. (2007) e RAMOS et al. (2007).
Agrupamento de refinamento iterativo (*Iterative Refinement Clustering – IRC*).	Batrinuet al. (2005).
Método k-means.	Marques et al. (2004), Figueiredo et al. (2005), Chicco et al. (2005), Chicco et al. (2006) e Tsekouras et al. (2007).
Min-Max Neuro-Fuzzy - MMNF	Lamedica et al. (2000).
Rede neural probabilística (*Probabilistic Neural Network - PNN*)	Gerbec et al. (2005).
Mapa auto-organizáveis (*Self Organizing Map – SOM*).	Marques et al. (2004), Chiccoet al. (2004), Figueiredo et al. (2005), Chicco et al. (2005), Gerbec et al. (2005), Valero et al. (2007), Tsekouras et al. (2008) e Räsänen et al. (2010).
Agrupamento por Vetor de suporte (*Support Vector Clustering – SVC*).	Chiccoet al. (2009).
Agrupamento por Acúmulo de evidências ponderadas (*Weighted Evidence Accumulation Clustering – WEACS*).	Ramos et al. (2007).

Fonte: Chicco et al. (2012).

Com o intuito de subsidiar um melhor entendimento quanto a gama de métodos que podem ser utilizados no reconhecimento de tipologias (padrões) em curvas de carga por meio da análise de agrupamentos, em seguida será apresentado o detalhamento dos métodos apresentados no Quadro 4, sobretudo no que tange aos algoritmos de agrupamentos adotados.

Tsekouras et al. (2007) descreveram uma metodologia de dois estágios que foi desenvolvida para a classificação de consumidores de energia elétrica, e que se baseou principalmente nos métodos *k-means*, quantização vetorial adaptativa de Kohonen, *FCM* e

agrupamentos hierárquicos. No primeiro estágio, perfis típicos de curvas de cargas de vários consumidores são estimados usando métodos de reconhecimento de padrões e estes resultados são comparados usando seis medidas de validação de grupos (erro quadrático médio, índice médio de adequação, índice de dispersão de agrupamento, matriz de similaridade, índice de *Davies-Bouldin*, razão entre a soma dos erros quadrático dentro do grupo e soma dos erros quadrados entre os grupos). No segundo estágio as curvas de cargas sofrem uma reclassificação não hierárquica e a cada curva de carga é designada ao grupo com perfil de curva de carga (obtido no primeiro estágio) mais parecido. O resultado do primeiro estágio pode ser usado para a previsão da carga de consumidores e determinação de tarifas. O resultado do segundo estágio pode fornecer informações das características estatísticas dos consumidores associados a cada perfil reconhecido. O desenvolvimento do método é aplicado no conjunto de consumidores de média tensão.

Chicco et al. (2010) ilustraram e discutiram uma abordagem original para a classificação de consumidores de energia elétrica de acordo com o padrão de carga diária. Esta abordagem explora a noção da entropia introduzida por Renyi para um conjunto específico de procedimentos de agrupamentos. Diferentemente deste procedimento, o método adotado para a classificação de consumidores de energia elétrica utilizou a noção da distância Euclidiana. O algoritmo testado inclui primeiramente um método de classificação baseado na entropia entre os grupos e na minimização de suas variações. Então, um novo procedimento é apresentado, baseando-se no cálculo da similaridade entre centróides com sucessivos refinamentos para permitir a identificação efetiva de padrões anormais de carga diária. Os resultados da classificação realizada usando o procedimento proposto são comparados com os resultados de outras técnicas avaliadas, com base num conjunto de medidas de validação de grupos. Basicamente, o novo procedimento exibe uma melhor performance de agrupamento tanto em relação aos métodos anteriormente mencionados, como para o método clássico baseado em entropia para diferentes números de grupos. Os resultados obtidos foram promissores no sentido de ajudar na identificação de padrões de cargas de grupos de consumidores alvos de tarifas customizadas.

Chicco et al. (2003) associaram a classificação de consumidores de energia elétrica aos seus hábitos de consumo a partir de uma extensa base de dados de curvas de cargas diárias oriundas de medidas de campo dos clientes. No espoco do estudo foi proposto originalmente duas medidas de validação de grupos: um índice médio de adequação (*Mean Index Adequacy*

– *MIA*) e um indicador da dispersão do agrupamento (*Clustering Dispersion Indicator – CDI*). Estas medidas serviram de suporte para o desenvolvimento de um algoritmo que realiza automaticamente a análise de agrupamento. Cada classe de consumidor é então representada por um perfil de carga. O estudo favoreceu a melhoria do processo de tomada de decisão de empresa de grande porte de distribuição de energia, sobretudo no que diz respeito a definições das tarifas de serviços.

Chicco et al. (2004) destacaram que o conhecimento dos padrões de consumo dos clientes representa um ativo de valor para distribuidores de energia elétrica e que várias abordagens podem ser aplicadas para reconhecer grupos de consumidores com comportamentos similares nos seus hábitos de consumo. Duas abordagens para a classificação de consumidores são adotadas, quais sejam, os algoritmos *follow-the-leader* e mapas auto-organizáveis.

Em outro trabalho, Chicco et al. (2005) apresentaram abordagens de análise de agrupamento envolvendo métodos estatísticos e determinísticos e métodos baseados em inteligência computacional tais como sistemas de inferência *fuzzy* e redes neurais artificiais.

Gerbec et al. (2005) apresentaram uma metodologia para a reconhecimento de padrões de perfis de carga dos consumidores usando rede neural probabilística. Numa fase preliminar deste procedimento foi feito um pré-processamento dos perfis de cargas medidos utilizando análise *wavelet* de multiresolução e grupos iniciais foram obtidos por meio da aplicação do método de agrupamento *FCM*. A metodologia proposta foi testada com sucesso no caso de uma empresa de distribuição.

Yu et al. (2005) propuseram um método para o reconhecimento de padrões em curvas de cargas através de um sistema de leitura de medição automática. Um algoritmo de agrupamento foi desenvolvido usando os dados de perfis de cargas medidos por meio de um intervalo de integralização de 15 minutos.

Carpaneto et al. (2006) forneceram um arcabouço teórico matemático para a definição e investigação de dados no domínio da frequência. Resultados obtidos no conjunto de consumidores pertencentes a um sistema de distribuição são apresentados e discutidos. Estes resultados mostraram que a representação proposta é efetiva na redução da quantidade de dados armazenados e são capazes de manter um nível satisfatório de classificação.

Chicco et al. (2006) ilustraram e compararam os resultados obtidos por vários algoritmos de agrupamentos não supervionados (follow-the-leader modificado, agrupamento hierárquico, *k-means, FCM*) e o mapa autoorganizável para grupos de consumidores agregados com comportamentos similares de consumo. Outras técnicas de redução de dimensionalidade são discutidas e comparadas pelos autores tais como mapa de Sammon, análise de componentes principais e análise de componentes curvelineares. A efetividade da classificação obtida com o algoritmo testado foi comparado utilizando um conjunto de medidas de validação de grupos. Os resultados obtidos referem-se a um conjunto de consumidores não residenciais.

Nazarko et al. (2005) apresentaram possibilidades de aplicação de Modelagem ARIMA com abordagem de agrupamento *fuzzy* para estimativa da carga de energia elétrica. O agrupamento das subestações (clientes) de acordo com as curvas de carga diárias é feito com base em valores médios. Os resultados evidenciam a importância dos métodos para a estimativa de carga destinada a cada grupo de subestações.

Ramos et al. (2007) trataram a respeito de uma metodologia de caracterização de consumidores de média tensão. A caracterização é feita a partir de uma base de dados via processo de descoberta de conhecimento (*Knowledge Database Discovering – KDD*). Técnicas de mineração de dados são usadas com a proposta de obter perfis típicos de consumidores de média tensão visando conhecer os hábitos de consumos. A fim de formar classe de diferentes consumidores, e encontrar um conjunto representativo de padrão de consumidores, um algoritmo de agrupamento hierárquico e uma combinação de uma família de abordagem de agrupamentos por acúmulo de evidências ponderadas (*Weighted Evidence Accumulation Clustering – WEACS*) são usados. Para receitas de contas de perfis típicos de consumidores desta classe, novas opções de tarifas foram definidas e novos preços de energia foram propostos.

Batrinu et al. (2005) apresentaram o método de agrupamento de refinamento iterativo (*Iterative Refinement Clustering – IRC*), originalmente desenvolvido a fim de compensar as limitação em termos de predefinição do número de grupos verificada no método *follow-the-leader*. Os resultados foram baseados em uma amostra de duzentos consumidores não residenciais.

Marques et al. (2004) utilizaram seis métodos de análise de agrupamentos para classificar o perfil de carga de um barramento de subestação de energia elétrica: *k-means*, quatro variações de mapas auto-organizáveis e *FCM*. Várias simulações com diferentes parâmetros foram usadas para melhorar a performance do agrupamento. Medidas de validação de grupos foram utilizadas para comparar os métodos adotados.

Figueiredo et al. (2005) apresentaram uma caracterização dos consumidores de energia elétrica utilizando técnicas de mineração de dados (*KDD*). O núcleo desse procedimento combina métodos supervisionados e não-supervisionados de aprendizagem de agrupamento de curvas de cargas. O estudo de caso é feito em duas etapas usando dados de uma empresa de distribuição de energia elétrica de Portugal.

Valero et al. (2007) apresentaram métodos de classificação avaliando diferentes metodologias de agrupamentos. O objetivo é investigar a capacidade dos mapas auto-organizáveis para classificar consumidores de energia elétrica e suas potenciais respostas para o sistema de distribuição e comercialização através da modelagem das curvas de cargas diárias. A demanda e a resposta do consumidor foram avaliadas em relação ao comportamento dos preços dos serviços. O resultado mostrou a potencialidade desta abordagem para a melhoria do gerenciamento dos dados e seleção de políticas coerentes com base numa previsão confiável de diferentes cenários de preços.

Tsekouras et al. (2008) descreveram uma metodologia de reconhecimento de padrões para a classificação das curvas de cargas diárias de grandes consumidores de energia elétrica a fim de estimar os perfis de cargas diários representativos. A abordagem baseia-se em métodos de reconhecimento de padrões, tais como, *k-means*, mapas auto-organizáveis, *FCM* e agrupamento hierárquico. Os parâmetros de cada método de agrupamento são adequadamente selecionados por um processo de otimização e seis medidas de validação de grupos são aplicadas para verificar seus desempenhos. Os resultados podem ser usados para a caracterização de demanda de curto prazo e previsão de curvas de cargas dos consumidores visando a especificação de tarifas adequadas e o estudos de viabilidade dos programas de gestão da demanda. Esta metodologia foi aplicada analogamente para clientes industriais e residências de média tensão na Grécia.

Rãsânen et al. (2010) destacaram que os recentes desenvolvimentos tecnológicos que monitoram o uso de energia elétrica de pequenos clientes proporcionam uma nova visão para

aumento da eficiência energética e operação de serviços específicos para os clientes de sistemas de distribuição de energia elétrica. Os autores apresentaram métodos de agrupamentos baseados em mapas auto-organizáveis, *k-means* e algoritmos hierárquicos, capazes de processar grande quantidade de dados de séries temporais no contexto da pesquisa de gerenciamento de cargas. Os métodos foram aplicados numa base de dados consistindo na medição de dados de consumo de energia elétrica de pequenos consumidores do norte da Finlândia.

Chicco et al. (2009) apresentaram uma original e eficiente aplicação do método de agrupamento baseado em vetor de suporte (*Support Vector Clustering – SVC*) para a classificação de padrões de cargas de energia elétrica. A proposta do método *SVC* combina o cálculo do vetor de suporte usando um procedimento clássico que adota um núcleo gaussiano com um algoritmo determinístico, especificamente desenvolvido para formar os agrupamentos. Este algoritmo explora a localização significativa dos vetores de suporte delimitados (*Bounded Support Vectors - BSVs*) para definir os valores aberrantes, identificando os agrupamentos em função da distância dos não-*BSVs* em relação aos *BSVs*. A comparação com outros métodos de agrupamento evidencia a eficácia da abordagem proposta para reconhecimento de grupos não sobrepostos. Esta eficácia foi confirmada pelo cálculo de várias medidas de validação de grupos.

Em todos os trabalhos citados o número de grupos (número de padrões) foi previamente definido. Trabalhos recentes sugerem que métodos baseados em inteligência artificial proporcionam uma melhor qualidade no agrupamento de curvas de carga, pois os dados passam inicialmente por uma etapa de aprendizagem, em que um conjunto de exemplos é apresentado à rede, a qual extrai automaticamente as características necessárias para representar a informação fornecida (MONEDERO et al., 2006; NIZAR et al., 2006; NAGI et al., 2008; NIZAR et al., 2008; SILVA et al, 2011). Uma crítica a este modelo é que seus parâmetros internos, em geral, não são conhecidos, e por isso, é denominado de modelo caixa-preta. Em outros termos, este modelo dificulta obter informações sobre o comportamento do sistema durante o processo de reconhecimento de padrões.

Monedero et al. (2006) descreveram um protótipo para a detecção de perdas não-técnicas, por meio de dois métodos de agrupamentos de dados: redes neurais e algoritmos com métricas estatísticas de similaridades. A perda não-técnica foi definida como qualquer energia ou serviço consumido que não é cobrado por causa de falha do equipamento de

medição ou uso fraudulento dos referidos equipamentos. A detecção de perdas não-técnicas (que inclui a detecção de fraude) é um campo onde a mineração de dados tem sido aplicada com êxito.

Nizar et al (2006) aplicaram métodos de caracterização de objetos para a obtenção de melhores padrões de demanda de carga em um sistema de distribuição usando *KDD*. O objetivo do estudo foi detectar e prever perdas não-técnicas no setor da distribuição associados a erros de medição e falhas administrativas, e extrair conhecimento sobre o comportamento do cliente e suas preferências. Com base no conjunto de dados disponíveis, os clientes foram classificados de acordo com a hora do dia, fatores de carga e as características de consumo.

Nagi et al. (2008) apresentaram uma nova abordagem para análise de perdas não-técnicas (*Non-Technical Loss - NTL*) de dados de empresas de energia elétrica, utilizando uma nova técnica baseada em inteligência de máquinas de vetores de suportes (*Support Vector Machine - SVM*). A principal motivação deste estudo foi auxiliar a empresa *Tenaga Nasional Berhad* (*TNB*) na Malásia para reduzir suas perdas não-técnicas no setor de distribuição devido às ocorrências de fraudes constatadas. O modelo proposto pré-seleciona os clientes suspeitos a serem inspecionados no local por fraude com base em irregularidades e comportamento de consumo anormal. Esta abordagem baseia-se também em KDD e envolve extração de características a partir de dados históricos de consumo dos clientes. A abordagem baseada em inteligência de máquinas de vetores de suportes usa informações do perfil de carga do cliente para expor comportamento anormal que é conhecido por ser altamente correlacionado com atividades de perdas não-técnicas. Os resultados mostram que o método proposto é mais eficaz em comparação com ações atuais tomadas pela empresa *Tenaga Nasional Berhad* no que tange ao combate das perdas não-técnicas.

Nizar et al. (2008) apresentaram uma nova abordagem para análise de perdas não-técnicas (*Non-Technical Loss - NTL*) para consumidores de energia elétrica usando uma técnica computacional a aprendizagem extrema de máquinas (*Extreme Learning Machine ELM*). A abordagem usa informações de perfis de cargas dos clientes para detectar comportamentos anormais que são conhecidos por serem altamente correlacionados com a *NTL.* Esta abordagem proporciona um método de extração de padrões a partir de dados históricos de consumo que incluem intervalos de 48 dias durante dois anos. Três algoritmos

foram utilizados para a classificação dos comportamentos de clientes, quais sejam, *ELM*, *Online Sequencial* (*OS) - ELM* e *SVM*.

Silva et al. (2011) apresentaram uma nova estrutura de mineração de dados para a exploração e extração de conhecimentos gerados a partir de dados de medições de energia elétrica. Os autores afirmam que as abordagens convencionais não são capazes de resolver por completo as possibilidades de extração de conhecimento dos fluxos contínuos de dados de medidores de consumo de energia elétrica e que para superar esses problemas é importante que as técnicas de mineração de dados incorporem funcionalidade para sumarização e análise incremental utilizando técnicas Inteligentes. Assim, foi proposta uma estrutura de caracterização incremental de padrões (*Incremental Summarization and Pattern Charaterization – ISPC*) para atender estas expectativas. Independentemente, outro algoritmo de caracterização incremental de padrões (*Incrementally Characterizes Patterns – IPCL*) foi também aplicado para as devidas comparações.

Alguns trabalhos compararam o desempenho de diferentes métodos de reconhecimento de padrões de curvas de cargas, e concluíram que o algoritmo *FCM* (BEZDEK et al., 2005) proporciona um melhor nível de coesão e de discriminação dos problemas associados com o agrupamento de curvas de cargas visando a tipificação da demanda (GERBEC et al., 2005; ZAKARIA et al., 2006; ANNUAR e ZAKARIA, 2010). Uma vez que o algoritmo *FCM* projeta para cada objeto o grau de pertinência para todos os grupos reconhecidos, favorecendo a uma melhor discriminação de grupos com uma maior qualidade de coesão e separação.

Annuar e Zakaria (2010) aplicaram o método *FCM* em conjunto com índices de validação de grupos (*Xie -Beni* e *Davies- Bouldin*) para seleção do melhor número de grupos. Os perfis de carga usados foram obtidos através da *Tenaga Nasional Berhad* (*TNB*), uma empresa de distribuição da Malásia.

3.3 MARCOS TEÓRICOS PARA ANÁLISE DE DADOS MULTIVARIADOS

No âmbito dos programas de eficiência energética, em particular os relacionados a troca de refrigeradores em comunidades de baixa renda (capítulo 3), pode-se avaliar, antes e depois da substituição de refrigeradores (economia de energia), o cumprimento de metas de eficiência energética através da análise conjunta do comportamento do consumo de energia (curva de carga) com outras séries temporais correlatas. Em relação a esse caráter

multivariado das séries temporais observadas, no universo de variáveis do setor elétrico, ressalta-se que, na gestão da demanda no sistema elétrico, as variáveis meteorológicas se destacam pela sua significativa influência nos padrões de consumo de eletricidade (LIN et al, 2006; CHICCO et al, 2012).

Lin et al. (2006) mostraram que os principais fatores climáticos como temperatura e umidade relativa do ar são causas importantes de influência na tendência ou variações dos perfis da carga do sistema, sendo que o primeiro fator (temperatura) possui maior impacto. Lin et al. (2006) usaram um banco de dados no período de 10 anos referente à empresa de energia a China Light and Power (CLP) e analisaram o efeito da temperatura e da umidade relativa do ar no pico de carga. Através do método k-means os autores realizaram análise de agrupamentos possibilitando um melhor reconhecimento de padrões de regime de operações dos transformadores.

A fim de melhorar a qualidade da avaliação final destes programas de eficiência energética, é conveniente que a curva de carga seja analisada em conjunto com outras séries temporais, aumentando o nível de conhecimento sobre o comportamento da demanda no sistema elétrico (TSEKOURAS et al., 2011). A temperatura ambiente tem um forte efeito sobre a eficiência térmica do equipamento, especialmente no que diz respeito ao modo de uso (frequência de abrir a porta do refrigerador), o que contribui para os desvios da temperatura interna em relação à temperatura de acondicionamento programada (CLINCH e HEALY, 2001). Particularmente, quanto aos refrigeradores, a eficiência energética também pode ser quantificada por meio do inverso do Coeficiente de Performance de Carnot (COP- 1) que expressa a relação entre o trabalho de arrefecimento e o calor absorvido da fonte fria (STOECKER, 1998).

Clinch e Healy (2001) destacam que consumidores residenciais têm se revelado como uma classe com significativo potencial de contribuição para eficiência energética global. Entretanto, antes da implementação de programas de energia em grande escala é importante avaliar se estes consumidores têm sustentabilidade econômica. Os autores forneceram um modelo para avaliações econômicas de eficiência energética a partir de um estudo de caso de um programa irlandês de substituição de equipamentos que possuem impactos significativos na conta de energia dos consumidores residenciais. O estudo demonstrou que a economia de energia contribuiu para a melhoria de condições ambientais e sociais e os resultados foram

suficientemente convincentes para mostrar os benefícios que podem ser obtidos com a implementação de programas de eficiência energética.

A fim de auxiliar a avaliação da eficácia do programa de eficiência energética de forma multivariada, em geral é feito o reconhecimento de padrões do consumo de energia elétrica com a inclusão de perfis adicionais a serem considerados de forma integrada (problema protótipo não pontual), onde cada objeto é representado por séries temporais multivariadas, favorecendo a uma abordagem integrada da análise de agrupamento de objetos, visando a compreensão dos possíveis comportamentos ou padrões que devem ser reconhecidos conjuntamente.

As publicações de artigos sobre a análise de séries temporais multivariadas é menos frequente do que as publicações de artigos de séries temporais univariadas, sobretudo, no setor elétrico abordando o reconhecimento de padrões via agrupamentos de curvas de cargas conjuntamente com séries temporais correlatas. Um indicativo importante disto é a baixa ocorrência de publicações numa biblioteca virtual nacional que reúne e disponibiliza a instituições de ensino e pesquisa no Brasil mais de 35 mil periódicos internacionais e nacionais com texto completo, 130 bases referenciais, 11 bases dedicadas exclusivamente a patentes, além de livros, enciclopédias e obras de referência, normas técnicas, estatísticas e conteúdo audiovisual. Até o presente momento, tomando-se como referência palavras-chave similares a "*Clustering multivariate time series*" em associação a "*electric sector*", apenas uma ocorrência de artigo publicado foi constatado (FERREIRA et al., 2015). Este artigo tem como um dos autores o próprio autor desta tese, e serviu de base para o desenvolvimento do Capítulo 6 deste trabalho.

3.4 CONTRIBUIÇÕES DO MÉTODO PROPOSTO

Assim, o ineditismo deste trabalho está em desenvolver e validar um método de reconhecimento de padrões de curvas de carga, baseado em técnicas de agrupamento de séries temporais, que incorpora diferentes métricas estatísticas para a formação dos grupos, tanto no caso univariado quanto no caso multivariado. Devido a sua característica semi-hierárquica, o método proposto é capaz de estabelecer um número ótimo de grupos (balizando-se pelo índice de silhueta) de séries temporais a serem reconhecidos, sobretudo, com o período horário de amostragem que pode revelar indícios de hábitos de consumo da energia elétrica.

O algoritmo *FCM* é utilizado como referência comparativa tanto para a versão univariada quanto para a versão multivariada, sendo que, para a versão multivariada, é adotada uma versão modificada do *FCM* baseada em uma métrica de similaridade que utiliza componentes principais (*Similarity Principal Componente Analysis – SPCA*) (SINGHAL e SEBORG, 2005).

4 HIPÓTESES, OBJETIVOS E METODOLOGIA

4.1 HIPÓTESES

4.1.1 Hipótese primária

A aplicação de um método de agrupamento de séries temporais, para reconhecer padrões de demanda de energia do refrigerador, favorece a uma identificação de oportunidades para conservação da energia mais coerente com a realidade dos hábitos de consumo, que são importantes norteadores para a gestão da demanda do sistema elétrico.

4.1.2 Hipótese secundária

O comportamento da curva de carga de energia do refrigerador apresenta uma sazonalidade horária que tem como principais causas o modo de uso e as condições do equipamento.

4.2 OBJETIVOS GERAL E ESPECÍFICOS

O objetivo geral deste trabalho é desenvolver um método (que denominamos de Sistema de Tipificação e Agrupamento de Curvas de cargas – STAC) para reconhecer padrões de séries temporais. Utilizando como estudo de caso as medições do consumo de refrigeradores e suas temperaturas interna e externa, realizadas no âmbito de programa de eficiência energética de troca de refrigeradores. Além disso, tomar como referência comparativa o método Fuzzy C-Means (FCM), já consolidado na literatura.

Especificamente pretende-se:

- Propor um método constituído de dois algoritmos: um aplicável a séries temporais univariadas e outro aplicável a séries temporais multivariadas, ambos desenvolvidos para o reconhecimento de padrões de séries temporais com o mesmo número de pontos amostrados e para o mesmo período de observação;

- Incorporar no novo método um procedimento de agrupamento que defina automaticamente o número ótimo de grupos de séries temporais, com capacidade de reconhecer curvas típicas que represente significativas diversidades sazonais;

- Tomar como estudos de casos os programas de eficiência energética envolvendo troca de refrigeradores em unidades de baixa renda em diferentes Estados do Brasil, e apresentar um novo método para o reconhecimento de padrões de hábito de consumo diário de energia elétrica capaz de tratar as curvas de consumo (curvas de carga) de forma isolada (série temporal univariada) ou em conjunto com séries temporais relacionadas a outras variáveis (séries multivariadas) de impacto sobre o sistema de distribuição.

- Utilizar o novo método (em ambas as suas versões, uni e multivariada) como ferramenta de apoio na avaliação do impacto de programas de eficiência energética por substituição de equipamentos (refrigeradores), viabilizando o reconhecimento de mudanças nos padrões de consumo decorrentes da implementação do programa.

4.3 METODOLOGIA

4.3.1 Pressupostos

Quanto aos pressupostos metodológicos o método de abordagem da presente pesquisa é o hipotético-dedutivo, ou seja, os objetivos da pesquisa foram norteados pelas hipóteses adotadas para a investigação do problema. Além disso, é de natureza aplicada com uma abordagem quantitativa, pois, através da analise de agrupamento das séries temporais relacionadas aos refrigeradores, busca-se reconhecer, sobretudo, padrões de demanda que possam revelar indícios de hábitos de uso destes equipamentos. Por conseguinte, quanto aos objetivos a pesquisa é também explicativa. No geral, diferentes procedimentos foram adotados para o adensamento teórico e coleta de dados, tais como, pesquisa documental, pesquisa bibliográfica, pesquisa de levantamento e estudo de caso.

As etapas para realização da pesquisa foram: estabelecimento da questão inicial, exploração do tema, problematização, coleta de dados, tratamento e análise dos dados, construção de modelos e constatação final. A questão inicial foi apresentada na introdução, quando foi descrito o processo decisório do setor elétrico com o foco no programa de eficiência energética implementado pelas distribuidoras de energia, incentivado pelo governo brasileiro através da ANEEL, cuja experiência de duas distribuidoras serviu de estudos de casos da tese. A exploração do tema foi apresentada no embasamento teórico da análise de agrupamentos, que usou fontes bibliográficas de livros e artigos na discussão da teoria que

norteou a caracterização, medição, algoritmo e validação do novo método proposto. Na seção sobre revisão bibliográfica também foi feita a exploração do tema, com maior predominância em artigos publicados em periódicos de grande relevância acadêmica, com o intuito de indicar a contribuição do novo método proposto para o estado da arte. A problematização e o planejamento da coleta de dados serão apresentados na próxima seção através da discussão das hipóteses, objetivos, pressupostos, amostragem e materiais e métodos. Por fim, nos capítulos que se seguem apresentam o novo método proposto e conclusão.

Conforme já mencionado, o novo método de reconhecimento de padrões de séries temporais do consumo horário de energia elétrica foi concebido em duas versões: na primeira foi desenvolvido um método de reconhecimento de padrões de curvas de carga isoladamente. A segunda versão compreende o reconhecimento de padrões de curvas de cargas juntamente com outras séries temporais associadas ao consumo de energia elétrica. A implementação computacional do novo método foi feita por meio de rotinas desenvolvidas no *software MATLAB*, sobretudo, o seu *TOOLBOX* para operações de álgebra linear e cálculos estatísticos.

Os dados que constituíram a amostra de trabalho neste estudo estão relacionados aos programas de eficiência energética implementados por duas empresas distribuidoras, quais sejam, Companhia Energética do Maranhão (CEMAR) e Companhia Energética de Alagoas (CEAL). Os programas de eficiência energética compreenderam, em ambos os casos, a substituição de refrigeradores nas residências de consumidores de baixa renda distribuídos em municípios dos respectivos Estados abrangidos pelas CEMAR e CEAL. Os dados foram coletados diretamente dos refrigeradores das unidades consumidoras antes (caso I) e depois (caso II) da substituição dos equipamentos. Refrigeradores novos foram doados às unidades consumidoras em substituição aos equipamentos antigos com menores recursos tecnológicos e prazos de vida útil esgotados.

Quatro municípios do estado do Maranhão, na área de abrangência da CEMAR, foram considerados e o Programa foi implementado durante o período de novembro de 2008 a julho de 2009. Este programa envolveu a substituição de 5250 refrigeradores velhos por novos. Uma amostra de 80 curvas de carga (refrigeradores velhos), apresentando um alto consumo de energia elétrica (caso I, o consumo médio de 82 kWh) e outra amostra de 80 curvas de carga, após a troca de refrigeradores (caso II, o consumo médio de 52 kWh) estavam disponíveis. O tamanho da amostra representou um nível de erro amostral de 10% e um nível de confiança de 95% no parâmetro de predição da população. O Protocolo Internacional de Medição e

Verificação de Desempenho (PIMVP) ou *International Protocol for Measurement and Verification of Performance - IPMVP* (EVO, 2007) recomenda um erro de amostragem de até aproximadamente 10%.

No segundo estudo de caso, foram considerados 5 (cinco) municípios do estado de Alagoas, na área de abrangência da CEAL, entre março de 2012 e julho de 2012 e o Programa envolveu a substituição de 5000 refrigeradores velhos por novos. Uma amostra de 54 curvas de carga, respectivamente, associadas às temperaturas externas e internas dos refrigeradores velhos com alto consumo de energia elétrica (caso I, média de consumo de 35 kWh) e outra amostra de 54 curvas de carga, respectivamente, associadas às temperaturas externas e internas dos refrigeradores novos (caso II, média de consumo de 18 kWh) estavam disponíveis. O tamanho da amostra representou também um nível do erro amostral da média de cerca de 10% de variação, mas o nível de confiança foi de 90% no parâmetro de predição da população.

Todos os dados estão relacionados com os dias úteis e o período considerado compreende quase todo o verão (estação com maior consumo de energia). Estas duas características contribuíram para reduzir os efeitos da sazonalidade nos dados. A escolha do dia útil da semana foi arbitrária, devido o fato deles terem comportamentos no consumo de energia mais parecidos entre si, e menos parecidos com o comportamento no consumo de energia nos dias de sábado e domingo. A disponibilidade de equipamentos de medição, logística operacional e restrições financeiras foram fatores determinantes neste caso.

Assim, para quantificar a energia elétrica economizada e a redução de demanda do processo de substituição de refrigeradores, foram realizadas medições em uma amostra estabelecida conforme o plano de amostragem definido pela norma NBR 5426, com regime de inspeção severa, Nível I conforme recomenda o PIMVP.

4.3.2 Amostragem

A amostra de trabalho foi constituída considerando-se a combinação de diferentes planos de amostragem, quais sejam, a amostragem por conglomerado e amostragem sistemática. Os municípios dos Estados foram agrupados tomando como referência os seus perfis de consumo (curvas de carga). Um município (centro ou protótipo) foi selecionado para representar cada grupo. As unidades de amostragem (refrigeradores) foram aleatoriamente selecionadas nos respectivos municípios. Estes planos de amostragem favoreceram a

minimização dos custos com a medição e verificação da variação de consumo dos equipamentos, dispersão geográfica entre os municípios-alvo do projeto e distribuição temporal das trocas dos refrigeradores velhos por novos.

Definidos os municípios-protótipos da medição e verificação, o sorteio das Unidades Consumidoras (UC's) para medição adotou o seguinte critério de amostragem aleatória sistemática para abranger todo o universo de forma distribuída:

- Ordenação e enumeração dos clientes do universo avaliado (clientes beneficiados do município) em ordem crescente de consumo;
- Sorteio aleatório de um número entre 1 e *K*, sendo *K* a razão entre a quantidade de elementos no universo avaliado e a quantidade de elementos da amostra;
- Seleção do elemento de número *K* como o primeiro a ser medido. Os demais foram selecionados a partir do incremento do valor de *K*.

Para a representatividade quantitativa, o tamanho da amostra foi dimensionado a partir do modelo para estimativa da proporção de sucesso para grandes populações. O processo de amostragem das unidades consumidoras foi balizado pelas conceituações preconizadas na teoria de amostragem, conforme recomenda o PIMVP (EVO, 2007). O erro de amostragem e nível de confiança foram baseados nas especificações adotadas pelo setor elétrico brasileiro, segundo as recomendações da Agência Nacional de Energia Elétrica (ANEEL). Esta agência estabelece diretrizes técnicas normativas para equipamentos refrigeradores conforme preconiza as principais normais internacionais, tais como, o PIMVP (ASHRAE, 2002; EVO, 2007).

Para a coleta de dados dos refrigeradores, foram realizadas medições utilizando registrador de grandezas elétricas e Termohigrômetro:

- Potência de cada refrigerador;
- Temperatura interna da geladeira, compartimento refrigerador;
- Temperatura do ambiente externo.

4.3.3 Materiais e Métodos

As medições que são feitas no âmbito de um projeto de eficiência energética, para determinar as reais reduções de consumo e demanda, devem ter seus resultados avaliados. O processo de medição da eficiência energética possui uma particularidade que merecem ser mencionada, qual seja não poder medir diretamente a eficiência energética, em consequência de ser medida após a implantação das ações de economia de energia. Em outros termos, torna-se imperativo fazer medições antes e após a ação de eficiência, e abstrair um modelo matemático sobre o comportamento de variáveis que impactam no consumo de energia. As ações adotadas, no presente estudo, tiveram como alvo as medições elétricas (consumo e demanda) dos refrigeradores e temperatura interna do equipamento.

Dentre as opções indicadas pelo Protocolo Internacional para Medição e Verificação de Performance (PIMVP) as opções A e B inserem-se no caso de substituições de equipamentos (Retrofit isolado). Na Opção A alguns parâmetros do uso de energia podem ser estimados e a análise de sua importância deve ser incluída no Plano de M&V.

No caso do projeto de eficiência energética em unidades consumidoras de baixa renda incluído nos programas de eficiência energética do presente estudo, todos os parâmetros podem ser medidos diretamente, sem a necessidade de estimativas, o que o torna enquadrável na Opção B.

Em outras palavras, podem ser enquadrados na Opção B programas de eficiência energética cujas ações (substituição de refrigeradores) não apresentam impactos identificáveis em outros usos de energia, o que torna possível a adoção de medições diretas e limitadas aos equipamentos que serão substituídos.

A coleta de dados teve como alvo as medições das grandezas da potência elétrica e temperatura associadas aos refrigeradores, e respectivamente, os sistemas de medição (Kit) SAGA 2000 e Termohigrômetro Extech RHT 10 foram utilizados nos períodos antes e após substituição.

As medições elétricas feitas através do Kit composto por um medidor SAGA 2000 (Figura 2) tem a instalação descrita no diagrama da Figura 3 e a coleta de dados é realizada no momento da desinstalação do medidor, com auxílio de um notebook devidamente configurado com o software do instrumento (PLAWIN 1000).

Figura 2 - Kit do medidor Saga 2000.

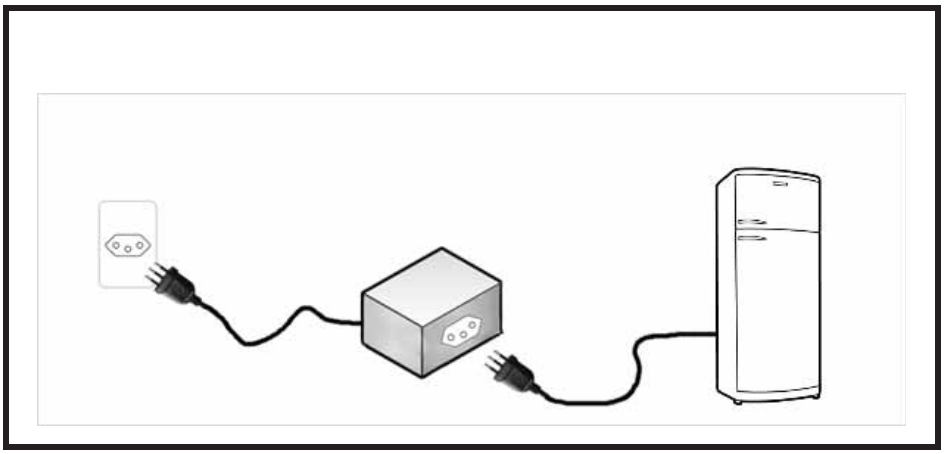


Figura 3 - Diagrama de Instalação do Kit de Medição Saga 2000.

Este Kit de medição deve ser instalado para medir o consumo do refrigerador, segundo descrição abaixo:

- Instalação: Conectar a tomada do equipamento ao Kit e a tomada do Kit à tomada do ambiente;
- Local da Instalação: Local seguro de modo a evitar acidentes;
- Duração da medição: mínimo de 03 dias corridos.

Após a instalação do kit, deve-se:

- Verificar se está medindo corretamente pressionando a tecla do mostrador e verificando se a luz vermelha (ao lado do mostrador) está acesa piscando;
- Conferir dados com a tabela de códigos afixada no Kit;

- Passar o selo (trava) no local específico para impedir a abertura do equipamento.

As medições elétricas feitas através do Kit composto por registrador de temperatura Termohigrômetro Extech RHT 10 (Figura 4). Deve ser instalado segundo a descrição abaixo:

- Instalação: Após parametrizado com tempo de integralização de 5 min e encapsulado/vedado em saco ou caixa plástica, de forma a isolar o equipamento do contato, mesmo que indireto com os alimentos; instalar na 1ª prateleira, próximo à parede interna posterior (não instalar dentro do congelador) – OBSERVAÇÃO: durante a rodada de medições, apenas um Termohigrômetro deve ser instalado em ambiente externo;
- Local da Instalação: Local seguro de modo a evitar acidentes;
- Duração da medição: 03 dias corridos, simultaneamente à medição de potência.

Figura 4 - Termohigrômetro Extech RHT 10.

Após a instalação do Termohigrômetro, deve-se verificar se está medindo corretamente verificando se a luz verde está acesa piscando.

As seguintes recomendações devem ser seguidas:

- O medidor não deve ser retirado da tomada para não interromper a medição;
- A rotina da família não será alterada;
- A desinstalação dos equipamentos só deve ser feita por técnicos autorizados.

Depois do período de leitura, devem-se coletar os dados do kit registrador de grandezas elétricas (chamada de memória de massa) de modo a validar a medição, seguindo os seguintes passos:

- Conectar o cabo óptico ao medidor (na porta óptica) e ao notebook;
- Abrir o programa PLAWIN 1000 no notebook (o atalho se encontra no desktop);

- Na barra de ferramentas, seguir a seguinte sequência: “Leitura de Dados” → “Composta” → “Verificação”;
- Após a leitura do arquivo, anotar o nome do arquivo informado pelo programa (na “Ficha Cadastro de Medição” e na “Planilha de medição). Por exemplo, “3204&GC.REN”;
- Fechar o programa PLA WIN;

Da mesma forma que o Kit registrador, deve-se coletar também a memória de massa do Termohigrômetro, da seguinte forma:

- Inserir o Termohigrômetro na porta USB do computador;
- Abrir o programa correspondente (RHT 10 ou outro);
- Clicar em “Connect”;

Uma ação importante para a confiabilidade dos dados é analisar a memória de massa, verificando se há, por exemplo, extensos períodos com valores zerados (indicativo de medição incorreta ou equipamento desligado). Nestes casos, confirmar com o usuário se o equipamento que foi medido foi desligado durante o período. Não tendo sido desligado, o indicativo é de que houve problema na medição, devendo haver reinstalação dos medidores (as medições de potência e temperaturas – interna e externa - só valem se forem simultâneas).

Depois de concluída análise da memória de massa, retirar o kit, da seguinte forma:

- Desligar o kit desde a tomada;
- Desligar o refrigerador desde o Kit;
- Ligar o refrigerador à tomada;
- Verificar se o refrigerador está funcionando adequadamente.

O Termohigrômetro deve ser retirado somente após descarregar memória de massa do kit registrador de grandezas elétricas. O download da memória de massa do Termohigrômetro deve ser feito ainda na unidade consumidora, de forma a consistir a medição, vincular à pasta específica, nomeada com o número do equipamento correspondente, evitando equívocos posteriores.

5 RECONHECIMENTO DE PADRÕES EM CURVAS DE CONSUMO DO SETOR ELÉTRICO – CASO UNIVARIADO

Este capítulo apresenta um novo método para Seleção, Tipificação e Agrupamento de Curvas de cargas – STAC (*selection, typification, and load curve clustering - STCL*) baseado em uma extração sistemática de características inerentes ao setor elétrico. A versão inicialmente apresentada trata o problema univariado (apenas curvas de consumo) e, no capítulo seguinte, o método é estendido para a situação multivariada (curvas de consumo juntamente com outras séries temporais). O estudo de caso analisado é um programa de eficiência energética realizado pela Companhia Elétrica do Maranhão (Brasil), que considerou, entre outros, a análise do impacto da substituição de refrigeradores nas casas dos consumidores de baixa renda, distribuídos em várias cidades localizadas no interior do Estado do Maranhão (Brasil). O método proposto incorpora múltiplos critérios no agrupamento e tipificação de curvas de cargas, ao contrário das abordagens tradicionais que essencialmente usam um critério central de distância entre curvas de carga para o reconhecimento de grupo. Este item 3.2 apresenta o método STAC. O item 3.3 apresenta uma comparação entre os resultados obtidos pelo método STAC e *FCM* tomando como base o estudo de caso proposto.

5.1 O MÉTODO STAC

O STAC é um método que envolve a seleção, tipificação e agrupamento de curvas de carga do setor elétrico (FERREIRA et al., 2011; FERREIRA et al., 2012; FERREIRA et al., 2013a; FERREIRA et al., 2013b) baseado numa extração sistemática de características. O método proposto incorpora vários critérios no agrupamento e tipificação das curvas de carga, ao contrário das abordagens tradicionais que essencialmente usam o critério de distância entre curvas de carga para o reconhecimento de grupos. As Figuras 5-6 apresentam o esquema geral do método STAC que compreende duas fases. Na primeira o reconhecimento de padrões é realizado através de sucessivas iterações. A primeira iteração realiza o agrupamento de toda a amostra com base em características específicas associadas com o perfil de consumo e alguns grupos de curvas de carga são reconhecidos. As iterações subsequentes consideram apenas as medianas (padrões) de cada grupo gerado na primeira iteração e verificam a semelhança entre estas medianas com base nos mesmos testes estatísticos considerados na primeira iteração. Esta verificação de semelhanças permite que alguns grupos sejam unidos. Assim, no final da primeira fase (depois da convergência), padrões ou tipos associados com curvas de carga são

reconhecidos. A segunda fase define os grupos finais associando cada curva de carga (base de dados) para um dos padrões reconhecidos na primeira fase.

Inicialmente, cada curva é normalizada dentro do intervalo [0;1] dividindo as medições horárias pelo pico de demanda de cada um. O consumo adimensional quantificado desta forma é denominado de por unidade (*pu*) (ANUAR e ZAKARIA, 2010).

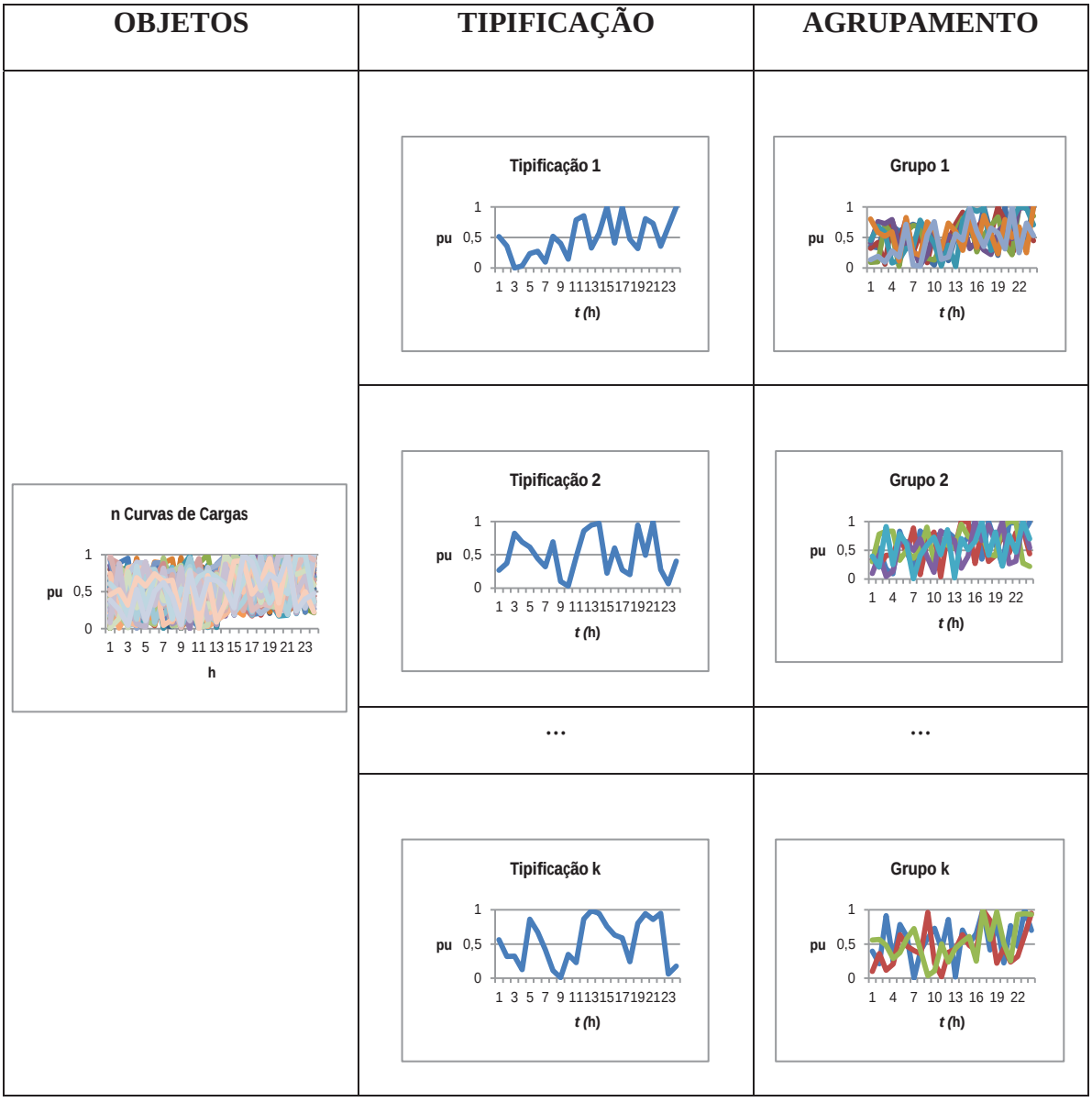


Figura 5 – Esquema geral do método STAC.

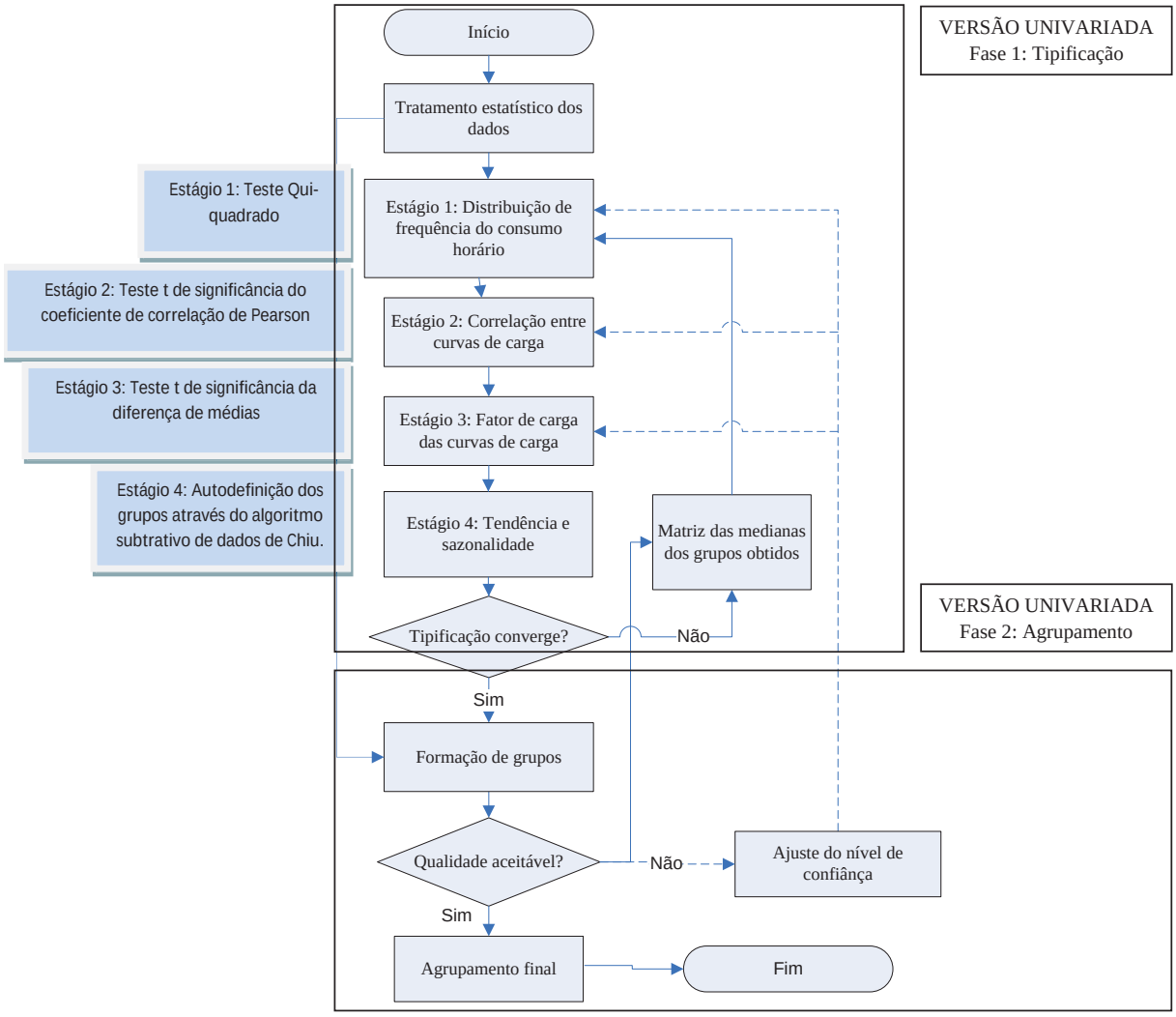


Figura 6 – Fluxograma do algoritmo do Método STAC - Versão univariada.

A Figura 6 e o Quadro 5 apresentam os critérios considerados na análise de similaridade realizada pelo algoritmo do método STAC até terceiro estágio da primeira fase, juntamente com o teste estatístico aplicado (SEPPÃLÃ, 1995; JANES, 2001; O'GORMAN, 1997). Esses critérios foram estabelecidos de acordo com os requisitos e indicadores praticados no setor elétrico de distribuição de energia (LIN et al., 2006).

Quadro 5 – Testes estatísticos - método STAC.

CRITÉRIOS DE AGRUPAMENTOS	TESTES ESTATÍSTICOS
Distribuição de frequência da demanda horária.	Estágio 1: Estatística do Qui-quadrado (SEPPÃLÃ, 1995; JANES, 2001).
Seleção das curvas de cargas por grau de correlação.	Estágio 2: Estatística *t* para o teste de significância do grau de correlação de Pearson entre duas amostras independentes (O'GORMAN, 1997).
Seleção do Fator de carga.	Estágio 3: Estatística *t* para o teste de significância da diferença de médias entre duas amostra independentes (O'GORMAN, 1997).

As três características apresentadas no Quadro 5 são aplicadas sucessivamente. Na primeira iteração, os grupos são formados com base na similaridade entre as curvas de carga e a curva com o maior consumo médio de energia (curva de referência). Após a primeira iteração, o método utiliza a mediana de cada grupo (considerada curva padrão), uma medida central mais resistente do que a média em relação aos valores atípicos que possam estar presentes nos seus grupos de origens, para caracterizar os perfis dos padrões (BOX et al., 2008). Desta forma, os mesmos testes são aplicados sucessivamente, considerando as medianas (padrões). A existência de uma similaridade entre medianas de acordo com os testes estatísticos aplicados implica na união de grupos e novas medianas (padrões) resultantes são obtidas.

Nos três primeiros estágios, os testes de hipóteses estatísticas entre as curvas de cargas são realizados de acordo com o Quadro 5. Existe um quarto estágio que adota uma métrica multivariada de dissimilaridade entre as curvas de carga em relação ao seu comportamento horosazonal. Cada grupo gerado no terceiro estágio é submetido a um agrupamento adicional de acordo com a sazonalidade de suas séries temporais. Este agrupamento é realizado em dois subestágios. No primeiro, de acordo com Daigo (2005), a existência da sazonalidade é detectada através da aplicação da análise fatorial combinado com a análise das componentes principais das curvas de carga de cada grupo reconhecido. Daigo (2005) apresentou o método de decomposição padrão (*Pattern Decomposition Method - PDM*) para analisar imagens no

qual cada pixel é expresso por meio de soma linear de padrões espectrais fixos. O uso de padrões espectrais fixos torna possível a comparação dos dados tanto no domínio da frequência como no domínio do tempo. O *PDM* adota um sistema de coordenadas oblíquas tal como é utilizado na análise fatorial. Os algoritmos desenvolvidos foram aplicados aos dados de uma determinada empresa do setor elétrico e os resultados demonstraram que a análise fatorial é bastante semelhante ao *PDM*.

Retomando a explicação do método proposto na versão univariada, a análise da sazonalidade foi realizada por meio da relação entre o consumo horário de todas as curvas durante o período de 24 horas. Os dados são representados por uma matriz $24 \times n_c$ sendo n_c é o número de curvas apresentadas em cada grupo reconhecido no final do estágio 3. De acordo com Yu et al. (2011), o método de análise fatorial aplicado a esta matriz proporciona a identificação de um número reduzido (menos do que 24 tal como sugerido pela análise das componentes principais) de fatores que caracterizam a sazonalidade de cada curva.

O segundo subestágio compreende a aplicação do Método de Agrupamento Subtrativo (CHIU, 1994) sobre estes fatores. O método proposto por Chiu (1994) é uma adaptação do método apresentado por Yager e Filev (1992) e compreende as seguintes etapas:

i. Etapa 1 – cálculo inicial dos potenciais (P_i) dos objetos em relação a um centro de grupo:

$$P_i = \sum_{j=1}^{n} e^{-\alpha \left\| x_i - x_j \right\|^2} \tag{80}$$

sendo

$$\alpha = \frac{4}{r_a^2} \tag{82}$$

e r_a é uma constante positiva. Assim, a medida do potencial de um dado objeto é função da distância deste objeto em relação aos demais. O raio determina o objeto de maior potencial, à medida que o mesmo se aproximar do centro do grupo.

ii. Etapa 2 – Definição do primeiro centro de grupo
O objeto com maior potencial é escolhido como o primeiro centro de referência (x_1).

iii. Etapa 3 – atualização dos potenciais dos objetos:

$$P_i = P_i - P_i^* e^{-\beta \left\| x_i - x_j \right\|^2} \tag{82}$$

sendo

$$\beta = \frac{4}{r_b^2} \tag{83}$$

e r_b é uma constante positiva. Assim, subtraiu-se um valor de potencial de um objeto por uma função da distância com decaimento exponencial a partir do primeiro centro de grupo, sendo esta ponderada também pelo potencial do primeiro centro. Recomenda-se $r_b = 1{,}5\ r_a$.

iv. Etapa 4 – Definição dos outros centros

Depois da atualização dos potenciais escolhe o objeto com maior potencial como referência para o segundo centro. Este processo se repete até que o maior potencial superior o limiar estabelecido pelo critério de parada.

v. Etapa 5 – critério de parada

Os novos centros são escolhidos até que o seguinte critério seja satisfeito:

$$P_k^* < \varepsilon P_1^* \tag{84}$$

Onde ε é um menor valor estipulado como tolerância.

O fator de carga citado no Quadro 5 é a relação entre as demandas média e máxima de uma curva de carga. O fator de carga é um índice de avaliação para o uso racional de energia elétrica pelo consumidor (NAZARKO e STYCZYNSKI, 1999). Em linhas gerais, tendo em conta a demanda máxima no período sob análise, quanto maior o fator de carga, mais racional e econômica é a utilização de energia, pois o planejamento da distribuição de energia é feito normalmente com base na demanda máxima. Dessa forma, as companhias energéticas precisam compatibilizar estes objetivos, ou seja, ao mesmo tempo incentivar o aumento do fator de carga dos equipamentos (ou ativos demandantes de cargas de energia) agregando as cargas destes equipamentos de tal forma que a carga resultante do sistema possua um fator de carga elevado, maximizando o uso da energia elétrica fornecida. Neste sentido, a gestão da demanda incentivando deslocamento de picos de determinadas classes de consumidores é um importante instrumento de equalização destes interesses.

A primeira fase (Figura 6) é repetida sucessivamente, a fim de verificar a eventual semelhança entre alguns dos padrões (média de cada grupo), indicando a necessidade de reagrupamento. Esta primeira fase é concluída quando existe uma convergência no número de padrões. A convergência é sempre assegurada, porque não é admitida a ocorrência de grupos contendo apenas um objeto até que o índice geral de silhueta sature ou assuma valor igual a

um. Assim, o número de padrões é um resultado do próprio método evitando a necessidade de uma estimativa inicial e no final da primeira fase (depois de convergência) apenas um padrão ou tipo é reconhecido para cada grupo (padrão ponto protótipo).

Na segunda fase do método STAC (Figura 6), cada uma das curvas de carga da amostra está associada a uma das curvas típicas reconhecidas na primeira fase usando uma abordagem de algoritmo de agrupamento não hierárquico, onde é adotado a métrica de distância Euclidiana e o método de formação de grupo da menor distância. Os grupos finais obtidos são submetidos a uma inspeção da qualidade de agrupamento. Uma das medidas adotadas para medir a qualidade de agrupamento é o índice silhueta (ROUSSEEUW, 1987).

5.2 ESTUDO DE CASO E RESULTADOS

Reitera-se que a amostra de trabalho neste estudo foi oriunda do programa de eficiência energética implementado pela Companhia Energética do Maranhão (CEMAR). O programa de eficiência energética consistiu na substituição de refrigeradores nas residências de consumidores de baixa renda distribuídos em municípios do Estado de Maranhão. Os dados foram coletados diretamente dos refrigeradores das unidades consumidoras antes (caso I) e depois (caso II) da substituição dos equipamentos. Refrigeradores novos foram doados às unidades consumidoras em substituição aos equipamentos antigos com menores recursos tecnológicos e prazos de vida útil esgotados.

Conforme já foi dito, quatro municípios do estado do Maranhão, na área de abrangência da CEMAR, foram considerados e o Programa foi implementado durante o período de novembro de 2008 a julho de 2009. Este programa envolveu a substituição de 5250 refrigeradores velhos por novos. Uma amostra de oitenta curvas de carga (refrigeradores velhos), apresentando um alto consumo de energia elétrica (caso I (Figura 7), o consumo médio de 82 kWh) e outra amostra de 80 curvas de carga, após a troca de refrigeradores (caso II (Figura 8), o consumo médio de 52 kWh) estavam disponíveis. O tamanho da amostra representou um nível de erro amostral de 10% e um nível de confiança de 95% no parâmetro de predição da população. O *IPMVP* (EVO, 2007) recomenda um erro de amostragem de até aproximadamente de 10%.

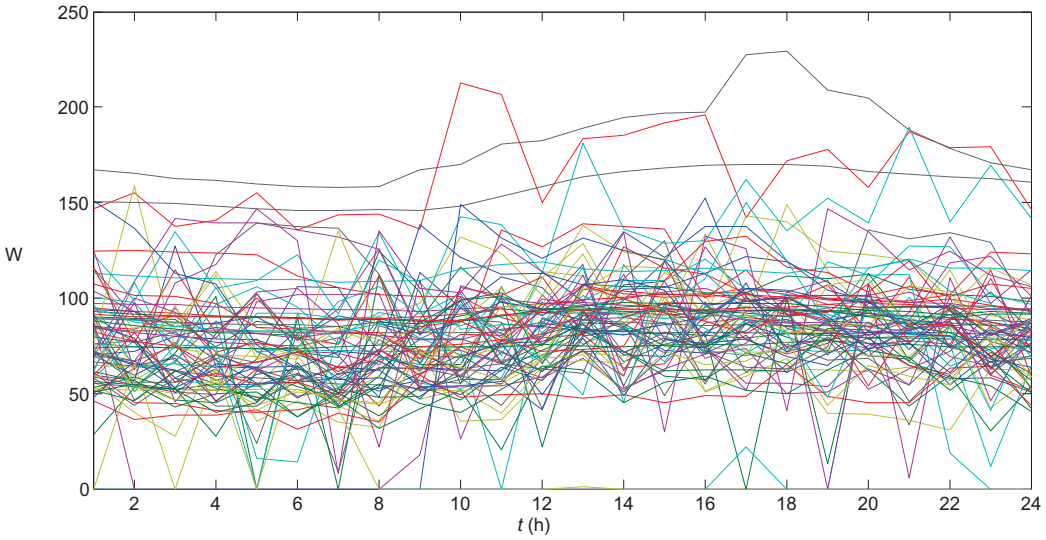


Figura 7 – Curvas de carga dos refrigeradores antes da troca (caso I).

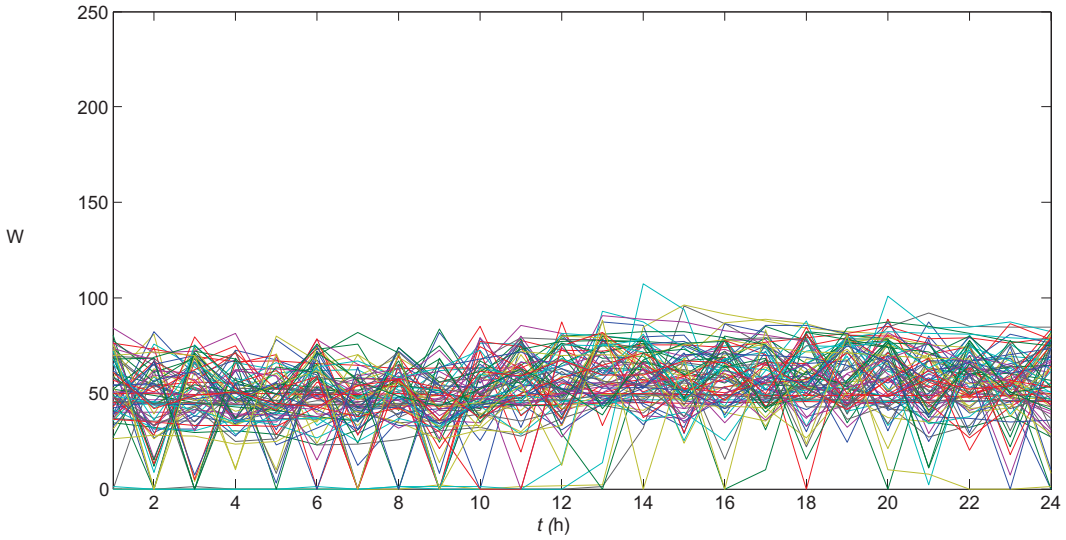


Figura 8 – Curvas de carga dos refrigeradores depois da troca (caso I).

Conforme o próprio nome, o método STAC compreende uma atividade inicial de seleção das curvas de carga (objetos). As curvas da amostra original foram analisadas para identificar e excluir curvas discrepantes (*outliers*). Esta análise consistiu na aplicação da análise fatorial juntamente com Análise das Componentes Principais (*Principal Componente Analysis –PCA*) (DAIGO, 2005) (também utilizada na primeira fase do método). A *PCA* foi inicialmente adotada para sugerir o número de fatores (*F*) que explicam 80% da variância total da amostra. A *PCA* proporciona uma redução na dimensionalidade de cada curva de carga (24 pontos), permitindo a seleção de fatores (inferior a 24) que pode representar o comportamento dinâmico de cada curva. A análise fatorial foi então aplicada para a obtenção

dos escores fatoriais (80 escores para cada fator). As Figuras 8-9 mostram o valor de distribuição dos escores fatoriais (*Z*) para os casos I e II. Os pontos identificados no *box-plot* pelo sinal "+" indicam curvas atípicas. Hubert e Vandervieren (2008) mostraram que o *boxplot* é uma ferramenta gráfica comumente utilizada para visualizar a distribuição dos dados contínuos unimodais. Segundo os autores, o *boxplot* mostra informações sobre o posicionamento, dispersão, assimetria e curtose dos dados. No entanto, os dados podem ser erroneamente declarados como valores atípicos quando eles possuem valores extremos. Neste trabalho é apresentado um exemplo de aplicação do *boxplot* resultando numa representação dos dados onde os seus valores atípicos são automaticamente e rapidamente detectados sem fazer qualquer pressuposto paramétrico sobre a distribuição dos dados. Exemplos e resultados de simulação mostram as vantagens deste novo procedimento (HUBERT e VANDERVIEREN, 2008). Em relação ao estudo de caso deste trabalho, cada ponto pode ocorrer em mais de um fator (mais de um "box-plot") associados à mesma curva. Esta análise permitiu a identificação de 13 curvas atípicas no caso I e 23 curvas atípicas no caso II. O aumento no número de fatores (número de "box-plot") no caso II está associado a uma maior variação sazonal causada pela menor demanda de energia dos motores dos novos refrigeradores. Este comportamento é verificado nas Figuras 9-10.

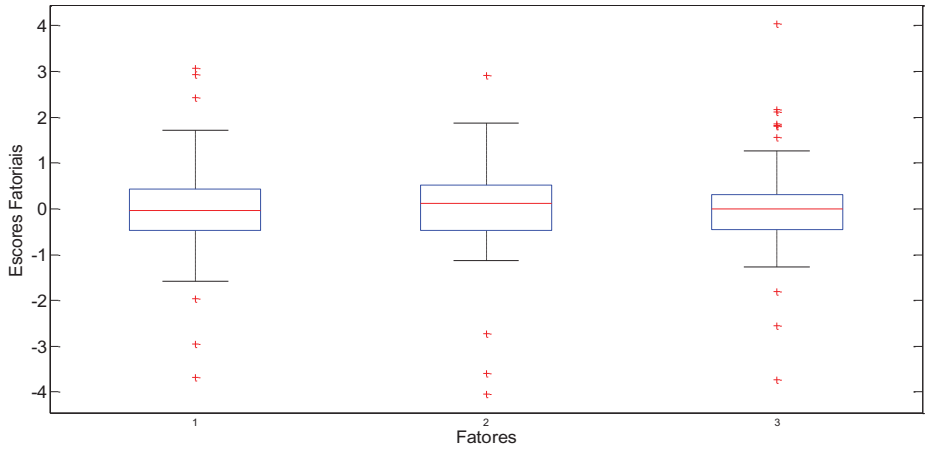


Figura 9 – Distribuição dos escores fatoriais de cada fator (caso I).

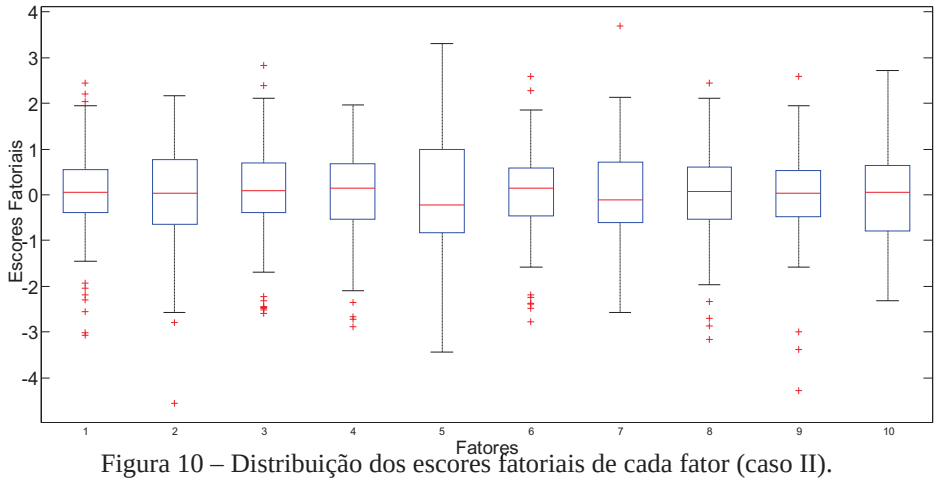


Figura 10 – Distribuição dos escores fatoriais de cada fator (caso II).

Os resultados obtidos com aplicação do algoritmo do método STAC foram comparados com o *FCM*. Conforme já discutido anteriormente, o FCM é um algoritmo já consolidado pertencente às famílias "*c-means*" de modelos de agrupamentos (BEZDEK et al., 2005; BENSAID et al., 1996), adequado para objetos de agrupamentos representados por séries temporais (LIAO, 2005).

O algoritmo *FCM* requer uma estimativa inicial para o número de grupos e, neste caso, o mesmo número de grupos fornecidos pelo algoritmo do método STAC foi considerado. O nível de confiança para os testes estatísticos aplicados nos três primeiros estágios da primeira fase do STAC foi de 99% em ambos os casos I e II.

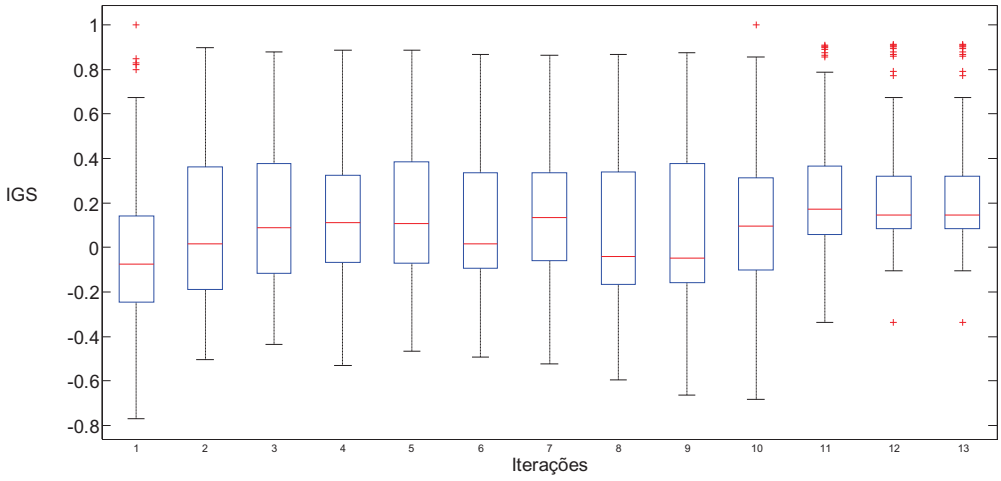


Figura 11 – Evolução dos Índices Globais de Silhueta (IGS) das iterações do método STAC (caso I).

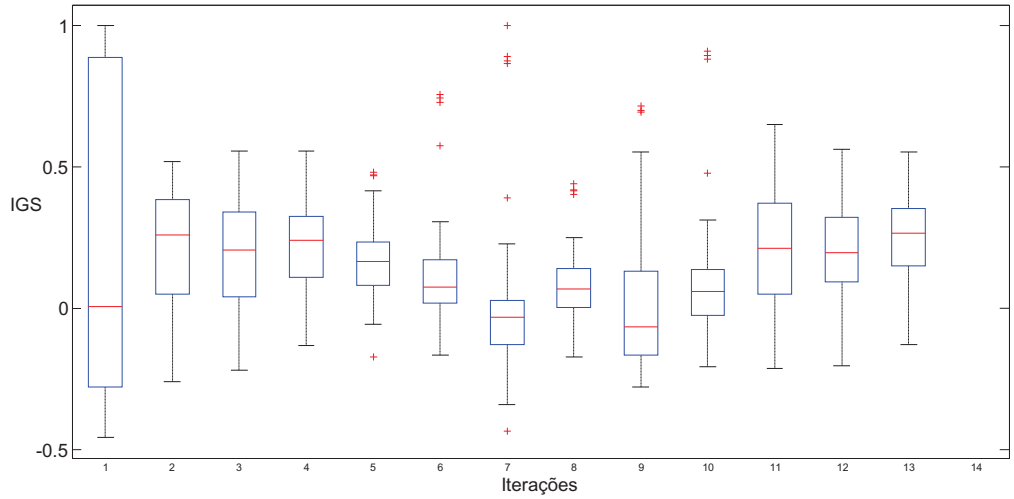


Figura 12 – Evolução dos Índices Globais de Silhueta (IGS) das iterações do método STAC (caso II).

A aplicação do algoritmo do método STAC para o caso I foi capaz de reconhecer, após a convergência (Figura 11), a existência de três padrões ou perfis de demanda (Figura 12). Por outro lado, dois dos três padrões reconhecidos pelo *FCM* foram semelhantes atestando o reconhecimento efetivo de apenas dois padrões (Figura 13). O algoritmo do método STAC foi capaz de reconhecer um terceiro padrão de consumo relacionado a um grupo de 13 curvas. Este resultado sugere a capacidade de STAC para lidar com uma amostra de objetos com um nível mais elevado de heterogeneidade (antes da substituição dos refrigeradores). Além disso, o terceiro padrão representa um perfil com menor consumo de energia, mesmo considerando o uso de refrigeradores velhos. A qualidade de agrupamentos obtida com o *FCM* foi ligeiramente inferior de acordo com o Índice Global de Silhueta IGS (igual a 0,25 e 0,28 para a *FCM* e STAC respectivamente). Neste trabalho, reitera-se que um Índice Global de Silhueta (IGS), representando a média dos índices de silhueta (ROUSSEEUW, 1987) obtidos para todos os objetos, foi adotado para avaliar a qualidade de agrupamento.

Para a amostra de curvas de carga após a substituição de refrigerador (caso II), após a convergência (Figura 11), ambos STAC e *FCM* reconheceram a existência de apenas um grupo e padrões semelhantes (Figura 14). Isso mostra que os perfis de demanda de energia elétrica tornou-se mais semelhante após a substituição de refrigeradores, indicando um aumento na uniformidade entre os consumidores.

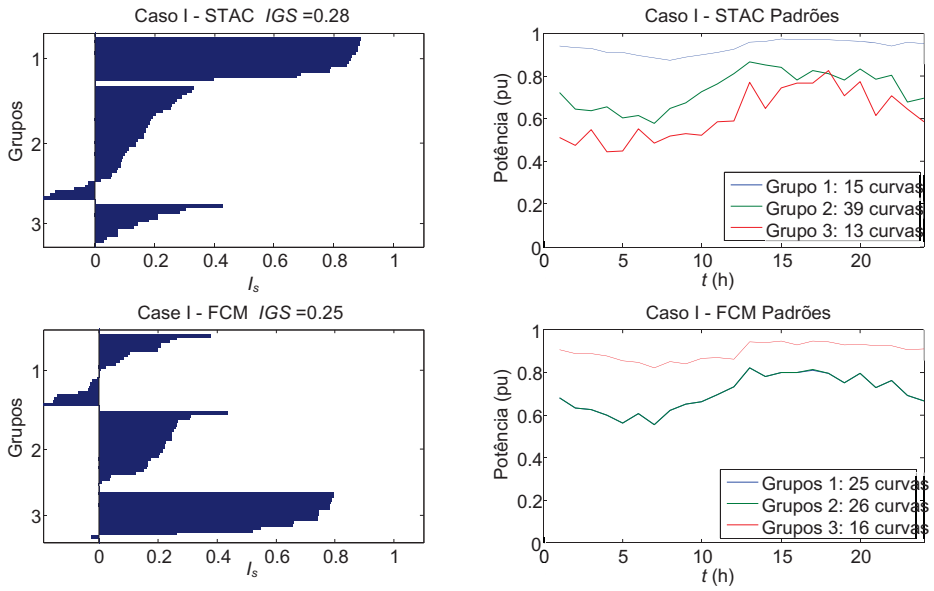


Figura 13 – Índices de Silhueta (I_s) e padrões reconhecidos pelos métodos STAC e *FCM* sem curvas atípicas na amostra (caso I).

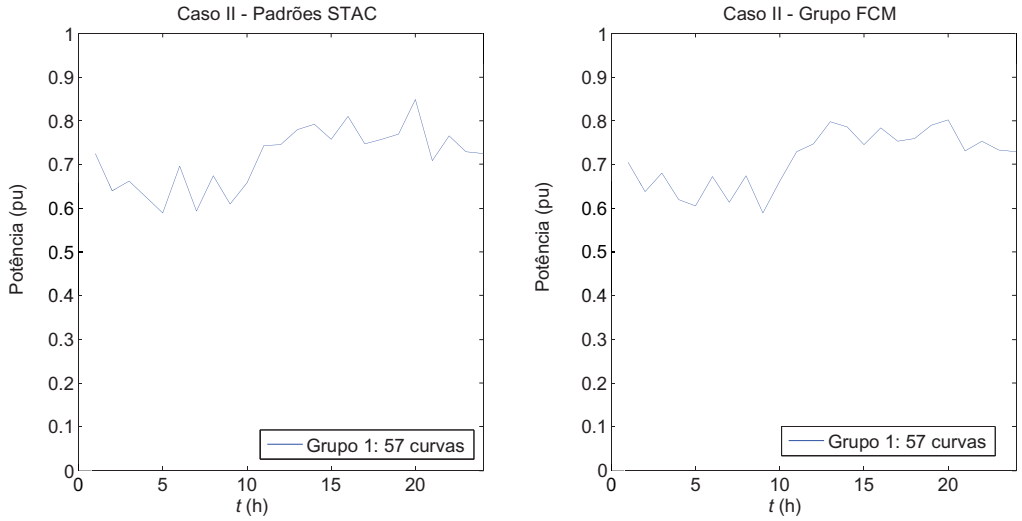


Figura 14 - Padrões reconhecidos pelos métodos STAC e *FCM* (caso II).

Os objetos (séries temporais) em cada grupo foram transformados em uma única série temporal, constituindo uma possível trajetória de um processo estocástico (ABDEL-AAL, 2006; SAINI e SONI, 2002). Neste caso, pode-se verificar as variações sazonais em cada grupo por meio de análise de autocorrelação (BOX et al., 2008). A existência de dois níveis de tendência nos padrões reconhecidos (especialmente em relação ao caso II) sugere neste caso a aplicação da análise de autocorrelação ($\rho(\tau)$) nas defasagens (τ) de primeira ordem da série original, a fim de mitigar os efeitos não-estacionários.

A Figura 15 apresenta o correlograma associado ao grupo modal (o número mais alto das curvas de carga) do caso I. Somente o primeiro valor da autocorrelação é significativo indicando que as variações sazonais são devidas a fatores aleatórios, em outros termos, possivelmente, o hábito de uso (o abrir e fechar da porta do refrigerador) não favoreceu ao aparecimento de variações sazonais porque a ineficiência dos refrigeradores antigos em conservar a temperatura (motivadas, por exemplo, pelo ineficiente isolamento térmico das portas dos refrigeradores) mascarou este efeito.

No caso II (Figura 16) aparecem de maneira aperiódica valores significativos da autocorrelação, ou seja, algumas autocorrelações não sucessivas são significativas, indicando que, o hábito de uso pode ser monitorado pelo comportamento sazonal da curva de carga dos refrigeradores, sobretudo, porque o isolamento térmico das portas dos refrigeradores foi equalizado (todos os refrigeradores são novos e possuem a mesma tecnologia de operação). O aumento das variações sazonais, também foi confirmado pela análise preliminar das curvas (análise fatorial). Vale reiterar que, o aumento da sazonalidade é previsível e coerente porque o novo refrigerador tem um isolamento térmico melhor e, portanto, o motor tem um comportamento intermitente de funcionamento.

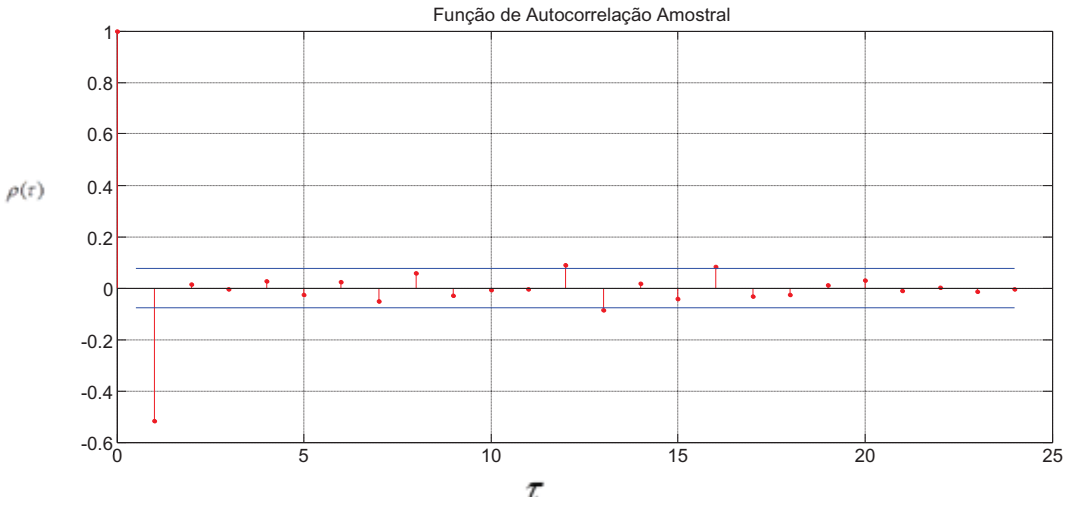


Figura 15 - Valores das autocorrelações ($\rho(\tau)$) das defasagens (τ) de primeira ordem (grupo modal caso I).

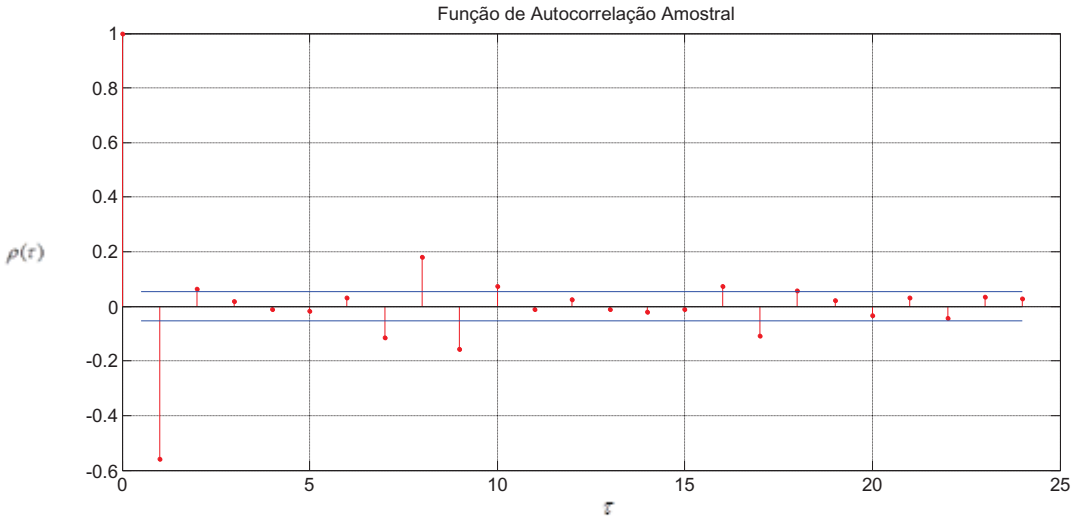


Figura 16 – Valores das autocorrelações ($\rho(\tau)$) das defasagens (τ) de primeira ordem (grupo modal caso II).

Uma análise adicional consiste na verificação da distribuição dos fatores de carga nos grupos reconhecidos. No caso I (Figura 13), o algoritmo do método STAC proporcionou grupos com medianas dos fatores de carga próximas dos fatores de carga dos respectivos padrões reconhecidos (0,65, 0,75 e 0,95 *pu* para os grupos 3, 2 e 1, respectivamente - Figura 17). Isto não se verifica nos resultados obtidos pelo método *FCM*, revelando uma inconsistência no reconhecimento dos padrões (curvas típicas) neste caso. De acordo com o STAC (e também *FCM*) a distribuição dos fatores de carga, no caso II mostrou valores mais baixos (Figura 18). Os motivos estão relacionados aos mesmos fatores que aumentam a sazonalidade, ou seja, novos refrigeradores fazem uso de tecnologia mais avançada, como um melhor isolamento térmico e um motor com menor demanda energética.

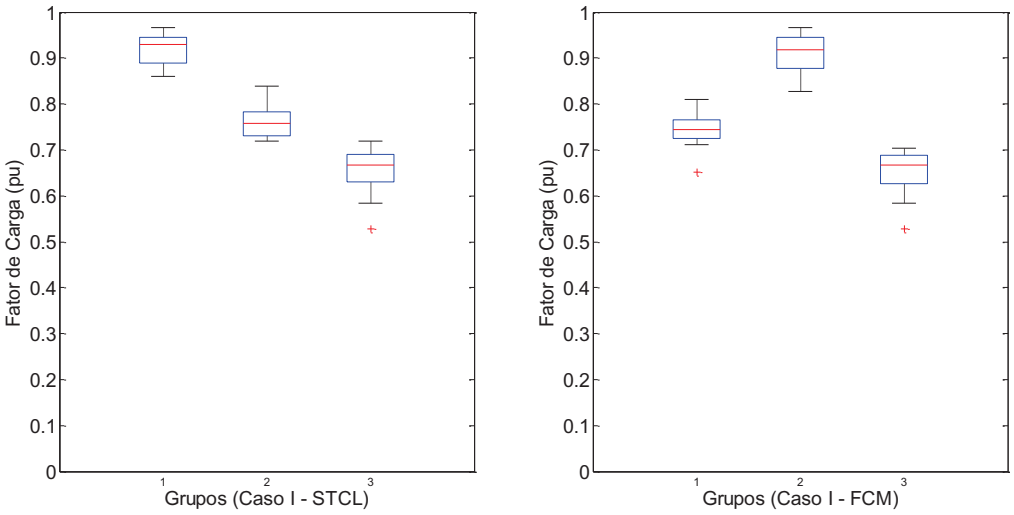


Figura 17 – Distribuição dos fatores de carga dos grupos reconhecidos (caso I).

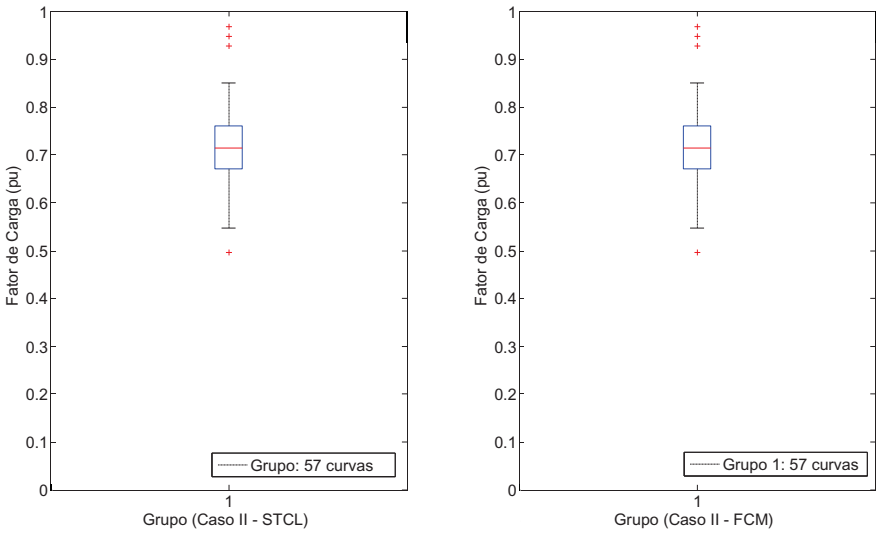


Figura 18 – Distribuição dos fatores de carga dos grupos reconhecidos (caso II).

6 RECONHECIMENTO DE PADRÕES EM CURVAS DE CONSUMO DO SETOR ELÉTRICO – CASO MULTIVARIADO

Este capítulo apresenta uma versão estendida do STAC para reconhecimento de padrões de séries temporais multivariadas (STAC-M), ou seja, adequada para seleção, tipificação e agrupamento de curvas de cargas e séries temporais correlatas. O estudo de caso analisado compreende um programa de eficiência energética realizado pela Companhia Elétrica de Alagoas (Brasil) que envolveu, entre outras ações, uma análise do impacto da substituição de refrigeradores nas residências dos consumidores de baixa renda. Neste caso, o método é capaz de reconhecer os padrões de consumo de energia associado aos perfis dinâmicos de temperaturas externas e internas do refrigerador e também proporcionar uma forma consistente para avaliar a eficácia do programa de eficiência energética. O item 6.2 apresenta o método e a métrica de validação adotada (ANUAR e ZAKARIA, 2010). O item 6.3 apresenta um estudo de caso e uma comparação do método desenvolvido com uma versão modificada do método *FCM* capaz de tratar o problema de protótipo não pontual (FONTES et al., 2012).

Conforme já foi mencionado, as séries temporais de variáveis meteorológicas, as temperaturas externas e internas, são as principais séries temporais que afetam ao comportamento das curvas de carga de energia elétrica dos refrigeradores. A temperatura externa interfere na eficiência térmica do equipamento, especialmente no que diz respeito ao modo de uso (frequência de abertura da porta do refrigerador), o que contribui para os desvios da temperatura interna ou de refrigeração (acondicionamento). O sistema de controle da temperatura no interior do refrigerador possui um sensor, o termostato, que mede a temperatura no interior e a compara com limites preestabelecidos. Se a temperatura estiver abaixo do limite mínimo aceitável, o motor do compressor é desligado, caso contrário, o motor é ligado até que a temperatura atinja o limite mínimo. Por sua vez, o isolamento térmico do refrigerador melhora a capacidade de manutenção da temperatura interna por um período maior. Por isso, a incorporação destas variáveis oferece a possibilidade de uma análise de agrupamento e reconhecimento de padrões mais amplo, uma melhor avaliação dos ganhos obtidos e do programa de eficiência implementado. Assim, um padrão observável de interesse estratégico do presente estudo é o consumo de energia dos refrigeradores (GELLER et al, 1998; GOLDMAN et al., 2005).

6.1 O MÉTODO STAC-M

O método STAC-M realiza a seleção, tipificação (reconhecimento de padrões) e agrupamento de séries temporais multivariadas (C*lustering and Pattern recognition of Multivariate time series - CPT-M*), baseando-se também numa extração sistemática de características dos objetos que exige um procedimento diferenciado na sua análise de dados (FERREIRA et al., 2015).

A versão multivariada do método STAC (STAC-M) é capaz de reconhecer padrões de consumo de energia associado aos perfis dinâmicos de temperaturas externa e interna dos refrigeradores, proporcionando uma forma de avaliação das interações existentes entre estas variáveis. Nesta versão, cada objeto constituído por uma STM é composto por três variáveis (três séries de tempo), ou seja, o consumo de energia (representados por curvas de carga), temperatura ambiente e temperatura de refrigeração, e as séries temporais de cada objeto são normalizadas no intervalo [0, 1] dividindo as medições horárias pelo valor do pico. As Figuras 19-20 apresentam exemplos de objetos (curvas de carga em conjunto com cada uma das temperaturas) extraídos das amostras associadas a refrigeradores antigos (em condições subnormais de uso) e refrigeradores novos (com as mesmas especificações técnicas).

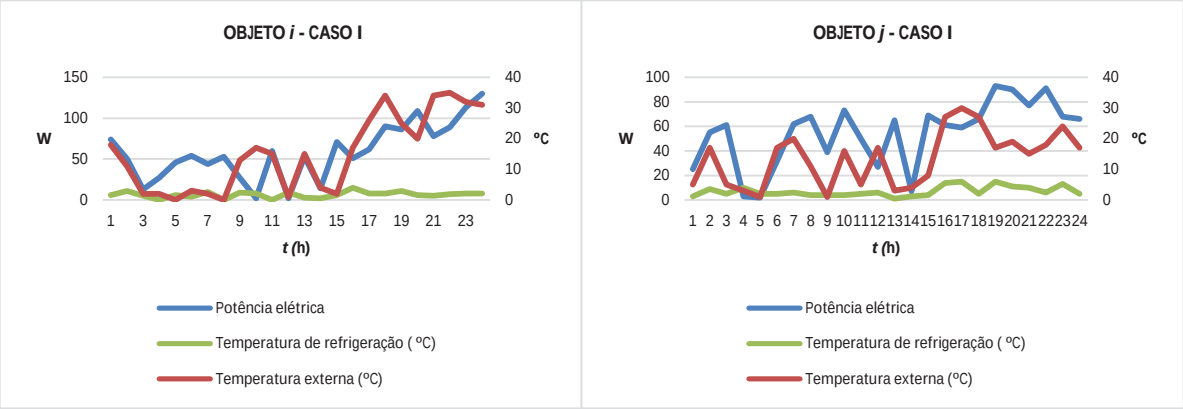


Figura 19 – Exemplos de objetos (séries temporais) – refrigeradores antigos.

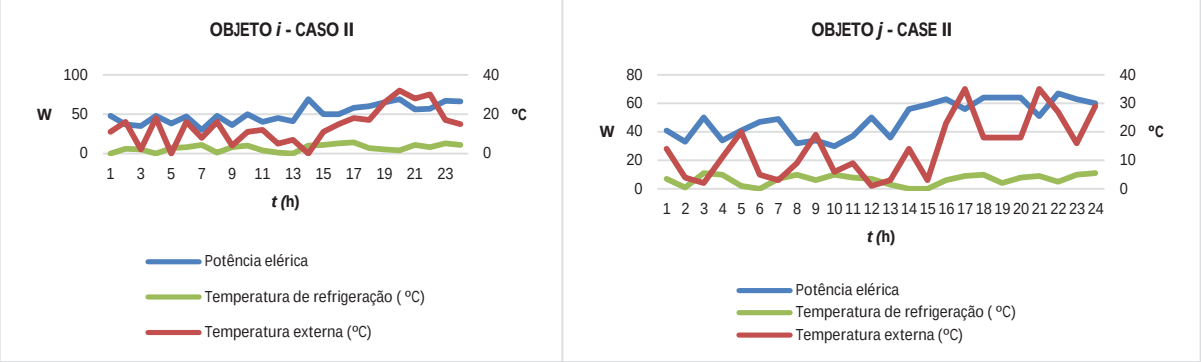
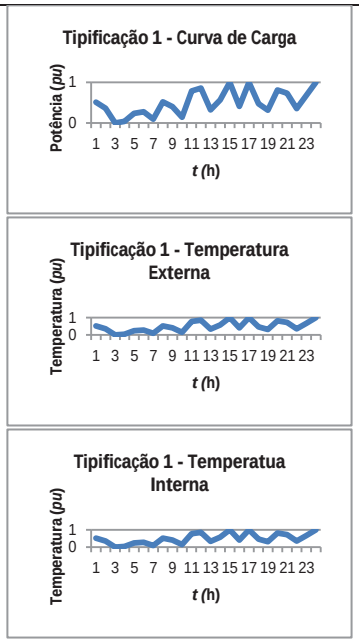
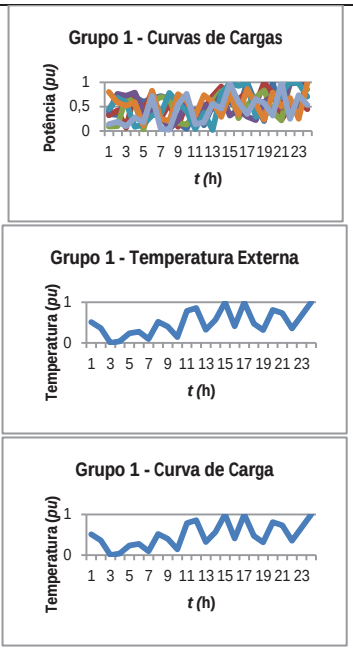


Figura 20 – Exemplos de objetos (séries temporais) – refrigeradores novos.

O procedimento de agrupamento também oferece o reconhecimento de um padrão (protótipo ou centro) para cada grupo. Cada padrão compreende a um conjunto de três séries temporais (STM) que não está associado a um equipamento específico, mas sim a um conjunto de refrigeradores (consumidores) representando, portanto, o comportamento térmico e de consumo de energia elétrica deste grupo.

A Figura 21 apresenta o esquema geral do método STAC-M e na Figura 22 o algoritmo do método STAC-M é apresentado.

SÉRIES	TIPIFICAÇÃO	AGRUPAMENTO
PADRÃO 1		
...	...	...

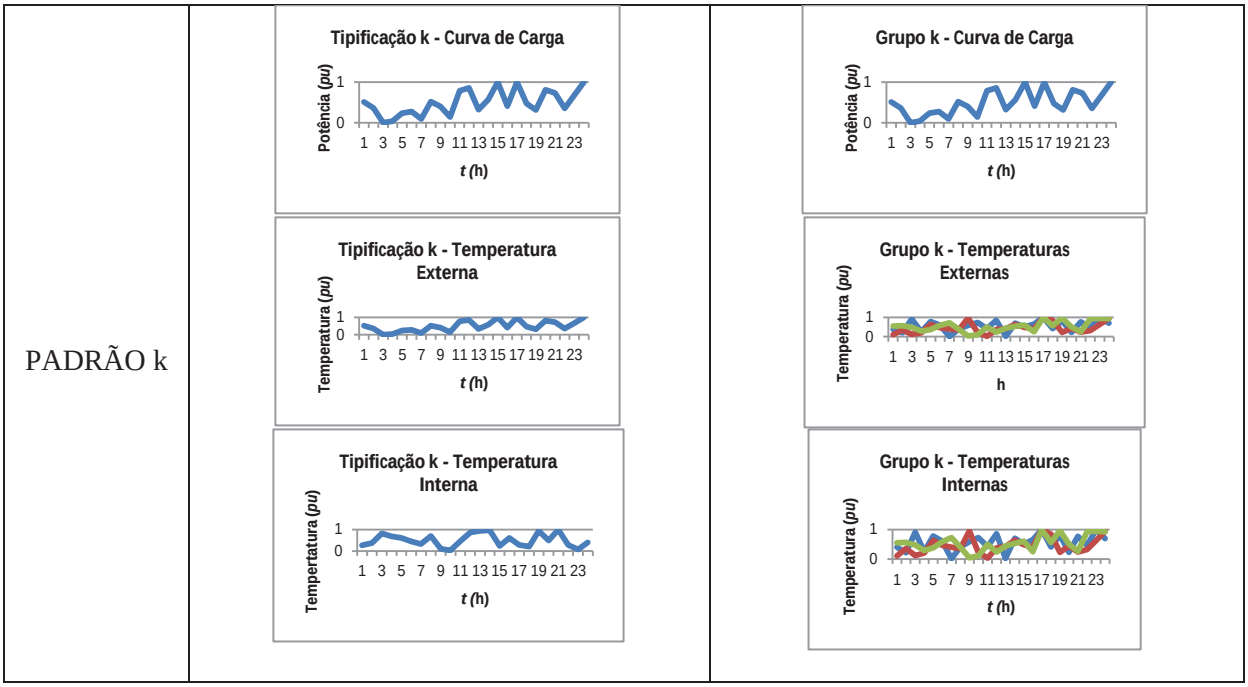


Figura 21 - Método do STAC-M.

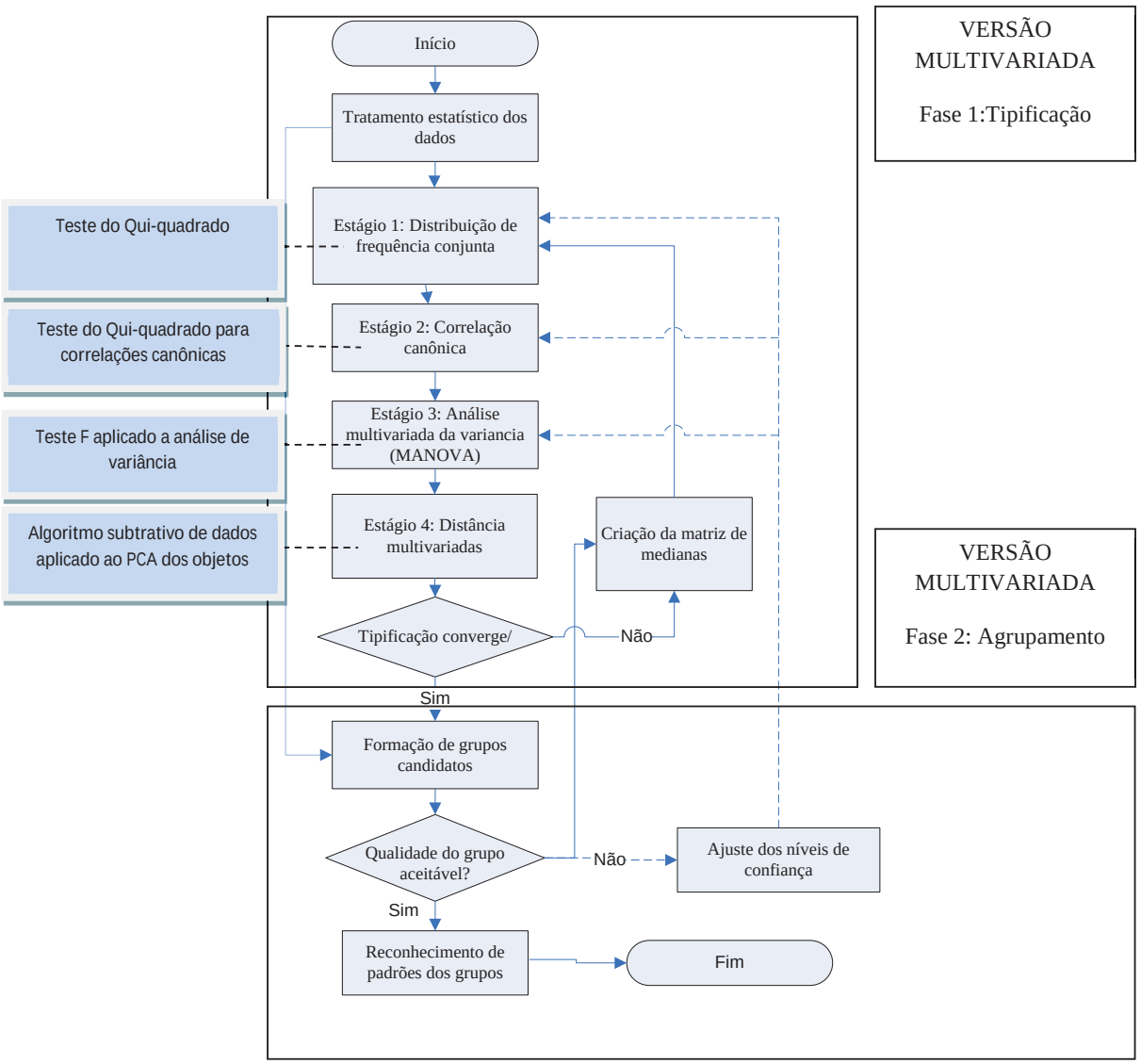


Figura 22 - Algoritmo do Método STAC-M – Versão Multivariada.

Quadro 6 – Testes estatísticos usados no método STAC-M.

STAC (Séries temporais univariadas)			STAC-M (Series temporais multivariadas)		
Critério de Agrupamento	Testes estatísticos	Pressupostos	Critério de Agrupamento	Testes estatísticos	Pressupostos
Distribuição de frequência da demanda horária (Estágio 1)	Teste de aderência Qui-quadrado [57]	• Amostra aleatória e independente • Frequência de classe menor que 5 • Distribuição uniforme dos desvios entre os valores observados e esperados	Distribuição conjunta da frequência de series temporais multivariadas (objetos)	Teste de aderência Qui-quadrado [58]	A mesma do método STAC
Seleção utilizando o nível de correlação entre as curvas de cargas (Estágio 2).	Teste *t* de *Student's* para verificar a correlação entre duas amostras Independentes [59]	• Distribuição normal • Igualdade entre os coeficientes de correlação • Homocedasticidade	Correlação canônica entre dois objetos.	Teste Qui-quadrado para verificar correlação canônica entre dois objetos [60]	• Linearidade • Normalidade multivariada • Homocedasticidade • Ausência de multicolinearidade
Selection by mean consumption (Estágio 3)	Teste *t* de *Student's* para verificar a diferença de médias entre duas amostras Independentes [59]	• Distribuição normal • Igualdade entre médias • Homocedasticidade	Multivariate analysis of variance (MANOVA) between pairs of objects.	*F*-test applied to multivariate analysis of variance [61]	• Distribuição normal multivariada • Igualdade entre vetores de médias

Analogamente ao STAC, o método proposto incorpora vários critérios em agrupamento e reconhecimento de padrões em curvas de carga, ao contrário das abordagens tradicionais que essencialmente usam o critério de distância entre as curvas de carga (FIGUEIREDO et al., 2005).

O método STAC-M também compreende duas fases e na primeira realiza o reconhecimento de padrões através de sucessivas iterações de toda a amostra (54 refrigeradores velhos com alto consumo de energia elétrica - caso I e 54 refrigeradores novos, com um consumo mais econômico - caso II).

A primeira iteração da primeira fase executa o agrupamento de objetos constituídos por três séries de tempo, quais sejam, curvas de cargas, temperaturas internas e externas dos refrigeradores. A primeira fase é repetida sucessivamente, a fim de verificar a eventual semelhança entre alguns dos padrões de objetos (mediana de cada uma das variáveis dentro do grupo).

Dessa forma, no final da primeira fase (após convergência), padrões ou tipos são reconhecidos para cada grupo, ou seja, esta primeira fase é concluída quando existe uma convergência no número de padrões. A convergência é sempre assegurada, porque, assim como no método STAC, não é admitida a ocorrência de grupos contendo apenas um objeto até que o índice geral de silhueta sature ou assuma valor igual a um.

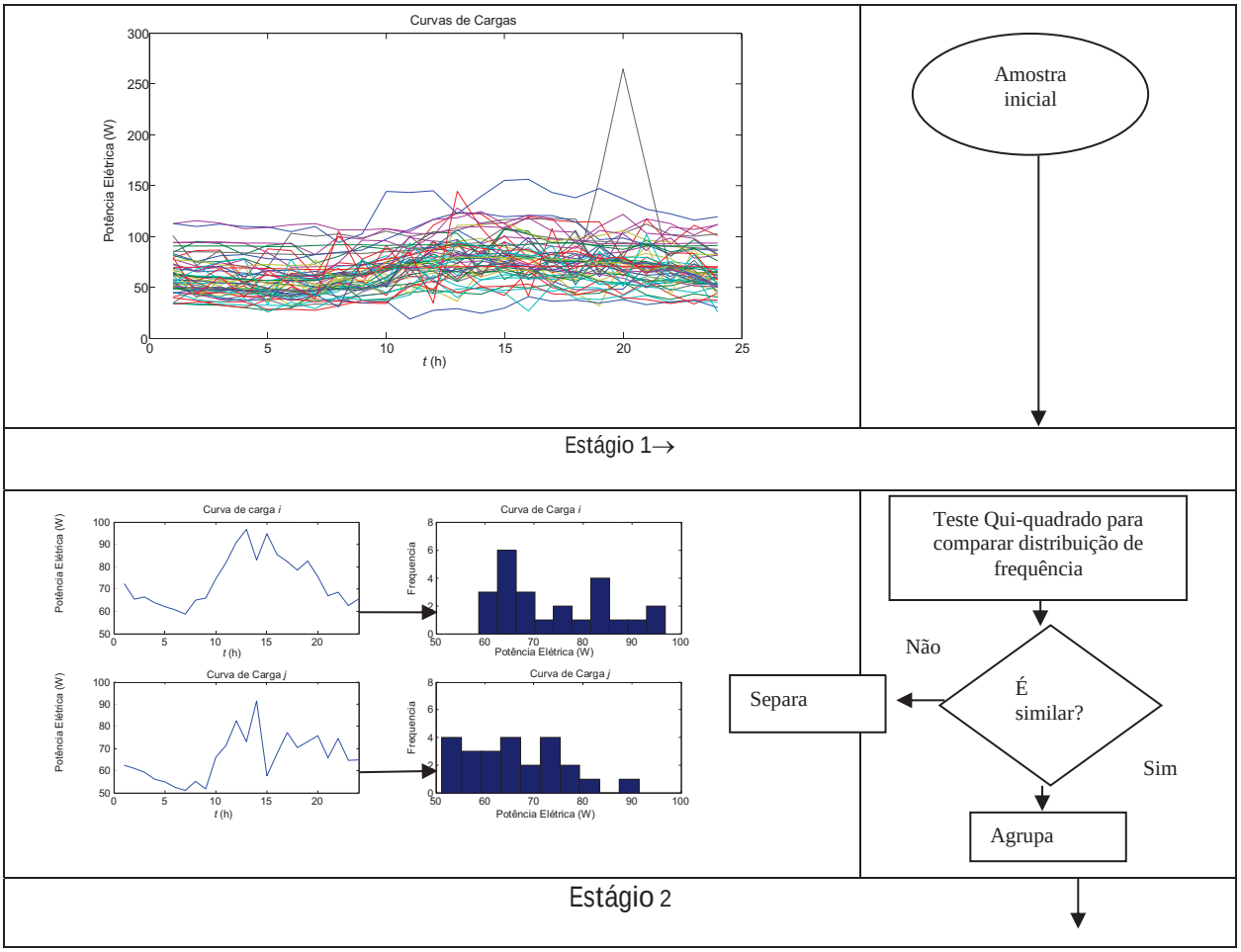
Assim, o número de padrões é um resultado do próprio método evitando a necessidade de uma estimativa inicial. Enquanto que na segunda fase do método STAC-M (Figura 22) os objetos da amostra original estão associados a um dos protótipos não pontual (padrão) reconhecidos na primeira fase, usando um procedimento de agrupamento não hierárquico. Cada objeto (matriz com três colunas) é representado pela sua primeira componente principal (o suficiente para capturar pelo menos 70% da variabilidade dos dados).

Em seguida, o mesmo é associado com o padrão mais semelhante (menor distância Euclidiana), favorecendo uma partição nítida de grupos com características parecidas entre seus objetos (BEZDEK et al., 2005). Se apenas uma componente principal não é capaz de capturar pelo menos 70% da variabilidade dos dados, a análise de similaridade deve considerar mais do que uma componente principal e uma métrica de similaridade adequada para as séries temporais multivariadas deve ser empregada tal como o *SPCA* (Análise de Similaridade das Componentes Principais) (Eq. (86)) (YANG e SHAHABI, 2004; LI e WEN, 2014, DOBOS e ABONYI, 2012).

Portanto, a segunda fase define os grupos finais associando cada objeto da amostra original a um dos padrões reconhecidos na primeira fase através de uma matriz de

dissimilaridade, obtida a partir do emprego do *SPCA*. A fim de se ter um melhor entendimento do STAC-M, cada uma destas fases serão mais detalhadas a seguir.

Especificamente, os quatro estágios da primeira fase são realizados utilizando testes de hipóteses estatísticas multivariadas e métricas de distâncias de acordo com os objetivos pretendidos. A fim de permitir uma melhor compreensão dos quatros estágios, uma apresentação gráfica (Figura 23) é feita baseando-se apenas em séries temporais univariadas (curvas de carga) utilizadas no reconhecimento de padrões do método STAC. A extensão para o método STAC-M é realizada através da utilização do teste estatístico correspondente à abordagem multivariada (QUADRO 6).



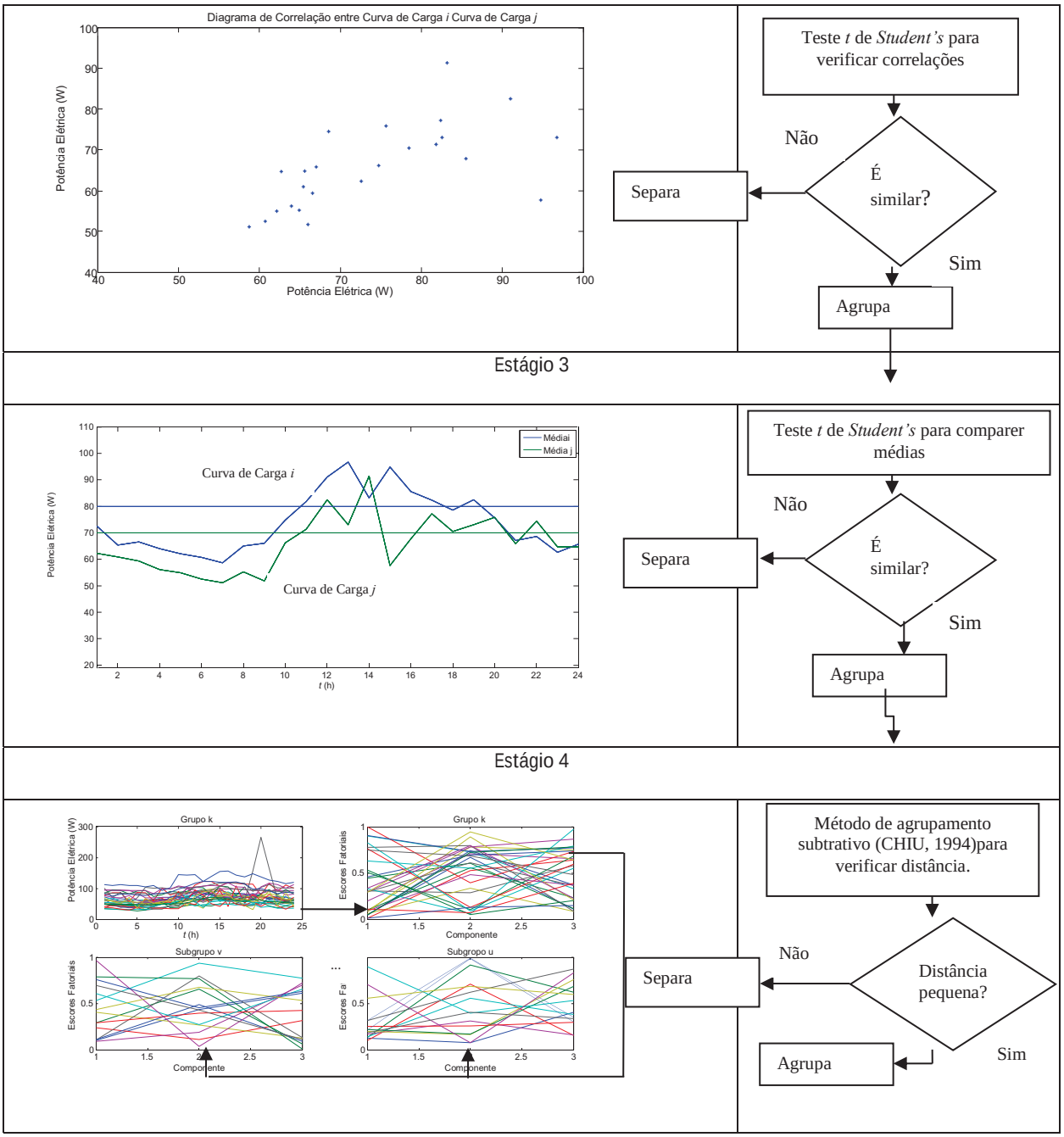


Figura 23 – Primeira fase do método STAC – Ilustração considerando o caso univariado.

Reportando-se mais uma vez à explicação para versão univariável (método STAC), a Figura 23 reitera que, no primeiro estágio, uma distribuição de frequência, considerando uma sequência finita de intervalos de valores de consumo de energia para cada curva de carga, está representada por meio de um histograma.

A similaridade entre os histogramas é verificada através do teste do qui-quadrado (QUADRO 6) e duas curvas de carga (objetos) podem ser agrupadas ou não. No segundo estágio, são obtidos os coeficientes de correlações binárias entre as curvas de carga que

pertencem ao mesmo conjunto reconhecido na primeira fase. A semelhança entre esses coeficientes são verificados através do teste *t* de *Student* e novos grupos são reconhecidos de forma hierárquica.

No terceiro estágio, o grau de similaridade entre as médias aritméticas das curvas de carga que pertencem ao mesmo conjunto reconhecido no estágio anterior é verificado também pelo teste *t* de *Student*. Finalmente, no quarto estágio, os rótulos (períodos de tempo estratificados em horas) dos dados dos grupos reconhecidos no terceiro estágio são substituídos pelas suas componentes principais e estes componentes apresentam um comportamento temporal resultante das forças sazonais intrínsecas aos dados.

Por conseguinte, novos grupos poderão ser obtidos com a aplicação do Método de Agrupamento Subtrativo (CHIU, 1994), pois um novo mapeamento das proximidades entre os objetos é estabelecido. Especificamente, neste último estágio, o procedimento de geração das novas variáveis é obtido através da aplicação do método *PCA* e, conforme já foi dito, baseia-se na derivação de uma matriz variância-covariância de amostra inicial de variáveis originais que é transformada em uma nova amostra de variáveis não correlacionadas.

Em outros termos, essas novas variáveis são combinações lineares das variáveis originais e são hierarquizadas em ordem decrescente de importância quanto a variância presente na amostra. Com isso, a primeira componente principal é a combinação linear normalizada com variância máxima (LI e WEN, 2014). Uma matriz de dados $Y \in \Re^{v \times w}$ (v linhas ou objetos e w as variáveis ou rótulos dos objetos) pode ser representada da seguinte forma:

$$Y = t_1 \times p_1^T + t_2 \times p_2^T + \ldots + t_\pi \times p_\pi^T + E \tag{85}$$

onde t_i é um vetor de escores ($t_i \in \Re^v$), p_i é um componente principal, ($p_i \in \Re^w$) e E é a matriz de resíduos (MITSA, 2010; DENG et al., 2013). Portanto, o método *PCA* reduz o conjunto original de variáveis para π componentes principais e os pequenos desvios são representados por uma matriz residual (E) (LI e WEN, 2014). O quarto estágio (Figura 20) compreende uma quantificação multivariada de dissimilaridade entre os objetos gerados no terceiro estágio de acordo com os seus comportamentos sazonais.

Retomando a discussão para o método STAC-M, vale dizer que o procedimento dos três primeiros estágios da primeira fase do método STAC também foi adotado utilizando os

testes estatísticos correspondentes à abordagem multivariada. O Quadro 6 apresenta os critérios considerados na análise de similaridade realizada, em cada estágio da primeira fase, juntamente com os testes estatísticos aplicados. Na primeira iteração, os grupos são formados com base na similaridade entre os objetos. Após a primeira iteração, o método pressupõe que as medianas das três séries temporais (protótipos não pontuais) de cada grupo mantenham as suas próprias características e os mesmos testes são aplicados considerando sucessivamente as medianas. A existência de uma similaridade entre as medianas de acordo com os testes estatísticos aplicados (QUADRO 6) implica na união de padrões e novos grupos podem ser obtidos.

No que tange o quarto estágio, o procedimento em questão sofreu uma modificação que exige uma explicação à parte. No método STAC-M o quarto estágio foi sequenciado em quatro subestágios (Figura 24). O primeiro subestágio inclui a redução de dimensões de três séries temporais para duas componentes principais (o suficiente para capturar pelo menos 95% da variabilidade dos dados de amostra) associados a cada objeto. Estas componentes principais são derivadas a partir das variáveis originais (consumo de energia, temperaturas externa e interna do refrigerador), através da decomposição da matriz de covariância das variáveis originais em duas matrizes (uma com autovalores dispostos em ordem de impacto decrescente e a outra com os respectivos autovetores ou componentes principais (YANG e SHAHABI 2004). No segundo subestágio, a semelhança entre cada par de objetos é obtida através da *PCA* adotando a métrica de similaridade *SPCA* (SINGHAL e SEBORG, 2005). O índice *SPCA* é uma métrica de distância que se aplica a séries multivariadas baseadas também em *PCA*. O índice *SPCA* é capaz de quantificar a similaridade entre grupos de séries temporais através da comparação de seus componentes principais, ou seja, através do cálculo do ângulo entre os autovetores (Eq. (86)).

$$SPCA_{pq} = 1/4 \sum_{j=1}^{2} \sum_{i=1}^{2} \cos^2 \theta_{ji} \tag{86}$$

θ_{ji} é o ângulo formado entre o "j^{th}" componente principal do "p^{th}" objeto e o "i^{th}" componente principal do "q^{th}" objeto$(p, q = 1, \ldots, n)$. $SPCA_{pq}$ é uma medida de similaridade entre os objetos p e q. Considerando uma amostra com n objetos os índices binários ($SPCA_{pq}$) irão compor uma matriz quadrada de ordem n. Por uma questão de clareza, a Figura 24 apresenta a matriz *SPCA* com valores hipotéticos. No terceiro subestágio, uma escala de representação multidimensional (MDS), com base na matriz de dissimilaridade (D) é obtida (BÉCAVIN et

al., 2011). A matriz de dissimilaridade (*D*) é composta das métricas de dissimilaridade entre cada par de objetos (D_{pq}):

$$D_{pq} = 1 - SPCA_{pq} \tag{87}$$

Esta análise começa com a seguinte transformação em cada elemento da matriz de dissimilaridade:

$$K_{pq} = (-1/2).\left(D_{pq}^2 - D_{.q}^2 - D_{p.}^2 + D_{..}^2\right) \tag{88}$$

onde

$$D_{p.}^2 = \sum_{q=1}^{n} D_{pq}^2 / n \,. \tag{89}$$

$$D_{.q}^2 = \sum_{p=1}^{n} D_{pq}^2 / n \,. \tag{90}$$

$$D_{..}^2 = \sum_{p=1}^{n}\sum_{q=1}^{n} D_{pq}^2 / n \,. \tag{91}$$

Isto resulta em uma matriz *K* (*nxn*) que pode ser decomposta de acordo com $K = V \cdot \Lambda \cdot V^T$. As colunas de *V* (*nxn*) são os correspondentes autovetores da matriz K e Λ é uma matriz diagonal (*nxn*) com os correspondentes autovalores. Os dois autovetores associados aos maiores autovalores são usados como eixo ortogonal no escalonamento multidimensional (Multidimensional Scaling - *MDS*). A distância entre os pontos na representação de escalonamento multidimensional (Figura 24 – subestágio 3) fornece o nível de dissimilaridade entre os objetos originais na amostra (BÉCAVIN et al., 2011). Como resultado, o *MDS* permite uma visualização de objetos de series temporais multivariadas num plano bidimensional. No quarto estágio um Método de Agrupamento Subtrativo (CHIU, 1994) é aplicado aos pontos apresentados na representação *MDS* (Figura 24 – subestágio 4) e novos grupos podem ser obtidos com este novo mapeamento das proximidades entre os objetos.

Na segunda fase do método STAC-M (Figura 22) os objetos da amostra original são associados a um dos protótipos não pontual (objetos padrões) reconhecidos na primeira fase, usando uma abordagem de algoritmo de agrupamento não hierárquico, onde é adotado a métrica de distância Euclidiana e o método de formação de grupo da menor distância. A diferença deste processo em relação ao processo da segunda fase do método STAC (versão univariável) é que cada objeto (três séries temporais) pode ser representado por um único

vetor (componente principal associado ao maior autovalor) e a distância Euclidiana é calculada entre a componente principal de um objeto original e a componente principal de um objeto padrão.

Também foi adotado o Índice Global de Silhueta (*IGS*) para avaliar a qualidade de agrupamento adaptado para séries temporais multivariadas (Eq. (76)).

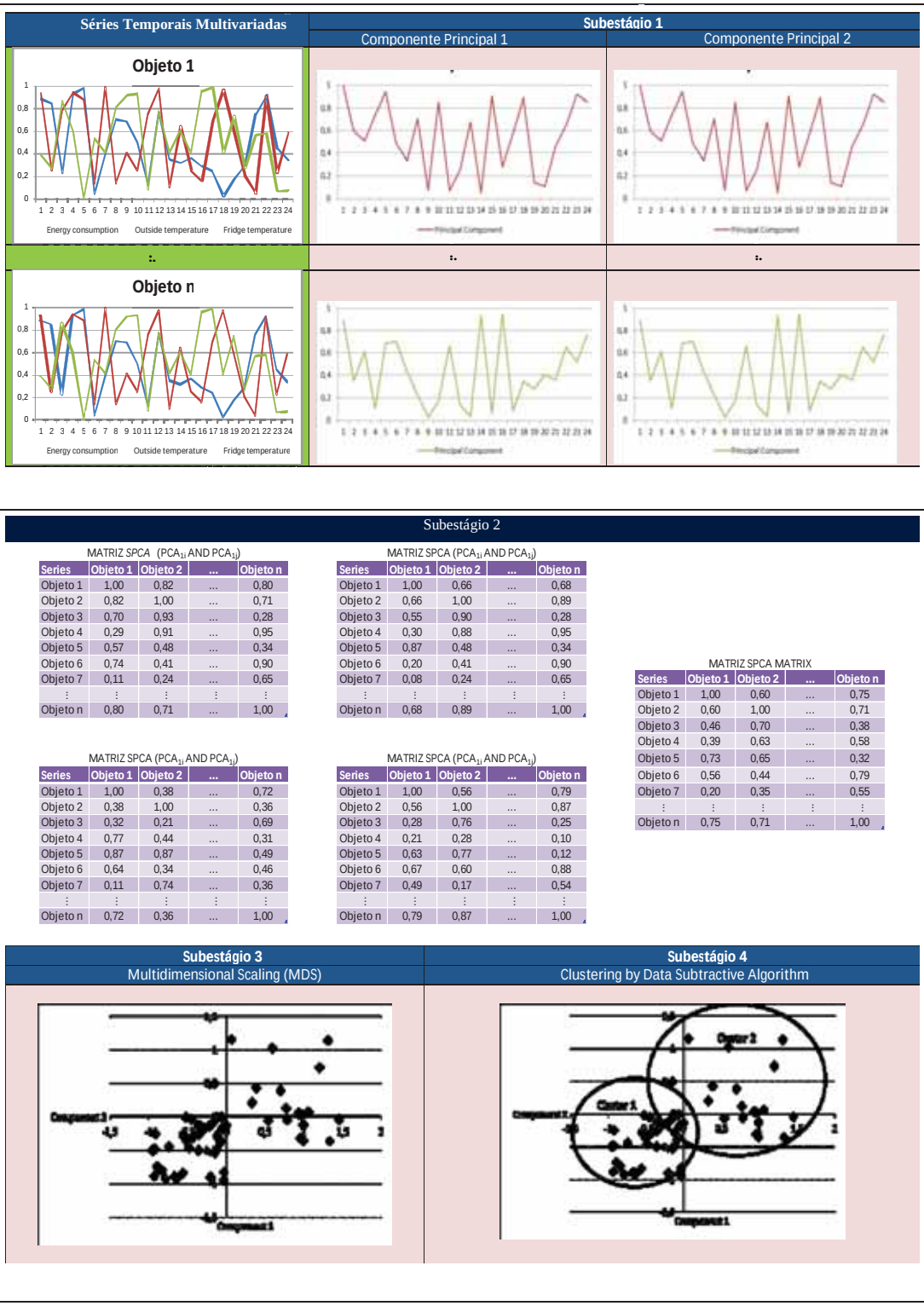


Figura 24 – O quarto estágio da primeira fase do método STAC-M.

6.2 ESTUDO DE CASO E RESULTADOS

O programa de eficiência energética que originou as séries temporais multivariadas coletadas envolveu essencialmente a substituição de 5.000 refrigeradores antigos por novos em comunidades de baixa renda. O procedimento de amostragem foi realizado com base na determinação do tamanho da amostra para estimar a média (ZHANG et al., 2004). Uma amostra de 54 refrigeradores antigos com alto consumo de energia elétrica (caso I, consumo médio de 35 kWh - Figura 25) e outra de 54 novos refrigeradores (caso II, consumo médio de 18 kWh - Figura 26) estavam disponíveis. O tamanho da amostra (54 consumidores em ambos os casos I e II) apresentou um erro de 8,7% e 6,0% para os casos I e II, respectivamente, e os níveis de confiança de 94,3% e 94,3% na predição do parâmetro de população. O *International Performance Measurement & Verification Protocol* (EVO, 2007) recomenda uma margem de erro de até aproximadamente 10%. Além disso, a população-alvo (5000 casas) compreende um subconjunto da classe de consumidores (consumidores de baixa renda) residenciais o que implica em uma menor variabilidade no comportamento sazonal do consumo. Neste sentido a variabilidade da população no caso II tende a ser ainda menor, pois a substituição dos refrigeradores (todos com as mesmas especificações técnicas) contribui para a homogeneização dos perfis de consumo.

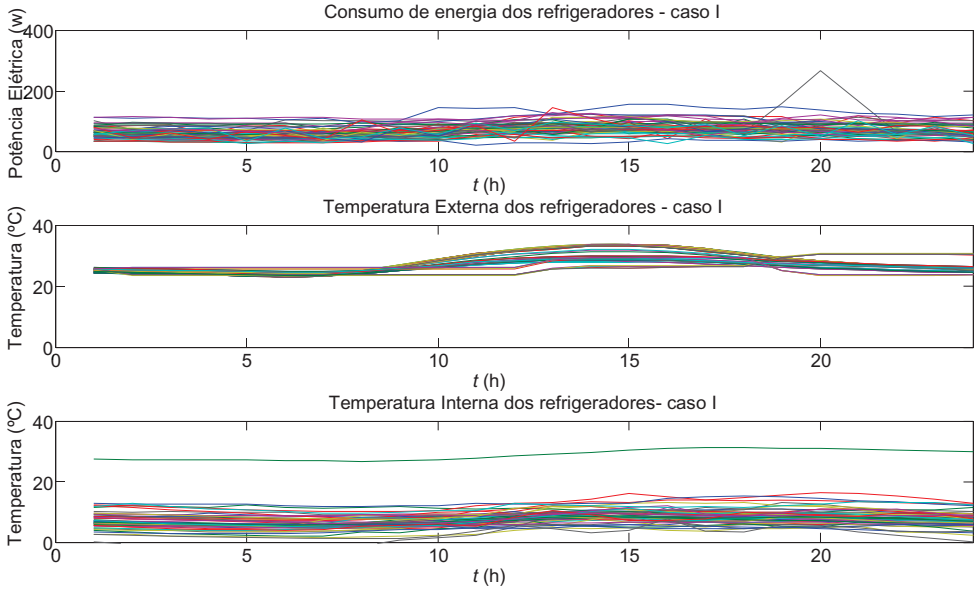


Figura 25 – Séries temporais referentes aos refrigeradores antes das trocas – caso I.

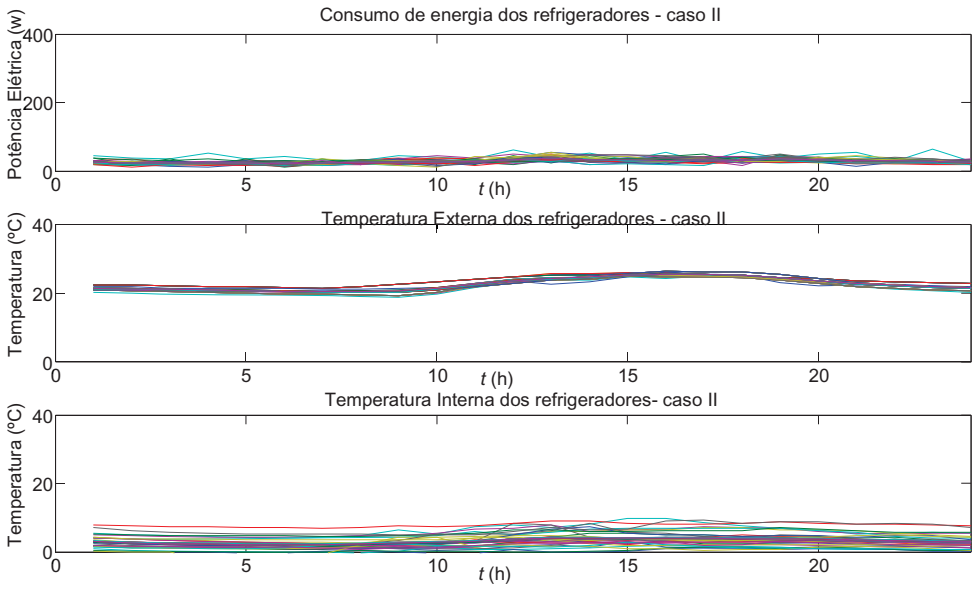


Figura 26 – Séries temporais referentes aos refrigeradores depois das trocas – caso II.

Inicialmente, os dados foram analisados para identificar e eliminar objetos atípicos. Este procedimento compreende a decomposição da amostra original (conjunto de dados) através de *PCA* (DAIGO, 2005) em uma amostra constituída de um número seletivo de combinações lineares (fatores) das variáveis originais. Esta decomposição foi capaz de descrever as principais variações no conjunto de dados, fornecendo a identificação e eliminação de objetos atípicos a partir da amostra original.

Com base nas amostras originais de cada caso (54 objetos e cada uma com 24 medições para cada variável, isto é, de curva de carga, temperaturas interna e externa), uma matriz de dados $Y \in \Re^{52\times 74}$ foi decomposta de acordo com a Eq. (86) e as amostras originais foram reduzidas para 4 (caso I) e 6 (caso II) fatores que consideram em cada caso, a descrição de, pelo menos, 50% da variabilidade nos dados. As Figuras 27-28 mostram a distribuição dos escores fatoriais para os casos I e II. Conforme já foi mencionado, os pontos identificados no box-plots pelo sinal "+" sugerem objetos atípicos. Estes sinais (+) podem identificar mais de uma vez a atipicidade de um mesmo objeto, por isso, o número de sinais (+), em geral, supera o número de objetos na amostra. Isto é devido ao fato que box-plots distintos podem identificar como atípico o mesmo objeto, provocando assim o aparecimento repetitivo do sinal "+". Portanto, os escores fatoriais foram úteis para a identificação de objetos atípicos de acordo com uma abordagem multivariada. Esta análise permitiu a exclusão de três objetos atípicos no caso I e 20 objetos atípicos no caso II. O aumento no número de fatores (número de box-plots) no caso II está associado a uma maior variação sazonal causada pela demanda

de energia mais baixa dos motores dos novos refrigeradores (tal como será mostrado mais adiante). Mesmo com a remoção de objetos atípicos estes erros aumentaram para 9,0% e 7,5% nos casos I e II, respectivamente.

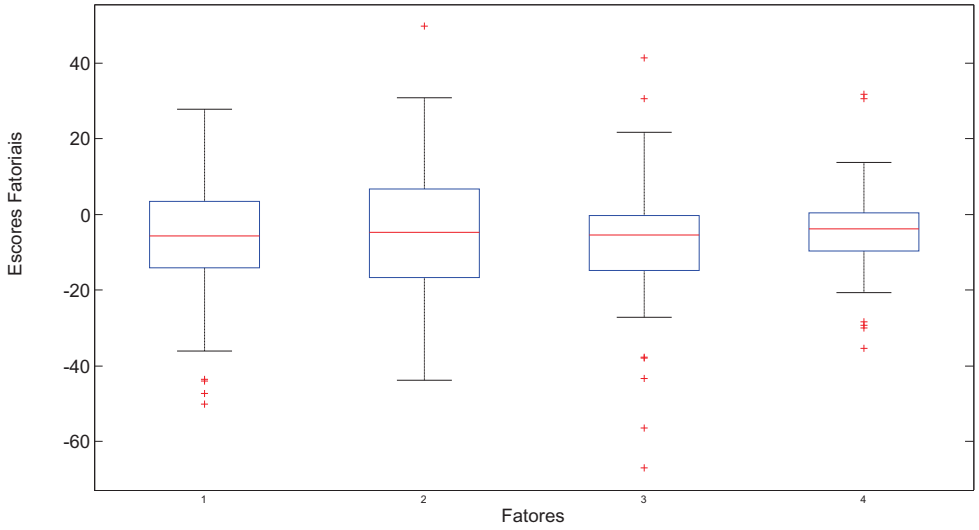


Figura 27 – Distribuição dos escores de cada fator (caso I).

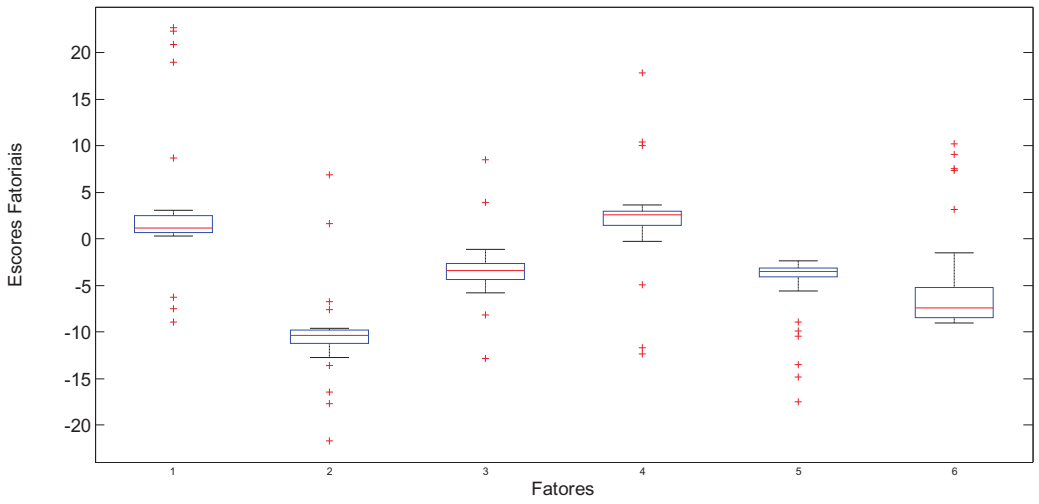


Figura 28 – Distribuição dos escores de cada fator (caso II).

Os resultados obtidos usando o método STAC-M foram comparados com o *FCM* (Eq. (61)). Neste caso, a formulação básica do algoritmo *FCM* (BEZDEK et al. 2005), apropriado

para um problema de protótipo pontual, teve de ser alterada, a fim de utilizar o *SPCA* em vez da métrica Euclidiana e um novo método de formulação para a *FCM* foi adotado (FONTES et al., 2012). Mesmo com esta versão modificada para aplicação em problemas de protótipos não pontual, apenas um parâmetro de ajuste do problema de otimização foi considerado (grau de fuzzificação da partição, $m \geq 1$). Este parâmetro foi definido igual a 2 de acordo com o que é sugerido na literatura (BEZDEK et al., 2005; HOPNNER et al., 2000).

A aplicação do método STAC-M em ambos os casos foi capaz de reconhecer a existência de dois grupos e dois padrões. Após as convergências (Figuras 29-30), o Índice Global de Silhueta (*IGS*) obtido pelo método STAC-M no caso I (*IGS* = 0,19) e no caso II (*IGS* = 0,46) e os índices obtidos pelo método *FCM* (caso I, *IGS* = - 0,12 e caso II, *IGS* = 0,21) sugerem que este último método apresentou, em ambos os casos, uma qualidade inferior ao STAC-M (Figuras 31-32). Além disso, os padrões reconhecidos pelo método *FCM* poderiam ser convertidos em um único padrão em ambos os casos I e II. Isso pode ser verificado por meio das Figuras 33-36 que apresentam, para cada grupo, o gráfico de paridade com os dois componentes principais associados a cada objeto e também para o padrão reconhecido (escalonamento multidimensional). A proximidade entre as componentes principais dos padrões reconhecidos pelo método *FCM* sugere que o STAC-M foi capaz de identificar os grupos mais heterogêneos. A disposição das componentes principais em ambos os casos também evidencia a melhor qualidade de agrupamento obtido pelo método STAC-M.

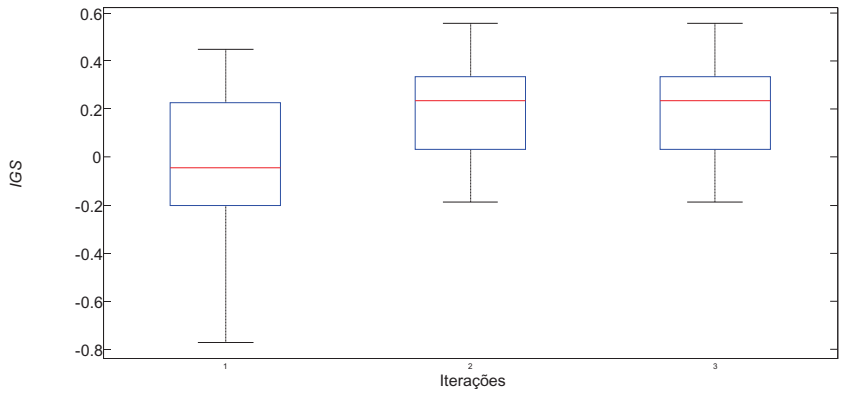


Figura 29 – Evolução dos Índices Globais de Silhuetas (*IGS*) das iterações do método STAC-M (caso I)

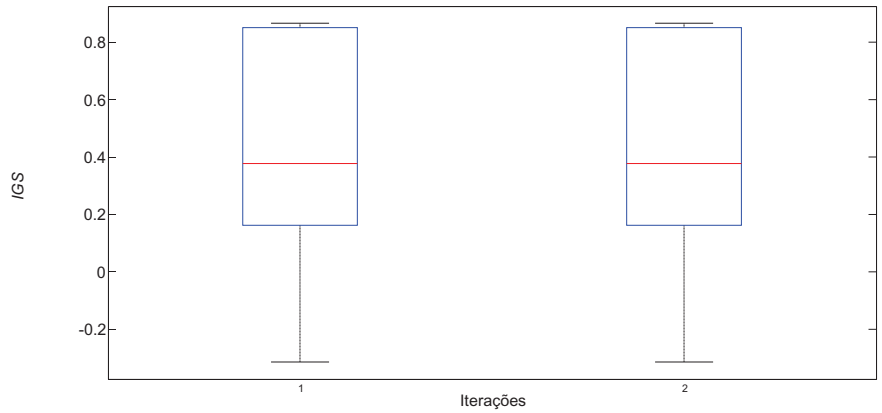


Figura 30 – Evolução dos Índices Globais de Silhuetas (*IGS*) das iterações do método STAC-M (caso II)

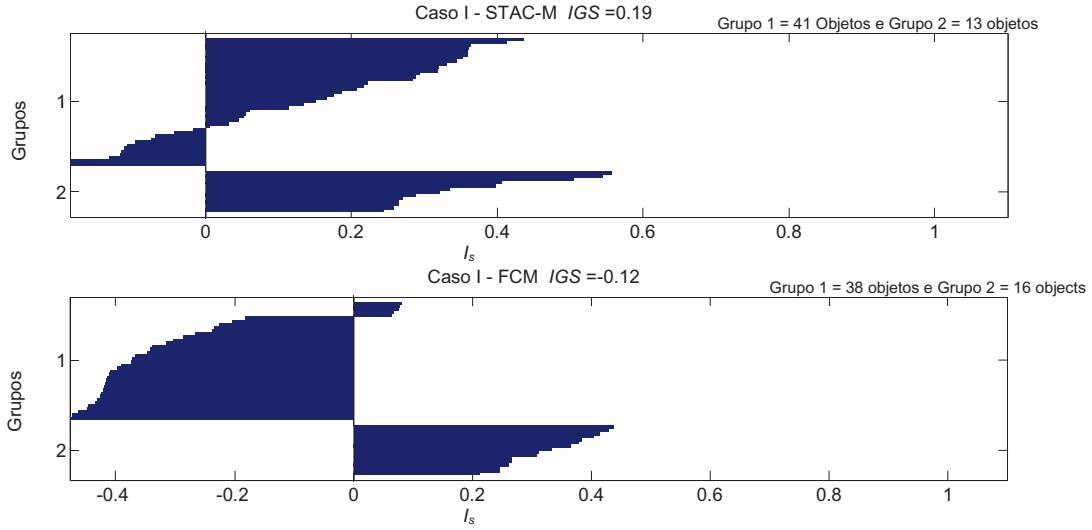


Figura 31 – Índices de silhueta (I_s) obtidos dos métodos STAC-M e *FCM* (case I).

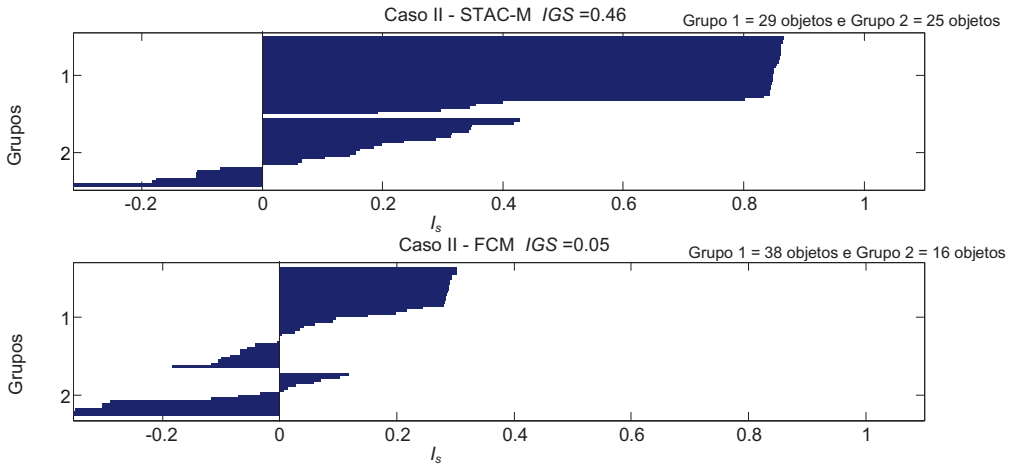


Figura 32– Índice de silhueta obtidos dos métodos STAC-M e *FCM* (case II).

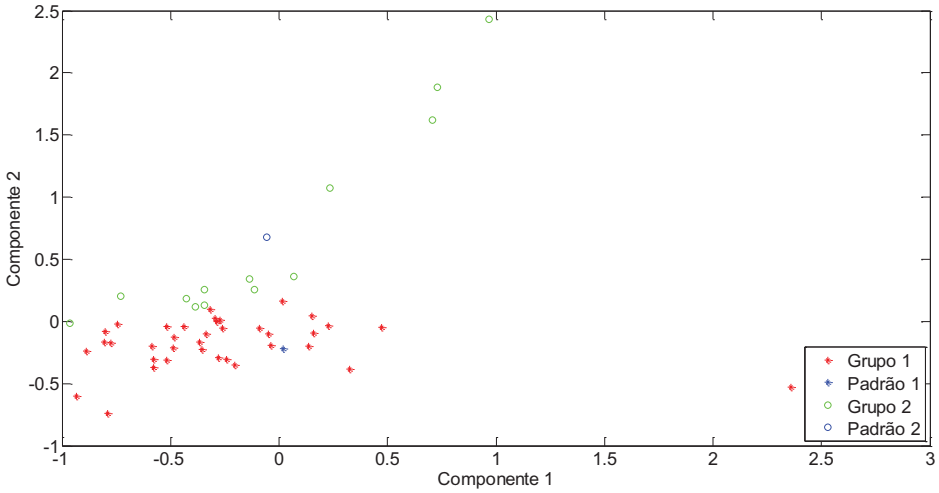


Figura 33 – Objetos e padrões obtidos pelo método STAC-M (caso I).

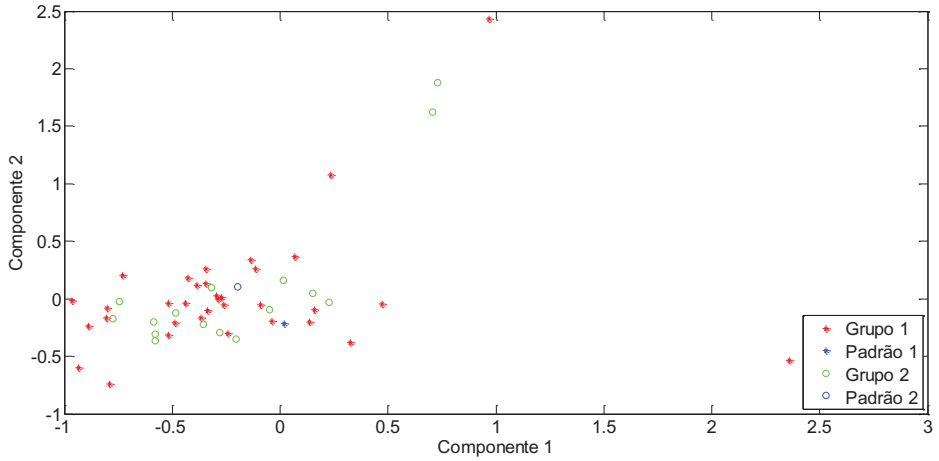


Figura 34 - Objetos e padrões obtidos pelo método *FCM* (case I).

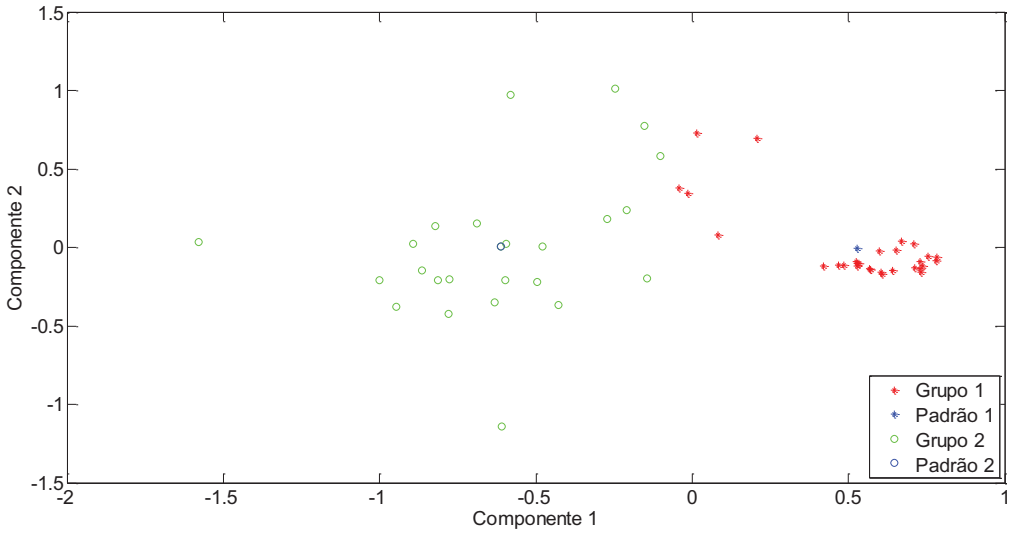


Figura 35 - Objetos e padrões obtidos pelo método STAC-M (caso II).

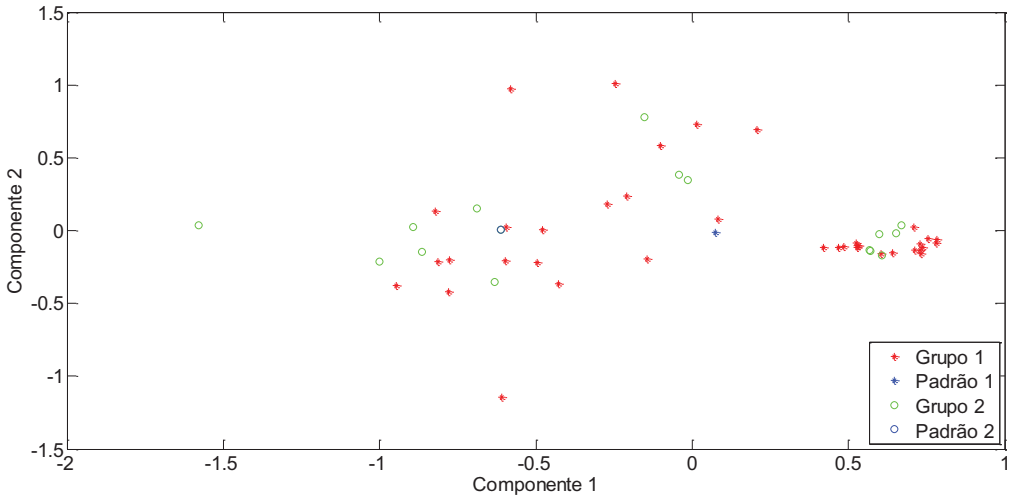


Figura 36 - Objetos e padrões obtidos pelo método *FCM* (case II).

As Figuras 3136 apresentam uma comparação entre os resultados obtidos a partir de métodos STAC-M e *FCM* considerando diferentes critérios de validação associados à qualidade de agrupamento.

Nas Figuras 31-32 a qualidade de agrupamento, de acordo com as características de coesão e separação dos grupos, é analisada através da distribuição do índice silhueta de todos os objetos em cada grupo reconhecido. Os resultados STAC-M para caso I (Figura 31-a) mostram que o grupo 1 tem cerca de 33% dos objetos com índices de silhueta negativos e o grupo 2 (quantidade de objetos menor que a metade do número de objetos do primeiro grupo)

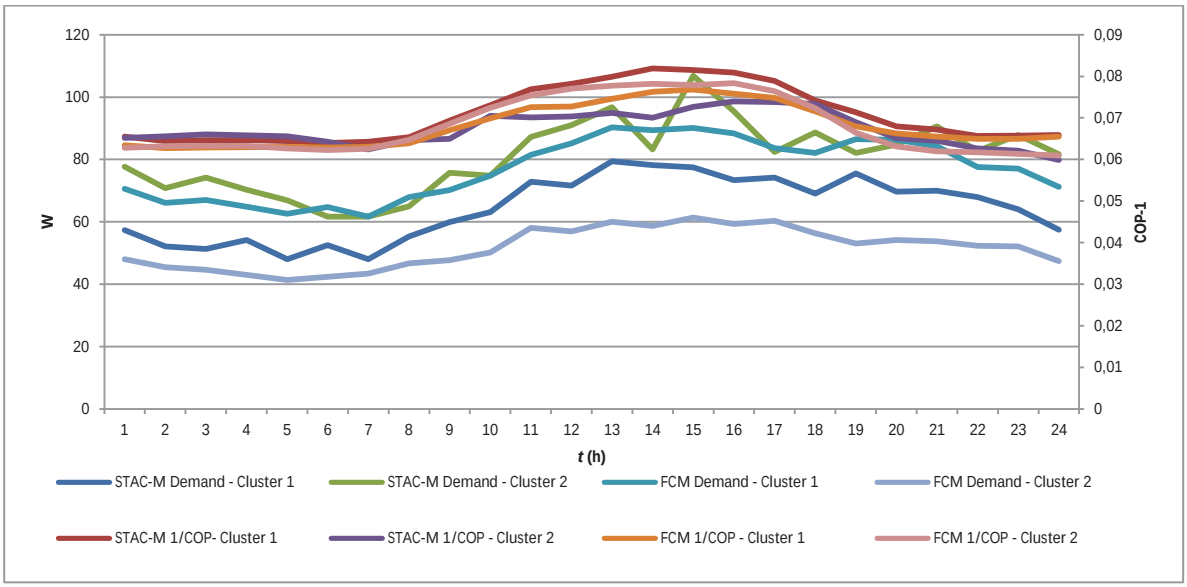
não apresenta índices silhueta negativos. Além disso, resultados do *FCM* para o caso I (Figura 31-b) mostram que o grupo 1 tem cerca de 90% dos objetos com índices de silhueta negativos. No caso II, o grupo 1 (Figura 32-a), reconhecido pelo STAC-M não apresenta índices silhueta negativos. Os resultados obtidos com *FCM* para o caso II mostram que o método STAC-M apresentou tendência a uma partição mais equilibrada em relação ao número de objetos (STAC-M, 29 e 25, *FCM*, 38 e 16).

Adotando uma abordagem semelhante ao que foi realizado na identificação de valores atípicos, a matriz de dados $Y \in \Re^{54\times 72}$ (54 objetos associados a três variáveis de séries temporais de 24 períodos de horas) foi decomposta em uma matriz de dados $W \in \Re^{54\times 2}$ constituída de duas componentes principais. As Figuras 33-36 apresentam os gráficos de paridade considerando as duas componentes principais referentes à matriz de objetos e aos respectivos padrões reconhecidos. Esta é também uma maneira de analisar a coesão de cada grupo e o nível de sobreposição entre os respectivos objetos.

No sentido de tornar mais consiste a análise comparativa entre os métodos STAC-M e *FCM*, calculou-se o inverso do coeficiente de desempenho (COP^{-1}), utilizado para quantificar a eficiência térmica através da relação $\frac{T_o - T_f}{T_f}$ (T_o e T_f são as temperaturas externa e interna do refrigerador). Sabe-se que a análise do COP^{-1} é mais adequada para o setor elétrico quando se lida com equipamento de arrefecimento, particularmente nos problemas que envolvem as temperaturas do refrigerador (interna e externa) (STOECKER, 1998; VINE, 2005).

Com base nos padrões (séries temporais) reconhecidos para cada variável (curvas de carga, temperatura interna e temperatura externa), as Figuras 37-38 apresentam uma análise conjunta envolvendo as eficiências térmicas e de energia (a primeira quantificada pela COP^{-1} e a segunda quantificada pela carga) dos refrigeradores para os casos I e II. O perfil diário da temperatura externa (T_o) na região analisada em conjunto com a adoção de um ponto de ajuste para a temperatura interna (T_f) estabelece um perfil típico para o COP^{-1}, em ambos os casos. Ambos os métodos (*FCM* e STAC-M) mostram um maior consumo de energia (menor eficiência energética), antes da substituição dos refrigeradores (caso I), a fim de garantir o perfil de eficiência térmica (praticamente o mesmo nos casos I e II). Este resultado atesta o sucesso do programa de eficiência energética.

No caso I, em particular, o método STAC-M (Figura 37) foi capaz de reconhecer dois padrões (ambos associados a demanda e ao COP^{-1}), que foram mais distintas umas das outras do que no método *FCM*, evidencia a capacidade de STAC- M para reconhecer pequenas diferenças em perfis de consumo. Mais especificamente, os padrões de COP^{-1} obtidos pelo método STAC-M são diferentes durante a tarde, enquanto que o método *FCM* não identifica esta diferença. Por outro lado, a proximidade entre os padrões reconhecidos no caso II, destaca, conforme esperado, a uniformidade no comportamento de consumo devido à substituição do refrigerador antigo por um novo. Mesmo neste caso, o STAC-M foi capaz de reconhecer dois padrões de consumo que mostram pequenas diferenças entre estes.



t (h)

Figura 37 – Padrões da eficiência do motor e da demanda (caso I).

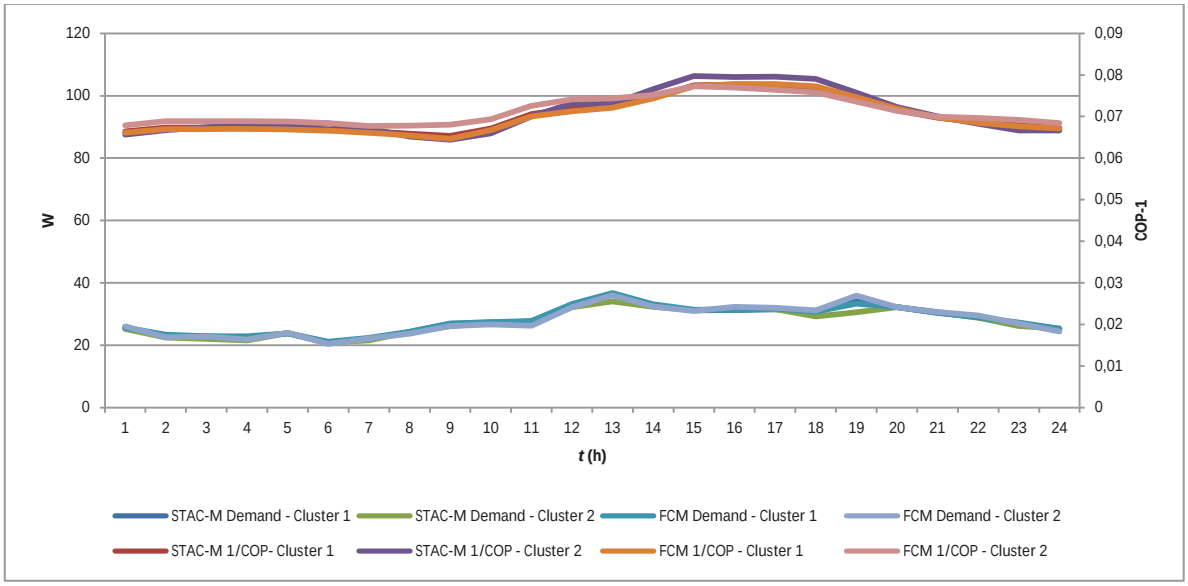


Figura 38 – Padrões da eficiência do motor e da demanda (caso II).

Os resultados mostram que o método STAC-M foi capaz de identificar uma maior diversidade de padrões de demanda e pode ser usado como uma potencial ferramenta para melhorar a tomada de decisão através de uma melhor classificação dos perfis de consumidores heterogêneos.

As Figuras 39-40 apresentam uma análise conjunta da sazonalidade, considerando apenas o grupo modal (grupo com o maior número de objetos) em cada caso. Cada objeto no grupo foi transformado em um único vetor representado por uma componente principal. Cada componente principal é uma série temporal que realiza uma possível trajetória de um espaço amostral, gerado por um processo estocástico (ABDEL-AAL, 2006).

A análise da sazonalidade no grupo pode ser realizada através da análise de autocorrelação do conjunto desses componentes principais (SAINI, 2002). A existência de pontos de inflexões nas tendências dos padrões reconhecidos (Figuras 37-38) sugere a aplicação da análise de autocorrelação nas diferenças de primeira ordem sobre estas componentes principais a fim de atenuar os efeitos não-estacionários (AGGARWAL, 2009). Os primeiros picos nas Figuras 39-40 são associados às condições não estacionárias e mostram alta autocorrelação que diminui ao longo do tempo atingindo níveis associados com condições de estado estacionário.

Nessas condições os picos ocorrem de forma intermitente (não consecutivas), mostrando o comportamento sazonal da série temporal analisada. De acordo com a Figura 39 há picos sazonais de ordem de defasagem igual a 5 (início da manhã) e 17 (final da tarde),

entre outros, que indicam mudanças no nível de consumo. Isso coincide com o pico sazonal do sistema elétrico brasileiro, possivelmente associado a hábitos de consumo, o que sugere a necessidade de ações para melhorar o perfil da demanda.

A Figura 40 (caso II) apresenta um maior número de picos, devido à maior frequência de desligamento do motor. Isto é consistente porque o novo refrigerador tem significativamente um melhor isolamento térmico, resultando num menor consumo de energia do motor (também atestada pela Figura 38).

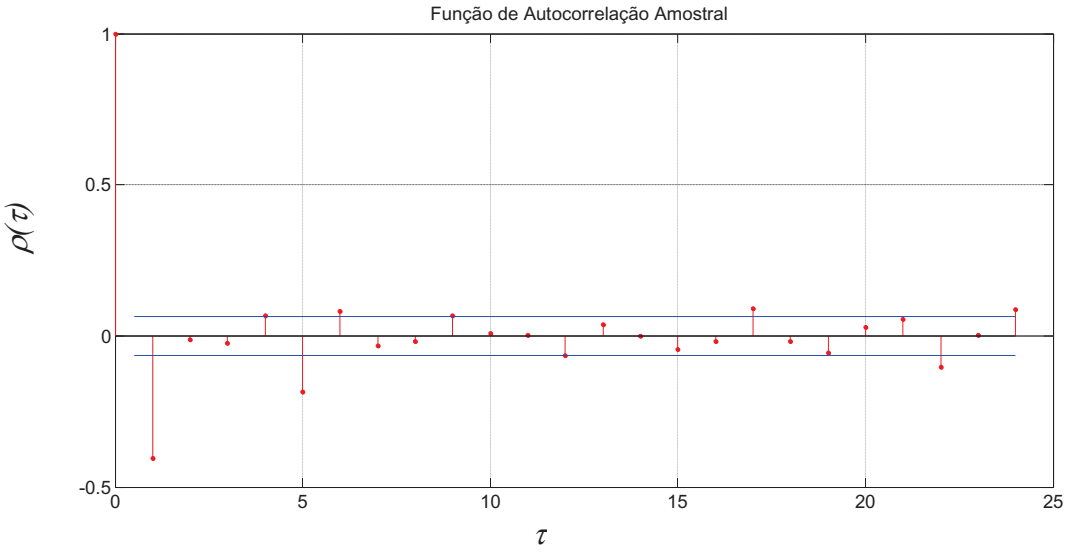


Figura 39 – Valores das autocorrelações ($\rho(\tau)$) das defasagens (τ) de primeira ordem (grupo modal - caso I).

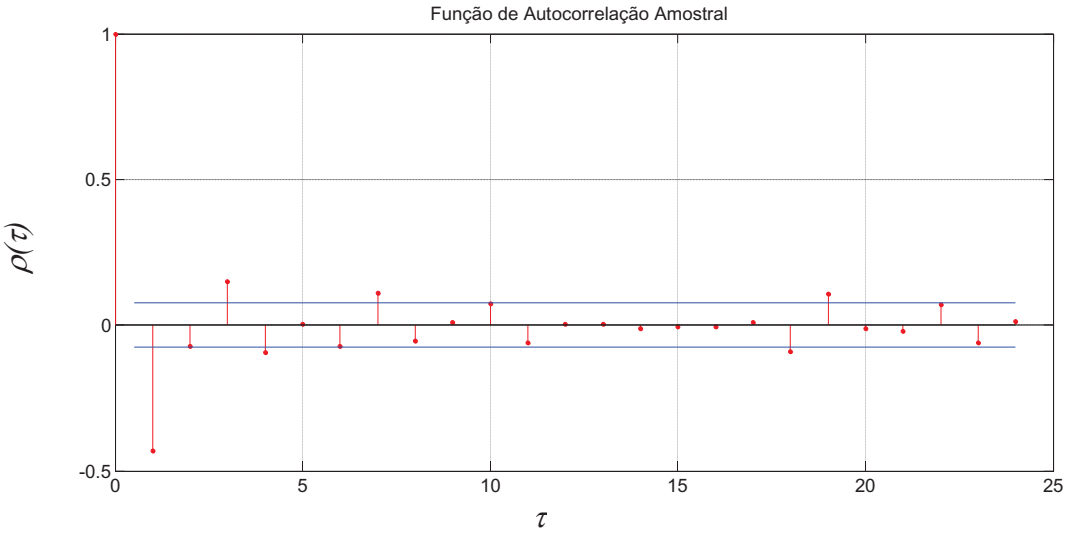


Figura 40 – Valores das autocorrelações ($\rho(\tau)$) das defasagens (τ) de primeira ordem (grupo modal - caso II).

O processo de tomada de decisão no contexto do caso em estudo está associado à avaliação do programa de eficiência energética realizada pela companhia energética. Esta avaliação, por sua vez, pode ser realizada através de um procedimento de ajuste dos padrões reconhecidos através de modelos empíricos a fim de estimar o ganho efetivo obtido pela substituição do equipamento.

O padrão (série temporal multivariada) reconhecida com o equipamento antigo (caso I, antes da substituição dos refrigeradores) pode ser aproximado por um modelo dinâmico no qual o consumo de energia (saída) é uma função do coeficiente de performance (COP^{-1}), ou seja, o consumo de energia é um efeito da eficiência térmica do equipamento.

A Figura 38 apresenta um cenário para o consumo padrão associado ao grupo modal do caso I (refrigeradores antigos), considerando os mesmos coeficientes de performance (COP^{-1}) (mesma condição de eficiência térmica) do padrão associado ao grupo modal do caso II (refrigeradores novos).

De acordo com o *IPMVP* (EVO, 2007), a linha de base (consumo ajustado) é uma referência útil para representar o padrão do grupo modal (reconhecido com base nas curvas de carga originais) e a linha de base ajustada (consumo previsto) fornece uma comparação com o padrão de perfil de consumo após a substituição dos refrigeradores, viabilizando uma avaliação objetiva do programa de eficiência energética.

Esta análise também suporta os processos de tomada de decisão ao nível da gestão, contribuindo (ou não) para expandir e consolidar o programa visando incluir outros consumidores. Além disso, o ganho ou a economia de energia pode ser útil na tomada de decisão em relação ao adiamento de gastos futuros relacionados com a expansão da oferta de energia considerando geração, transmissão e distribuição. Os programas de economia de energia também são importantes para a obtenção de subsídios do governo (ANEEL, 2008).

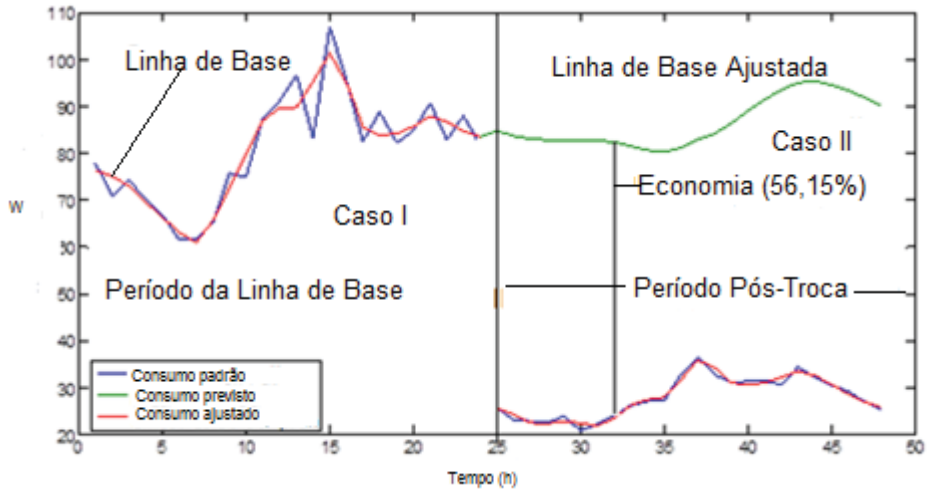


Figura 41 – Consumo padrão / Consumo ajustado (Grupo modal - Caso I e grupo modal - Caso II) e Consumo previsto refrigeradores antigos com base nas condições térmicas do Caso II.

O consumo previsto apresentado na Figura 38 foi calculado conforme um modelo *Autoregressive Exogenous - ARX* (autoregressivo com entrada exógena) (BOX et al., 2008; SADAEI et al., 2014; KRISTIANSEN, 2014; XIONG e BAO, 2014), de acordo com a Eq. (92). O procedimento de identificação também compreendeu uma pré-filtragem dos dados a fim de reconstruir o sinal com suavização de ruído adequado.

$$A(q)y(k) = B(q)u(k) + v(k) \tag{92}$$

sendo,

$$A(q) = 1 - a_1 q^{-1} - \ldots - a_{n_y} q^{-n_y}; \tag{93}$$

$$B(q) = b_1 q^{-1} + \ldots + b_{n_u} q^{-n_u}; \tag{94}$$

e q^{-1} é o operador de defasagem $\left(q^{-1} \cdot u(k) = u(k-1)\right)$, v (k) é ruído branco, u é a entrada, y é o sinal de saída, k é o instante de tempo e n é a ordem de defasagem. Neste trabalho o melhor ajuste foi alcançado com $n_y = n_u = 1$.

$$\left(1 - q^{-1}\right) \cdot COP^{-1}(k) = q^{-1} \cdot PC(k) + v(k) \tag{95}$$

onde *PC* é a entrada (consumo de energia).

Com base nos resultados e no procedimento de amostragem, a redução esperada no consumo de eletricidade fornecida pelo programa de eficiência energética realizado pela Companhia Energética de Alagoas (Brasil) seria em torno de 56%, o que representa um ganho de 250000,00 kWh, considerando o total da população analisada (cinco mil consumidores residenciais).

7 CONCLUSÃO

Este trabalho apresentou um método que incorpora dois algoritmos capazes de reconhecer padrões de séries discretas observadas em instantes igualmente espaçados, um aplicável a séries temporais univariadas e outro aplicável a séries temporais multivariadas. Os dados estão relacionados aos programas de eficiência energética implementados por duas empresas distribuidoras de energia elétrica, e em ambos os casos, a substituição de refrigeradores se referiu às residências de consumidores de baixa renda das distribuidoras.

Como referência comparativa, tanto para a versão univariada quanto para a versão multivariada, foi utilizado o algoritmo *FCM* (*Fuzzy C-Means*), sendo que, para a versão multivariada, foi adotada uma versão modificada do *FCM* baseada em uma métrica de similaridade que utiliza componentes principais (Similarity Principal Componente Analysis – SPCA).

O novo método proposto contribui para uma compreensão mais profícuados hábitos de uso do refrigerador. Os modelos concebidos permitiram uma melhor e mais confiável caracterização dos perfis típicos de consumo do refrigerador.

Esta ferramenta pode orientar a política de aporte de recursos e a gestão da operação no combate ao desperdício de energia. Do ponto de vista da modelagem matemática, este trabalho apresentou um método de reconhecimento de padrões de séries temporais usando um método agrupamento que incorpora um conjunto de métricas geométricas e estatísticas de distâncias entre séries temporais uni e multivariadas. O objeto de pesquisa foram os dados de medições do programa de eficiência energética voltado para a substituição dos refrigeradores antigos, em condições de uso subnormais, por refrigeradores novos de consumidores de baixa renda de áreas pertencentes aos estados do Maranhão e Alagoas. Este programa é uma iniciativa do governo federal, o qual desde 2009 estabeleceu a meta da troca de 10 milhões de refrigeradores antigos em todo o país. O novo método foi proposto em duas versões: uma realizou o reconhecimento de padrões temporais univariáveis visando a seleção, tipificação e agrupamento de curvas de cargas de refrigeradores de unidades consumidoras do estado de Maranhão; a outra fez o reconhecimento de padrões de séries temporais multivariadas visando a seleção, tipificação e agrupamento de curvas de cargas juntamente com as temperaturas externas e internas dos refrigeradores de unidades consumidoras do estado de Alagoas.

Na primeira versão, o método (STAC) reconheceu uma maior diversidade dos padrões de demanda horária de energia elétrica, apresentando-se como uma ferramenta potencial para a melhoria do processo de tomada de decisão através de uma melhor tipificação dos perfis de consumidores heterogêneos no setor de energia elétrica. Os resultados também demonstraram a sua superioridade em relação ao método não-hierárquico tradicional *Fuzzy C-Means* (*FCM*) quanto a auto-definição do número ótimo de grupos a ser reconhecidos, qualidade de coesão e separação dos grupos.

Este novo método de seleção, tipificação, e agrupamento univariado de curvas de carga (STAC) extraiu características baseadas em indicadores que favorecem a um melhor conhecimento dos hábitos de consumo da energia elétrica. O método STAC foi capaz de identificar uma maior diversidade nos padrões de demanda e também representa uma ferramenta potencial para a melhoria do processo de tomada de decisão através de uma melhor classificação dos perfis de consumidores heterogêneos no setor de energia elétrica. Vale destacar que este método é adequado para dados numéricos e sua aderência ao setor elétrico o torna uma ferramenta potencial para aplicações nesta área. Os resultados demonstram o seu bom desempenho no reconhecimento de padrões em amostras com dados heterogêneos (situação comum no setor elétrico). Ao contrário dos modelos *c-means* de agrupamento, o número de grupos é também um resultado obtido pelo método STAC.

O caso estudado analisou um programa de eficiência energética realizado pela Companhia Energética do Maranhão (CEMAR - Brasil), que analisou o impacto da substituição de 5250 refrigeradores velhos por novos para os consumidores de baixa renda. Vale reiterar que, o conceito de eficiência energética adotado nesta situação, de acordo com a (*ASHRAE*, 2002), foi o montante de energia final economizado para realizar o serviço energético ofertado (frio). Os resultados obtidos pelo algoritmo do método STAC, em comparação com um algoritmo bem conhecido de agrupamento (*FCM*), revelam a viabilidade e o potencial no reconhecimento de padrões e na geração de conclusões coerentes com a realidade do setor de energia elétrica. Isso apoia a implementação de ações de eficiência baseados em características reais dentro do mercado consumidor e também pode apoiar a tomada de decisões ao nível da gestão.

Por sua vez a adaptação para o caso multivariado (método STAC-M) foi também capaz de identificar uma maior diversidade de padrões, mostrando o potencial deste método no melhor reconhecimento de padrões de consumo, considerando o efeito de outras variáveis

simultaneamente. Os resultados foram comparados com o *FCM* modificado para agrupamento de séries temporais multivariadas e também foi constatada a superioridade do STAC-M, considerando os mesmos critérios adotados na comparação de desempenho do STAC. O método proposto incorpora vários métricas de distâncias de objetos, aumentando o poder de reconhecimento de padrões das curvas de carga, ao contrário das abordagens tradicionais que essencialmente se baseiam em uma única métrica de distância entre os objetos.

Neste aspecto, o novo método faz a seleção, tipificação e agrupamento de curvas de cargas e séries temporais multivariadas correlatas (STAC-M) adequadas para o setor de energia elétrica, e em especial, a análise de perfis de consumo associados com os equipamentos de refrigeração. O algoritmo é apropriado para dados sem rótulo e compreende quatro etapas que extraem características essenciais da série temporal multivariada de usuários residenciais, com ênfase no perfil sazonal e temporal, entre outros. Os resultados demonstraram o seu bom desempenho no reconhecimento de padrões em amostras com dados heterogêneos (situação comum no setor de energia elétrica). Ao contrário dos processos de tipificação dos modelos de agrupamento *c-means*, o número de grupos é também um resultado obtido pelo método STAC-M.

O método STAC-M foi capaz de reconhecer diferentes grupos usando vários critérios para o reconhecimento de padrões que são difíceis de identificar utilizando métodos tradicionais, e a qualidade de agrupamento considera os níveis de coesão e de separação entre os grupos. A abordagem de múltiplos critérios também contribui para o reconhecimento de sazonalidades na série temporal que é útil para a identificação de modelos de previsão para apoiar a tomada de decisão relacionada com a implementação de programas de eficiência energética. Por outro lado, a segunda fase do método STAC-M compreende a formação de grupos através de apenas um critério geométrico (distância).

O caso estudado se referiu a um programa de eficiência energética realizado pela Companhia Energética de Alagoas (CEAL-Brasil), que analisou o impacto da substituição de 5000 refrigeradores antigos por novos para os consumidores de baixa renda. Os resultados obtidos pela STAC-M, em comparação com um método bem conhecido de agrupamento (*Fuzzy C-Means*, *FCM*), revelam a viabilidade e o potencial no reconhecimento de padrões e na geração de conclusões coerentes com a realidade do setor de energia elétrica. Isso apoia a implementação de ações de eficiência com base em características reais dentro do mercado consumidor e também pode apoiar a tomada de decisões ao nível da gestão.

A capacidade de incluir temperaturas interna e externa, juntamente com as curvas de carga dos refrigeradores fornece a possibilidade de analisar em conjunto as eficiências térmicas e de energia do equipamento e identificar o comportamento anormal de consumo associado à temperatura ambiente e do desempenho de seu controle térmico.

Vale reiterar que, o processo de agrupamento do método proposto, em ambas versões, é constituído de um conjunto de técnicas matemáticas/estatísticas para caracterizar e medir as relações dos objetos, por meio de algoritmos de agrupamento de series temporais que aplicam os modelos e testes especialmente desenvolvidos para aumentar a qualidade de agrupamento. Destarte, ressalta-se que, antes das substituições (caso I), o STAC e o *FCM* tiveram, respectivamente, um índice global de silhueta (medida da qualidade de agrupamento) de 0,28 e 0,25. Após as substituições (caso II) dos refrigeradores, ambos métodos reconheceram a existência de apenas um grupo e padrões semelhantes. Na versão multivariada, no caso I, o STAC-M teve o índice global de silhueta de 0,19, e no caso II, o índice global de silhueta foi de 0,46. Os índices obtidos pelo método *FCM* modificado, no caso I e no caso II, respectivamente, foram de -0,12 e 0,21. O método proposto em ambas versões apresentaram uma identificação de uma maior diversidade de padrões; o reconhecimento da sazonalidade através de uma abordagem multicritérios; o melhoramento da tomada de decisão através de uma melhor classificação dos perfis de consumidores heterogêneos; e a definição do número de clusters através de uma abordagem baseada em grupos semi-hierárquica, revelando-se assim como uma importante contribuição para o estado-da-arte.

O novo método propicia maior facilidade para uma análise quantitativa dos hábitos de consumo de energia elétrica e, de forma complementar às informações tradicionais do mercado, indica padrões cujas violações do uso racional de energia elétrica não justificam certos investimentos. Vale a pena dizer que, através de pesquisa, cada vez mais se tem conquistado uma abordagem mais consistente de reconhecimento dos padrões ideais de consumo. Estas análises mais precisas possibilitam o progresso mais correto da criação de boas medidas de economia de energia.

A partir dos desafios e resultados obtidos neste trabalho, são sugeridas as seguintes possibilidades de trabalhos futuros:

- Desdobramento do método proposto com a incorporação de um procedimento de identificação de sistemas após os reconhecimentos de grupos e padrões

realizados, permitindo obter modelos de previsão para diferentes dinâmicas temporais;

- Análise do impacto do período de amostragem e da incerteza de medição (consumo e temperaturas) sobre a qualidade dos grupos e padrões reconhecidos.

REFERÊNCIAS

ABDEL-AAL, R.E. *Modeling and forecasting electric daily peak loads using abductive networks.* ***Electrical Power and Energy Systems***, v. 28, p. 133-141, 2006.

ABONYI, J.; FEIL, B.; NEMETH, S.; ARVA, P. *Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series.* ***Fuzzy Sets Syst***. 149, p. 39–56, 2005

AGGARWAL, S. K.; SAINI, L. M.; K, A. *Electricity price forecasting in deregulated markets: A review and evaluation.* ***Electrical Power and Energy Systems***, v. 31, p. 13-22, 2009.

ANEEL. Lei 9.991/2000, de 24 de julho de 2000;

_____. Manual para Elaboração do Programa de Eficiência Energética – MPEE. Brasília, 2008;

_____. Lei 12.212/2010, de 20 de janeiro de 2010;

ANUAR, N.; ZAKARIA, Z. *Cluster Validity Analysis for Electricity Load Profiling.* ***IEEE International Conference on Power and Energy****, Kuala Lumpur Malaysia*, p. 35-38, 2010.

ARBELAITZ, O.; GURRUTXAGA, I.; MUGUERZA, J.; PÉREZ, J. M.; PERONA, I. *An extensive comparative study of cluster validity índices.* ***Pattern Recognition*** v. 46, p. 243-256, 2013.

ASHRAE. *American Society of Heating, Refrigerating and Air-Conditioning Engineers*

- *ASHRAE Guideline 14-2002: Measurement of Energy and Demand Savings*, 2002.

BARAGONA R. A *simulation study on clustering time series with meta-heuristic methods.* ***Quad. Stat***. 3,p. 1–26, 2001.

BATRINU, F., G. CHICCO, R. NAPOLI, F. PIGLIONE, M. SCUTARIU, P. POSTOLACHE AND C. TOADER. *Efficient iterative refinement clustering for electricity customer classification. In Proc.* ***IEEE power Tech***, St. Petersburg, Russia, p. 1-7, 2005.

BÉCAVIN, C., TCHITCHEK, N.; MINTSA-EYA, C.; LESNE, A.; B, A. *Improving the efficiency of multidimensional scaling in the analysis of high dimensional data using singular value decomposition.* ***Bioinformatics*** v. 27, n. 10, p. 1413-1421, 2011.

BENSAID, A.; HALL, L. O.; BEZDEK JAMES. C.; CLARKE, L. P. *Partially Supervised Clustering for Image Segmentation.* ***Pattern Recognition,*** v. 29 n. 5, p. 859-870, 1996.

BEZDEK, JAMES C.; KELLER, J.; KRISNAPURAM, R.; PAL, N. R. ***Fuzzy Models and Algorithms for pattern recognition and image process.*** Springer: New York, NY, 2005.

BIERNACKI, C.; CELEUX, G.; GOVAERT, G. *Assessing a mixture model for clustering with the integrated completed likelihood.* ***IEEE Trans. Pattern Anal***. Mach. Intell. 22,p. 719 –725, 2000.

BOX, G.E.P.; JENKINS, G.M.; REINSEL, G. C. ***Times Series Analysis***- *Forecasting and Control.4th ed., John Wiley & Sons*, *Inc*.:, 2008.

BUSSAB W. O.;MORETTIN, P. A. **Estatística básica**. 5ª ed., EditoraSaraiva: São Paulo, 2006.

CAMARGO, C. C. BRASIL. **Transmissão de Energia Elétrica –Aspectos Fundamentais**. 3ª ed. Editora da UFSC: Florianópolis, 2006.

CAMPELLO, R. J. G. B. e HRUSCHKA, E. R. A *fuzzy extension of the silhouette widthcriterion for cluster analysis.* ***Fuzzy Sets and Systems***, 157: 2858 – 2875, 2006.

CARPANETO, E.; CHICCO, G., NAPOLI, R.; SCUTARIU, M. *Electricity customer classification using frequency-domain load pattern data.* ***Electrical Power Energy System*** v.28 n. 1, p. 13-20, 2006.

CHAOVALIT, P.; GANGOPADHYAY, A.; KARABATIS, G.; CHEN, Z. Discrete Wavelet Transform-Based Time Series Analysis and Mining. ***Journal ACM Computing Surveys (CSUR)*** Surveys Homepage archive, 2011; 43(2): 6:1-6:37.

CHERKASSKY, V. MULIER, F. ***Learning From Data: Concepts, Theory, and Methods***. *Wiley-Interscience*, 1998.

CHICCO, G.; NAPOLI, R.; POSTOLACHE, P.; SCUTARIU, M.; TOADER, C. *Customer characterization options for improving the tariff offer.* **IEEE Trans. Power System** v.18 n. 1, p. 381-387, 2003.

CHICCO, G.; NAPOLI, R.; POSTOLACHE, P.; SCUTARIU, M.; TOADER, C. *Load pattern-based classification of electricity customers.* ***IEEE Trans. Power System***. v. 19 n. 2, p. 1232-1239, 2004.

CHICCO, G.; NAPOLI, R.; POSTOLACHE, P.; SCUTARIU, M.; TOADER, C. *Emergent electricity customer classification.* ***IEEE Proc Generation Transmission Distribution*** v.152 n. 2, p. 164-72, 2005.

CHICCO, G., NAPOLI, R.; PIGLIONE, F. *Comparisons among clustering techniques for electricity customer classification.* ***IEEE Trans. Power System*** v.21 n. 2, p. 933-940, 2006.

CHICCO, G.;ILIE, I. S. *Support vector clustering of electrical load pattern data.* ***IEEE Trans. Power System,*** v.24 n. 3, p. 1619-1628, 2009.

CHICCO, G.; SUMAILI, AKILIMALI J. *"Renyi entropy-based classification of daily electrical load patterns".****Generation Transmission Distribution***, v.4 n. 6, p. 736-745, 2010.

CHICCO, G. *Overview and Performance of the Clustering Methods for Electrical Load Pattern Grouping".****Energy*** v.42 n. 1, p. 68-80, 2012.

CHIU, S.L. *A cluster estimation method with extension to fuzzy model identification. In* ***Proceedings of the Third IEEE Conference on Fuzzy Systems*** Orlando –Florida, USA: p. 1240-1245, 1994.

CLINCH, J. P.; HEAL, J. D. *Cost-benefit analysis of domestic energy efficiency.* ***Energy Policy*** v.29, p. 113-124, 2001.

COCHRAN, William G. ***Sampling techniques***. 3th ed. New York: Wiley, 428 p., 1977.

COPPI, R., D' URSO, P.; GIODANI, P.A *Fuzzy Clustering Model Multivariate Spatial Time Series.* ***Journal of Classification*** v. 27,p. 54-88, 2010.

D'URSO, P.; A., ELIZABETH; J., MAHARAJ. *Wavelets-base clustering of multivariate time series.* ***Fuzzy Sets and System,*** v. 193,p. 33-61, 2012.

DAIGO, M. *Factor analysis and pattern decomposition method*. ***SPIE***, p. 1-8, 2005.

DASWIN, S.; XINGHUO, Y. A *Data Mining Framework for Electricity Consumption Analysis From Meter Data.* ***IEEE Transactions on industrial informatics***, vol. 7, nº 3, august, p. 399-407, 2011.

DAVIES, D.L.; BOULDIN, D.W. *A cluster separation measure.* ***IEEE Trans. Pattern Anal. Machine Intell.*** 1 (4), p. 224-227, 1979.

DOBOS, L.; ABONYI, J. *On-line detection of homogeneous operation ranges by dynamic principal component analysis based time-series segment.* ***Chemical Engineering Science***, 2012; 75: p 96-105.

DUNN, J.C. *Well separated clusters and optimal fuzzy partitions.* ***Journal of Cybernetics***, v. 4, p. 95-104, 1974.

EVO. Protocolo Internacional de Medição e Verificação do Desempenho Energético – PIMVP, 2007.

FERREIRA, A. M. S.; MARAMBIO, J. E. S.; LUZ, A. D.; CHAGAS, E. H. C. ; C. MUCCINI, M. J. ; SOARES JR, F. A. ; SANTOS, S. O. . Metodologia para Planejamento e Acompanhamento de Programas de GLD em Mercado com Crescimento não Tradicional. In: II Congresso de Inovação Tecnológica em Energia Elétrica, 2003, Salvador. II Congresso de Inovação Tecnológica em Energia Elétrica, 2003.

FERREIRA, ADONIAS M. S.; CAVALCANTE, CARLOS A. M. T.; FONTES, CRISTIANO H. O. ;MARAMBIO, JORGE E. S. O setor elétrico brasileiro. **Politécnica – Revista do Instituto Politécnico da Bahia**, v. 3, p. 18-24, 2010.

FERREIRA, ADONIAS M. S.; CAVALCANTE, CARLOS A. M. T.; FONTES, CRISTIANO H. O. ; MARAMBIO, JORGE E. S. Um Novo Método de Tipificação da Demanda Horária de Energia Elétrica. In: **XXXI Encontro Nacional de Engenharia de Produção - ENEGEP**, Belo Horizonte 2011.

FERREIRA, A. M. S.; FONTES C. H. O.; CAVALCANTE C. A. M. T.; MARAMBIO, J. E. S. M. *A New Proposal of Typing Load Profiles to Support the Decision Making in the Sector of Electricity Energy Distribution.* ***In International Conference on Industrial Engineering and Industrial Management (ICIEOM)***, p. 18.1-18.7, 2012.

FERREIRA, ADONIAS M. S.; CAVALCANTE, CARLOS A. M. T.; FONTES, CRISTIANO H. O. ; MARAMBIO, JORGE E. S. *A new method for pattern recognition in*

load profiles to support decision-making in the management of the electric sector. ***International Journal of Electrical Power & Energy Systems*** JCR, v. 53C, p. 824-831, 2013a.

FERREIRA, ADONIAS M. S.; CAVALCANTE, CARLOS A. M. T.; FONTES, CRISTIANO H. O.;MARAMBIO, JORGE E. S. *Pattern recognition of Load Profiles in Managing Electricity Distribution.* ***International Journal of Industrial Engineering and Management****,* v. 4 n. 3, 2013b.

FERREIRA, ADONIAS M. S.; CAVALCANTE, CARLOS A. M. T.; FONTES, CRISTIANO H. O. ; MARAMBIO, JORGE E. S. *Pattern recognition as a tool to support decision making in the management of the electric sector. Part II: a new method based on clustering of multivariate time series.* ***International Journal of Electrical Power & Energy Systems*** JCR, v. 67, p. 613-626, 2015.

FIGUEIREDO, V.; RODRIGUES, F.; VALE,Z.; GOUVEIA, J. B. *An electric energy consumer characterization framework based on data mining techniques.* ***IEEE Trans. Power System*** v.20 n. 2, p. 596–602, 2005.

FONTES, C. H. O; CAVALCANTE, C. A. M. T.; PEREIRA, O. J.; BARRETO, S. S.; PACHECO, L. A.;EMANUEL, W. *Pattern Recognition using Multivariable Time Series for Fault Detection in a Thermoeletric Unit.* ***Computer-Aided Chemical Engineering,*** v.31, p. 315-319, 2012.

FU, T.-C.; CHUNG, F.-L.; NG, V.; LUK, R. *Pattern discovery from stock time series using self-organizing maps.* ***KDD 2001 Workshop on Temporal Data Mining***, August 26–29, San Francisco, p. 27–37, 2001.

GAN, G.; MA, C.; WU, J. ***Data Clustering: Theory, Algorithms and Applications****, ASA-SIAM Series on Statistics and Applied Probability*, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.

GATH, I.; GEVA, A. B. *Unsupervised optimal fuzzy clustering.* ***IEEE Trans. Pattern Anal. Mach. Intell***, p. 773-781, 1989.

GELLER, H.; JANNUZZI, G. M.; SCHAEFFER, R.; TOLMASQUIM, M. T. *The efficient use of electricity in Brazil: progress and opportunities.* ***Energy Policy*** v.26 n. 11, p. 859-872, 1998.

GEMINAGNANI, M. M. F.; OLIVEIRA, C. C. B; TAHAN, C. M. V. *Proposition and Comparative Analysis of Alternative Selection and Classification of Load Curve for Defining Types for tariff studies.* ***Décimo Tercer Encuentro Reginal Iberoamericano de Cigré– XIII ERIAC***, p. 1-6, 2009.

GERBEC, D., GASPERIC, S.; SMON, I.; GUBINA, F. *A methodology to classify distribution load profiles.* ***IEEE Trans. Power System***, v.1, p. 848-851, 2002.

GERBEC, D.; GASPERIC, S., SMON, I.; GUBINA, F. *Determining the load profiles of consumers based on fuzzy logic and probability neural networks.* ***IEEE Proc.-Generation Transmission Distribution***, v.151 n. 3, p. 395-400, 2004.

GERBEC, D., S. GASPERIC, SMON, I.;GUBINA, *F. Allocation of the load profiles to consumers using probabilistic neural networks.* ***IEEE Trans. Power System***, v.20 n. 2, p. 548-555, 2005.

GOUTTE, C.; TOFT P.; ROSTRUP, E. *On clustering fMRI time series.* ***Neuroimage***, 9 (3), p. 298–310, 1999.

GOUTTE C.; HANSEN, L.K.; LIPTROT, M.G.; ROSTRUP, E. *Feature-space clustering for fMRI meta-analysis, Hum.* ***Brain Mapping*** 13,p. 165 –183, 2001.

GOLDMAN, C. A.; HOPPER, N. C.; OSBORN, J. G. *Review of US ESCO industry market trends: an empirical analysis of project data.* ***Energy Policy,*** v. 33, p. 387-405, 2005.

GOODMAN, L., KRUSKAL, W. *Measures of associations for cross-validations.* ***J. Am. Stat. Assoc.****,* v. 49, p. 732-764, 1954.

GOLAY, X.; KOLLIAS, S.; STOLL, G.; MEIER, D.; VALAVANIS, A.; BOESIGER, P. *A new correlation-based fuzzy logic clustering algorithm for fMRI.* ***Mag. Resonance Med***. 40, p. 249-260, 1998.

HAIR, J. F.; BLACK, B.; BABIN, B.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. 5ª ed. Bookman: Porto Alegre, 2005.

HALKIDI, M.; BATISTAKIS, Y; VAZIRGIANNIS, M. *On clustering validation techniques,* ***Journal of Intelligent Information Systems****,* v. 17, p. 107-145, 2001.

HAN, J.; PEI, J.; YIWEN, Y. *Mining Frequent Patterns Without Candidate Generation.* ***Proceedings ACM-SIGMOD International Conference on Management of Data****,* ACM Press, p. 1-12, 2000.

HILLIER, F. S.; LIEBERMAN, G. J. **Introdução à pesquisa operacional**. Editora Campus, São, 1988.

HOPNNER, Frank; KLAWOON, Frank; KRUSE, Rudolf; RUNKLER, Thomas.***"Fuzzy Cluster Analysis - Methods for Classification, Data Analysis and Image Recognition"***, John Wiley & Sons, LTD, 2000.

HUBERT, L., SCHULTZ, J. *Quadratic assignment as a general data-analysis strategy.* ***British Journal of Mathematical and Statistical Psychologie***, v. 29, p. 190-241, 1976.

HUBERT, M; VANDERVIEREN, E. *An adjusted boxplot for skewed distributions.* ***Computational Statistics and Data Analysis***, v. 52, p. 5186-5201, 2008.

JACCARD, P. *The distribution of flora in the alpine zone.* ***New Phytologist***, v. 11, p. 37–50, 1912.

JAIN, A.K.; MURTY, M. N.; FLYNN, P. J. *Data Clustering: A Review.* ***ACM Computing Surveys***, v. 31 n. 3, p. 264-323, 1999.

JAIN, A.K.; MURTY, M.N. ; FLYNN, P.J. ***Data Clustering: A review****. Columbus. ACM Computing Surveys*, Vol. 31, n° 3, Setembro 1999.

JANES, JOSEPH. *Categorical relationships: chi-square.* ***Library Hi Tech***, v. 19 n. 3, p. 296-298, 2001.

JOHNSON, R.A.; WICHERN, D.W. ***Applied Multivariate Statistical Analysis****. 6th edition*, Pearson: New Jersey, 2007.

JUNIOR, L. C. Z. **Fundamentos de Sistemas Elétricos de Potência**. Editora Livraria da Física:São Paulo, 2006.'

KAKIZAWA, Y.; SHUMWAY, R. H.; TAMIGUCHI, N. *Discrimination and clustering for multivariate time series.* ***J. Amer. Stat. Assoc***. 93 (441). p. 328-340, 1998.

KALPAKIS, K.; GADA, D.; PUTTAGUNTA, V. *Distance measures for effective clustering of ARIMA time-series.* ***Proceedings of the 2001 IEEE International Conference on Data Mining***, San Jose, CA, November 29–December 2, p. 273–280, 2001.

KAUFMAN, L.; ROUSSEEUW, P.J. ***Finding Groups in Data: An Introduction to Cluster Analysis***. Wiley, New York, 1990.

KAVITHA, V.; PUNITHAVALLI, M. Clustering Time Series Data Stream – *A Literature Survey.* ***International Journal of Computer Science and Information Security***, v. 8, p. 289-294, 2010.

KEOGH, E.J.; KASETTY, S. *On the need for time series data mining benchmarks: a survey and empirical demonstration.* ***Data Mining and Knowledge Discovery***, v.7 n. 4, p. 349-371, 2003.

KHALIL, B.; OUARDA,T. B. M. J.; ST-HILAIRE, A. *Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis.* ***Journal of Hydrology,*** v. 405, p. 277–287, 2011.

KOŠMELJ, K.; BATAGELJ, V. *Cross-sectional approach for clustering time varying data.****J. Classification 7*** (1990) 99-109.

KRISTIANSEN, T. *A time series spot price forecast model for the Nord Pool market.* ***International Journal of Electrical Power & Energy Systems***, 2014; 61:20-26.

KUMAR, M. PATEL, N. R.; WOO, J. *Clustering seasonality patterns in the presence of errors.* ***Proceedings of KDD***, Edmonton, Alberta, Canada, 2002.

LAMEDICA, R.; FRACASSI, G.; MARTINELLI, G.; PRUDENZI, A.; SANTOLAMAZZA, L. *A novel methodology based on clustering techniques for automatic processing of MV feeder daily load patterns*. *In Proc.* ***IEEE PES Summer Meeting***, Seattle, WA, p. 96-101, 2000.

LI, C., BISWAS, G. *Temporal pattern generation using hidden Markov model based unsupervised classification.* ***In: D.J. Hand, J.N. Kok, M.R. Berthold (Eds.), Lecture Notes in Computer Science***, vol. 164, IDA '99, Springer, Berlin, p. 245–256, 1999,.

Li, C.; Biswas, G.; Dale, M.; Dale, P. *Building models of ecological dynamics using HMM based temporal data clustering—a preliminary study.* ***In: F. Hoffmann et al. (Eds.), IDA 2001, Lecture Notes in Computer Science***, vol. 2189, p. 53–62, 2001.

LI, S.; WEN, J. *Application of pattern matching method for detecting faults in air handling unit system,* ***Automation in Construction***, 2014;43: 49-58.

LIAO, T. WARREN. *Clustering of time series data-a survey.* ***Pattern Recognition,*** v. 38 n. 11, p. 1857-1874, 2005.

LIAO, T. WARREN; BOLT, B.; FORESTER, J.; HAILMAN, E.; HANSEN, R.C. KASTE; O'MAY, J. *Understanding and projecting the battle state.* ***23rd Army Science Conference***, Orlando, Fl, p. 2-5, 2002

LIN, J.K.; TSO, S. K.; HO, H. K.; MAK, C. M.; YUNG, K. M.; HO, Y. K. *Study of climatic effects on peak load and regional similarity of load profiles following disturbances based on data mining.* ***Electrical Power and Energy Systems***, v.28, p. 177-185, 2006.

MAHARAJ, E.A. Clusters of time series. **J. Classification** 17,p. 297–314, 2000.

MARQUES, D. Z.; ALMEIDA, K. A.; DEUS, A. M.; PAULO, A. R. G.; SILVA, L. A. *Comparative analysis of neural and fuzzy cluster techniques applied to the characterization of electric load in substations. Proc.* ***IEEE/PES Transmission and Distribution, in Conference and Exposition: Latin America***, p. 908-913, 2004.

MINGOTI, S. A.; **Análise de Dados Através de Métodos de Estatística Multivariada uma Abordagem Aplicada**. Editora UFMG: Minas Gerais, 2005.

MITSA, T.. *Temporal Data Mining. Boca Raton, FL : Chapman & Hall/CRC* ***Data Mining and Knowledge Discovery Series***, 2010.

MÖLLER-LEVET, C. S.; KLAWOMONN, F. CHO. WOLKENEN, O. *Fuzzy clustering of short time series and unevenly distributed sampling points.* ***Proceeding of the 5thInternatonal Symposium on intelligent Data Analysis***, p. 28-30.

MONEDERO, I.; BISCARRI, F.; LE´ON, C.; BISCARRI, J.; MILL´AN, R. Midas: *Detection of nontechnical losses in electrical consumption using neural networks and statistical techniques.* ***In Proceedings of the International Conference on Computational Science and Applications****, Springer Berlin/Heidelberg: Lecture Notes in Computer Science*, p. 725-734, 2006.

MORETTIN; P. A.; TOLOI, C. M. **Análise de Séries Temporais**. 2ª ed., Edgard Blücher: São Paulo, 2006.

NAGI, J.; MOHAMMAD, A.; YAP, K.; TIONG, S.; AHMED, S.; *Nontechnical loss analysis for detection of electricity theft using support vector machines. In* ***Proceedings of the 2nd IEEE International Power and Energy Conference***, p. 907–912, 2008.

NAZARKO, JOANICJUSZ; STYCZYNSKI, ZBIGNIEW A. *Application of statistical and neural approaches to the daily load profiles modeling in power distribution systems.* ***IEEE***, p. 320-325, 1999.

NAZARKO, J.; JURCZUK, A.; ZALEWSKI, W. *ARIMA models in load modeling with clustering approach. Proc.* ***IEEE power Tech****, in St. Petersburg, Russia*, p. 27-30, 2005.

NIZAR, A. H.; DONG, Z. Y.; ZHAO, J. H. *Load profiling and data mining techniques in electricity deregulated market. In* ***Presented at the IEEE Power Engineering Society (PES) General Meeting***, 2006, Montreal, Quebec, Canada, p. 1-7, 2006.

NIZAR, A. H.; DONG, Z. Y.; WANG, Y. *Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Model.* ***IEEE Trans. on Power Systems,*** v.23 n. 3, p. 946-955, 2008.

OATES, T.; FIROIU, L.; COHEN, P.R. *Clustering time series with hidden Markov models and dynamic time warping.* ***Proceedings of the IJCAI-99 Workshop on Neural, Symbolic, and Reinforcement Learning Methods for Sequence Learning***, 1999.

O'GORMAN, T. W. *A comparison of an adaptive two-sample test to the t-test, rank-sum, and log-rank tests* .***Communications in Statistics - Simulation and Computation***, v. 26, p. 1393-1411, 1997.

OWSLEY,L. M. D.; ATLAS, L. E.;BERNARD,G. D. *Self-organizing feature maps and hidden Markov models for machine-tool monitoring.* ***IEEE Trans. Signal Process.*** 45 (11),p. 2787–2798,1997.

PAL, N.; BISWAS, J. *Cluster validation using graph theoretic concepts.* ***Pattern Recognition***, v. 30, n. 6, p. 847-857, 1997.

PAUWELS, E.J.; FREDERIX, G., *Finding salient regions in images: nonparametric clustering for image segmentation and grouping.* ***Computer Vision and Image Understanding****,* v. 75, p. 73-85, 1999.

PIATETSKY, G. *Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics".* ***Data Mining Knowledge Discovery*** 15, p. 99-105, 2007.

PICCOLO, D. *A distance measure for classifying ARMA models.* ***J. Time Ser.*** *Anal.* 11 (2), p. 153 –163,1990.

RAMONI, M.; SEBASTIANI, P.; COHEN, P. *Multivariate clustering by dynamics.* ***Proceedings of the 2000 National Conference by dynamics, on Artificial Intelligence (AAAI-2000)****.* San Francisco-CA, p. 633–638, 2000.

RAMONI, M.; SEBASTIANI, P.; COHEN, P. *Bayesian clustering by dynamics.* ***Mach. Learning*** 47 (1), p. 91– 121, 2002.

RAMOS, S.; VALE, Z.; SANTANA, J.; DUARTE, J. *Data mining contributions to characterize MV consumers and to improve the suppliers-consumers settlements* . ***In Proc IEEE/PES*** *Gen Meeting*, p. 24-28, 2007.

POLICKER, S.; GEVA, A. B. *Nonstacionary time series analysis by temporal clustering. IEEE Trans. Syst. Man Cybernet,* 30 (2), p. 339-343, 2000.

RAND, W.M. *Objective criteria for the evaluation of clustering methods.* ***Journal of the American Statistical Association.***, p. 846-850, 1971.

RANI, S.; SIKKA, G. *Recent Techniques of Clustering of Time Series Data: A survey.* ***International Journal of Computer applications,*** v.52, p. 1-9, 2012.

RÄSÄNEN, T.; VOUKANTSIS, D.; NISKA, H.; KARATZAS, K.; KOLEHMAINEN, M. *Data-base method for creating electricity use load profiles using large amount ofcustomer-specific hourly measured electricity use data.* ***Applied Energy,*** v.87 n. 11, p. 3538-3545, 2010.

REIS, E. **Estatística Multivariada Aplicada**. 2ª ed., Edições Silabo: Lisboa – Portugal, 2001.

ROUSSEEUW, P.J. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* ***Journal of Computational and Applied Mathematics,*** v.20, p. 53-65, 1987.

RUEDA, L.; HERRERA, MYRIAM. *"A New Linear Dimensionality Reduction Technique Based on Chernoff Distance". IBERAMIA-SBIA*, p. 299-308, 2006.

SAINI, L. M; SONI, M. K. *Artificial neural network-based peak load forecasting using conjugate gradient methods.* ***IEEE Trans Power System***, v. 17, p. 907-12, 2002.

SADAEI, H. J.; ENAYATIFAR, R.; ABDULLAH, A. H.; GANI, A. *Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search.* ***International Journal of Electrical Power & Energy Systems***, 2014; 62:118-129.

SEPPÃLÃ, A. *Statistical distribution of customer load profile.* ***IEEE,*** v.95, p. 696-701, 1995.

SHAW, C.T.; King, G.P. *Using cluster analysis to classify time series.* ***Physica*** D 58,p. 288 – 298,1992.

SHUMWAY, R. H. *Time-frequency clustering and discriminant analysis.* ***Stat. Probab. Lett.*** 63, p. 307-314, 2003.

SILVA, D.; YU, X. A; ALAHAKOON, D.; HOLMES, G. *Data Mining Framework for Electricity Consumption Analysis From Meter Data.* ***IEEE Transactions on industrial informatics,*** v. 7 n. 3, p. 399-407, 2011.

SINGHAL, A.; D. E. SEBORG. *Clustering Multivariate Time-series Data.* ***Journal of Chemometrics***, v.19, p. 427-438, 2005.

STENVENSON, W. D. **Elementos de Análise de Sistemas de Potência**. McGraw-Hill: São Paulo, 1986.

STOECKER, WILBERT F. ***Industrial refrigeration handbook***. McGraw-Hill:New York, 1998.

THEODORIDIS, S.; KOUTROUBAS, K. ***Pattern Recognition***. *London*: *Academic Press*, 1999.

TOPCHY, A.; Jain, A.; Punch, W. *Combining multiple weak clusterings. Proc. Third IEEE* ***International Conference on Data Mining (ICDM'03)****,* p. 331-338, 2003.

TRAN, D., WAGNER, M. *Fuzzy c-means clustering-based speaker verification.* ***In: N.R. Pal, M. Sugeno (Eds.), AFSS 2002, Lecture Notes in Artificial Intelligence***, 2275, p. 318–324, 2002.

TREBUNA, P.; HALCINOVÁ, J. *Mathematical Tools of Cluster Analysis.* ***Applied Mathematics****,* v. 4, p. 814-816, 2013.

TSEKOURAS, G. J.; HATZIARGYRIOU, N. D.; DIALYNAS, E. N. *Two-stage pattern recognition of load curves for classification of electricity customers.* ***IEEE Trans. Power System,*** v.22 n. 3, p. 1120-1128, 2007.

TSEKOURAS, G. J.; KOTOULAS, P. B.; TSIREKIS, C. D.; DIALYNAS, E. N.; HATZIARGYRIOU, N. D. *A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers.* ***Electric Power System*** *Res* v. 78 n. 9, p. 1494-1510, 2008.

TSEKOURAS, G. J.; TSAROUCHA, M. A.; TSIREKIS, C. D.; SALIS, A. D.; DIALYNAS, E. N.; HATZIARGYRIOU, N. D. *A database system for power systems customers and energy efficiency programs.* ***Electrical Power and Energy Systems,*** v.33, p. 1220–1228, 2011.

VALERO, S.; ORTIZ, M.; SENABRE, C.; ALVAREZ, C.; FRANCO, F. J. G.; GABALDON, A. *Methods for customer and demand response policies selection in new electricity markets.* ***IET Generation, Transmission Distribution***, v. 1 n. 1, p. 104-110, 2007.

VINE, E. *An International Survey of the Energy Service Company (ESCO) Industry.****Energy Policy***, v. 33, p.691-704, 2005.

VLACHOS, M.;LIN, J.;KEOGH, E.;GUNOPULOS, D. A *wavelet-based anytime algorithm for k-means clustering of time series.* ***Proceedings of the Third SIAM International Conference on Data Mining.*** *San Francisco*, CA, May 1–3, 2003.

XIE, X.; BENI, G. *A validity measure for fuzzy clustering.* ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, v. 13 n. 8, p. 841-847, 1991.

XIONG, Y.; YEUNG, D.-Y. *Mixtures of ARMA models for model-based time series clustering.* ***Proceedings of the IEEE International Conference on Data Mining****, Maebaghi City, Japan*, p. 9–12, 2002.

XIONG, T.; BAO, Y. *Interval forecasting of electricity demand: A novel bivariate EMD-based support vector regression modeling framework.* ***International Journal of Electrical Power & Energy Systems***, 2014; 63:353-362.

WANG, L.; MEHRABI, M.G.; JR, E. Kannatey-Asibu. *Hidden Markov model-based wear monitoring in turning.* ***J. Manufacturing Sci. Eng.*** 124, p. 651– 658, 2002.

WIJK, J. J. VAN; SELOW, E. R. VAN. *Cluster and calendar based visualization of time series data.* ***Proceedings of IEEE Symposium on Information Visualization****, San Francisco*, CA, p. 25-26, 1999.

WILPON, J.G.; RABINER, L.R. *Modified k-means clustering algorithm for use in isolated word recognition.* ***IEEE Trans. Acoust. Speech Signal Process***. 33 (3) (1985) 587 –594.

WISMÜLLER, A.; LANGE, O.; DERSCH, D. R.; LEINSINGER, G. L.; HAHN, K.; PÜTZ, B.; AUER, D. *Cluster analysis of biomedical image time series,* ***Int J. Comput. Vision*** 46 (2), p. 103-128.

WITTEN, I.H. ; FRANK, Eibe. ***Data mining: practice machine learning tools and techniques****,* 2nd edition. Elsevier: San Francisco. 2005. p. 2-5.

WU, E. H. C.; LI, P. L. H. *Independent component analysis for clustering multivariate time series data, in: X. Li, S.Wang, Z. Y. Dong (Eds.), ADMA,* ***Lecture Notes in Artificial Intelligence***, v. 3384, Springer-Verlag, Berlin, Heidelberg, pp. 474–482,2005.

YAGER, R.; FILEV, D. *"Approximate clustering via the mountain method". Iona College Tech. Report #MII-1305, 1992. Also to apper* ***in IEEE trans. On Systems, Man & Cybernetics***, 1992.

YANG, K.; SHAHABI, C. *A PCA-based Similarity Measure for Multivariate Time Series. In* ***MMDB '04 Proceedings of the 2nd ACM international workshop on Multimedia databases,*** p. 65-74, 2004.

YU, I. H.; LEE, J. K.; KO, J. M.; KIM, S. I. *A method for classification of electricity demands using load profile data. Proc. Fourth Annual ACIS Intern, in* ***Conference Computer and Information Science***, p.164-168, 2005.

YU, DAREN; YU, XIAO; HU, QINGHUA; LIU, JINFU; WU, ANQI. *Dynamic time warping constraint learning for large margin nearest neighbor classification.* ***Information Sciences***, v. 181, p. 2787-2796, 2011.

ZAKARIA, Z.; LO, K L; SOHOD, H. M. *Application of Fuzzy Clustering to Determine Electricity Consumers Load Profiles. In* ***First International Power and Energy Conference****, Putrajaya, Malaysia*, p. 99-103, 2006.

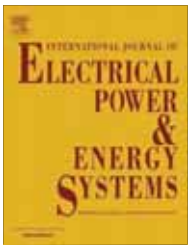
ZALEWSKI, W. *Aplication of Fuzzy Inference to Electric Load Clustering.* ***IEEE in International Conference on Power Systems***, p. 1-5, 2006.

ZHANG, J.; DOBSON, I.; ALVARADO, F. L. *Quantifying transmission reliability margin.* ***International Journal of Electrical Power and Energy Systems***, 2004; 26(9): p 697-702.

ZIMBA, P. V.; MISCHKE, C. C.; B, S. S. *Pond age–water column trophic relationships in channel catfish ictaluruspunctatus production ponds.* ***Aquaculture***, v. 219, p. 291 – 301, 2003.

ANEXO A – PRIMEIRO ARTIGO (PUBLICADO)

FERREIRA, ADONIAS M. S.; CAVALCANTE, CARLOS A. M. T.; FONTES, CRISTIANO H. O. ; MARAMBIO, JORGE E. S. *A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector.* ***International Journal of Electrical Power & Energy Systems*** JCR, v. 53, p. 824-831, 2013a.



ABSTRACT

This works presents a method for the selection, typification and clustering of load curves (STCL) capable of recognizing consumption patterns in the electricity sector. The algorithm comprises four steps that extract essential features from the load curve of residential users with an emphasis on their seasonal and temporal profile, among others. The method was successfully implemented and tested in the context of an energy efficiency program carried out by the Energy Company of Maranhão (Brazil). This program involved the replacement of refrigerators in low-income consumers' homes in several towns located within the state of Maranhão (Brazil). The results were compared with a well known time series clustering method already established in the literature, Fuzzy C-Means (FCM). The results reveal the viability of the STCL method in recognizing patterns and in generating conclusions coherent with the reality of the electricity sector. The proposed method is also useful to support decision-making at management level.

Keywords: Load profiles; Clustering; Pattern recognition; Electric sector.

1 INTRODUCTION

Multivariate analysis is a powerful tool for knowledge extraction especially when applying pattern recognition techniques based on data. In this context, the analysis of energy consumption data measured in homes can identify opportunities for improvement in the load factor [1] and energy efficiency of the distribution system through specific actions by the customer [2]. Methods of Data Mining (DM) that can extract useful information from data can be used to develop decision-making tools so as to improve production systems and management technology [3-6].

Some works present the use of clustering and load curve typification (pattern recognition) methods in the electric power sector. Gerbec et al. [7] performed a load curve typification using a hierarchical clustering method highlighting the advantage of this method in choosing the appropriate number of groups. The non-hierarchical method [8] emphasizes the

minimization of internal variance within a cluster and also the reduction of similarity between different groups. Gemignani et al [9] combined the hierarchical and non-hierarchical clustering methods to improve clustering efficiency in the recognition of different consumption patterns at the same level of tension. Zalewski [10] used fuzzy logic for clustering and load curve typification. The author performed the clustering of load profiles in order to classify substations into homogeneous groups according to consumption peak. Nizar et al [11] combined two methods, namely, Feature Selection and Knowledge Discovery in Databases (KDD) [12], to obtain better patterns of load demand in a distribution system. A recent study about knowledge extraction from electric power consumer data [10] presents an overall analysis and prediction of energy consumption trends (Incremental Summarization and Pattern Characterization - ISPC).

Some studies compare the performance of various methods of typification and conclude that the fuzzy C-Means (FCM) provides the best level of cohesion and discrimination of the problems associated with clustering in load curves. From this very point of view some authors have recently highlighted the FCM method in applications involving pattern recognition (typification) in load curves [14-16].

This study proposes a new method of selection, typification, and load curve clustering (STCL) based on a systematic extraction of features. This method is capable of identifying a greater diversity in demand patterns and also represents a potential tool for the improvement of the decision-making process through better classification of heterogeneous consumer profiles in the electric power sector. The case study analyzed is an energy efficiency program [17] carried out by the Electric Company of Maranhão (Brazil), that considers, among others, the analysis of the impact of replacing refrigerators in low-income consumers' homes distributed in several towns located within the state of Maranhão (Brazil). The proposed method incorporates multiple criteria in the clustering and typification of load curves unlike traditional approaches that essentially use the criterion of distance between load curves for cluster recognition. Section 2 presents the STCL method and the evaluation metrics adopted. Section 3 presents the case study and results obtained from the application of FCM and STCL methods demonstrating the ability and superiority of the latter in describing the problem.

2 THE STCL METHOD

The STCL method (Fig. 1) comprises two phases. The first carries out pattern recognition through successive iterations. The first iteration performs the clustering of the whole sample

(load curves from the database) based on specific features associated with the consumption profile, and some clusters of load curves are obtained. The subsequent iterations consider only the medians (patterns) of each group generated and verify the similarity between these medians based on the same statistical tests considered in the first iteration such that some patterns may be collapsed. Thus, at the end of the first phase (after convergence), patterns or types associated with load curves are recognized. The second phase defines the final groups associating each load curve (database) to one of the patterns recognized in the first phase.

Initially, each curve is normalized within the interval [0, 1] dividing the hourly measurements by the peak demand of each. The dimensionless consumption quantified in this way is called power per unit (pu) [16].

Table 1 presents the criteria considered in the similarity analysis performed in each step of the first phase together with the statistical test applied. These criteria were established according to the requirements and indicators practiced in the electric energy distribution sector [10, 18-21].

The three features (three stages) presented in Table 1 are applied successively. In the first iteration, the clusters are formed based on similarity between the load curves and the curve with the highest average power consumption (reference curve). After the first iteration, the method assumes that the median of each group keeps its own features and the same tests are successively applied considering the medians (patterns). The existence of a similarity between medians according to the statistical tests implies the union of groups and new medians are obtained.

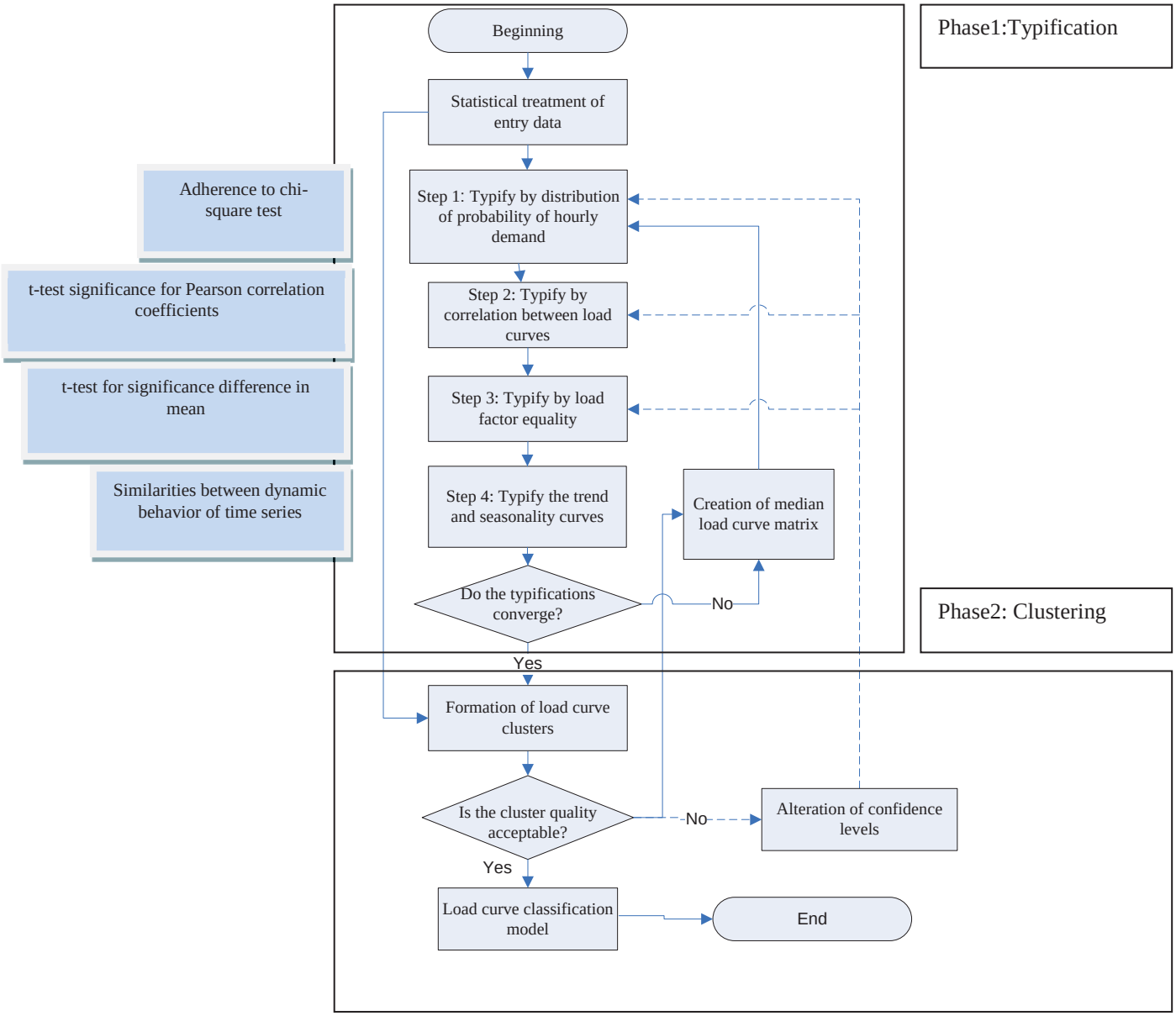


Figure 1 –The STCL method

Clustering criterion	Statistical test
Distribution of probability of hourly demand	Chi-square for goodness of fit [16]
Selection by correlation level between load curve	Independent two-sample t-test significance for Pearson correlation coefficients [17]
Selection by mean consumption or Load Factor	Independent two-sample t-test for significance difference in mean [17]

Table 1 – Statistical tests used in sequence in the process of typing.

The LF presented in Table 1 is the ratio between the average and maximum demands of a load curve. The LF is an evaluation index for the rational use of electric power by the consumer [20]. From the viewpoint of the energy distribution system, considering the consumption expectation, the lower the LF, the more rational and economical the energy use will be. From the viewpoint of the equipment, the lower the LF, the more efficient its operation will be.

In the first three stages, the testing of biserial statistical hypotheses between load curves is carried out according to Table 1. There is a fourth stage that comprises a multivariate quantification of dissimilarity between load curves in relation to their dynamic behavior throughout the day. The load curves of each set generated in the third stage are submitted to an additional clustering according to seasonality. This clustering is carried out in two sub stages. In the first, the existence of seasonality is detected through the application of factorial analysis combined with Principal Component Analysis (PCA) of the load curves of each cluster recognized [22]. The analysis of seasonality is performed through the relationship between the hourly consumption of all the curves during the period of 24 hours. The data is represented by a nc×24 matrix where nc is the number of curves presented in each cluster recognized at the end of stage 3. The factorial analysis method applied to this matrix provides the identification of a reduced number (less than 24 and suggested by PCA) of factors that characterize the seasonality of each curve [23]. The second sub-stage comprises the application of a clustering method (subtractive data algorithm [24]) on these factors.

The first phase is repeated successively in order to verify any possible similarity between some of the patterns (median of each group), indicating the need for re-clustering. This first phase is concluded when there is a convergence in the number of patterns. Thus, the number of patterns is a result of the method itself avoiding the need for an initial estimation.

In the second phase of the STCL method (Fig. 1) each load curve of the whole sample is associated to one of the typical curves recognized in the first phase according to the shortest Euclidian distance. The final clusters obtained undergo evaluation. One of the metrics adopted to measure the clustering quality is the Silhouette Index [25]. This index measures the cohesion within and differences between the clusters regardless of the clustering method applied.

Considering N_K load curves (objects) belonging to the K cluster and a total of G clusters ($G \geq 2$) , the Silhouette Index for each load curve is

$$S_i^L = \frac{b_i^L - a_i^L}{\max\left\{a_i^L, b_i^L\right\}} \qquad i = 1,\ldots,\sum_{k=1}^{G} N_k \tag{1}$$

where $S_i^L \quad \left(-1 \leq S_i^L \leq 1\right)$ is the Silhouette Index of i curve belonging to the L cluster ($1 \leq L \leq G$). a_i^L (equation1) is the average distance between the i curve and all other load curves belonging to the same cluster. b_i^L (equation1) is the minimum average distance between the i curve and the load curves belonging to the other clusters.

$$a_i^L = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{N_L - 1} d_{ij}^{L,L}}{N_L - 1} \qquad i = 1,\ldots,N_L \tag{2}$$

$$b_i^L = \min_{\substack{Z=1 \\ Z \neq L}}^{W} \left(\frac{\sum_{j=1}^{N_L} d_{ij}^{L,Z}}{W} \right) \quad where \quad W = \sum_{\substack{K=1 \\ K \neq L}}^{G} N_K \tag{3}$$

$d_{ij}^{L,K}$ is the Euclidian distance between the i curve, belonging to the L cluster, and the j curve belonging to the K cluster.

A Silhouette Index close to unity and positive is desired. Negative values imply that there is more homogeneity between the clusters and less internal cohesion. In this work, the mean between the Silhouette Indices [25] obtained for all curves (General Silhouette Index – GSI) was adopted to evaluate the clustering quality.

3 CASE STUDY & RESULTS

The SCTL method was applied in order to analyze possible changes in the consumption profiles of an energy efficiency program carried out by the Energy Company of Maranhão (CEMAR-Brazil) during the period of November 2008 to July 2009. This program essentially involved the replacement of 5,250 old refrigerators for new ones in low-income communities. The sampling process comprised two steps, namely, sampling by clustering and systematic sampling. In the former, the municipalities belonging to the State of Maranhão were classified in clusters according to the similarity between the consumption profiles (load curves). One municipality (center or prototype) was selected to represent each cluster. In the systematic sampling, sampling units (refrigerators) were selected based on the average monthly consumption available in the CEMAR'S records. A sample of eighty load curves (old refrigerators), presenting a high consumption of electric energy (case I, average consumption of 82 kWh) and another sample of 80 load curves after the replacement of the refrigerators (case II, average consumption of 52 kWh) were obtained. In order to reduce the effects of seasonality, all the data are related to working days and the period considered comprises almost the entire summer (the season with the highest energy consumption). The sample size represents an error level of 10% variation in sample means and a confidence level of 95% in the prediction of the population parameter. The International Performance Measurement & Verification Protocol (IPMV) recommends a sampling error of up to ± 10% [22].

Initially the data was analyzed to identify and exclude outlier curves. This analysis comprised the application of factorial analysis together with PCA [22, 27] (also used in the first phase of STCL method). PCA was used initially to suggest the number of factors that explain 80% of the total variance of the sample (above this level the addition of more factors represented little contribution to explaining the total variance). The PCA provides a reduction in the dimensionality of each load curve (24 points) enabling the selection of factors (less than 24) that can represent the dynamic behavior of each curve. The factorial analysis was then applied to obtain the factorial scores (80 Scores for each factor). Figs. 2 and 3 show the distribution value of factorial scores of the factor for the cases I and II. According to the box-plots, the points identified by the sign "+" indicate outlier curves. Each point may occur on more than one factor (more than one box-plot) associated to the same curve. This analysis enabled the identification of 13 outlier curves in case I and 23 outlier curves in case II. The increase in the number of factors (number of box-plots) in case II is associated to the greater seasonal

variation caused by the lower energy demand of the motors of the new refrigerators. This behavior is further corroborated below.

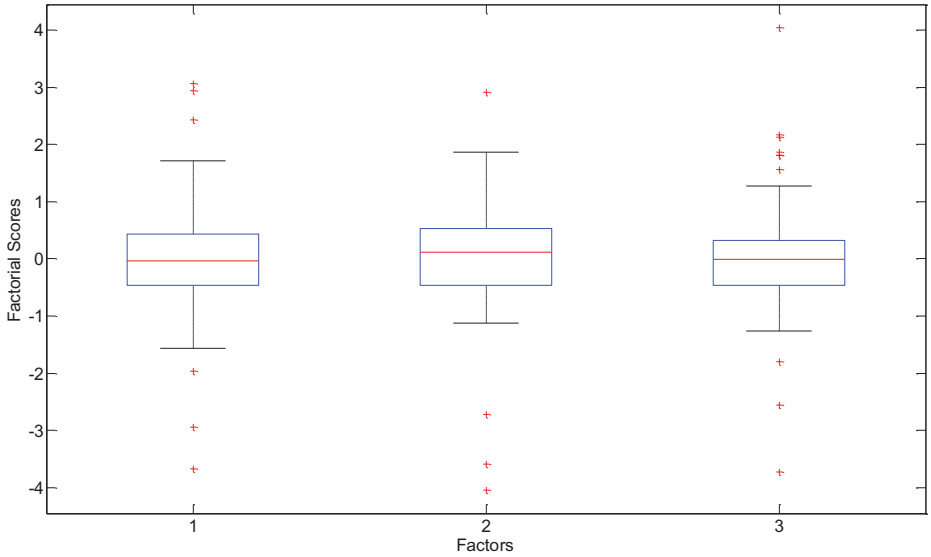


Figure 2 – Distribution of values in each factor (case I).

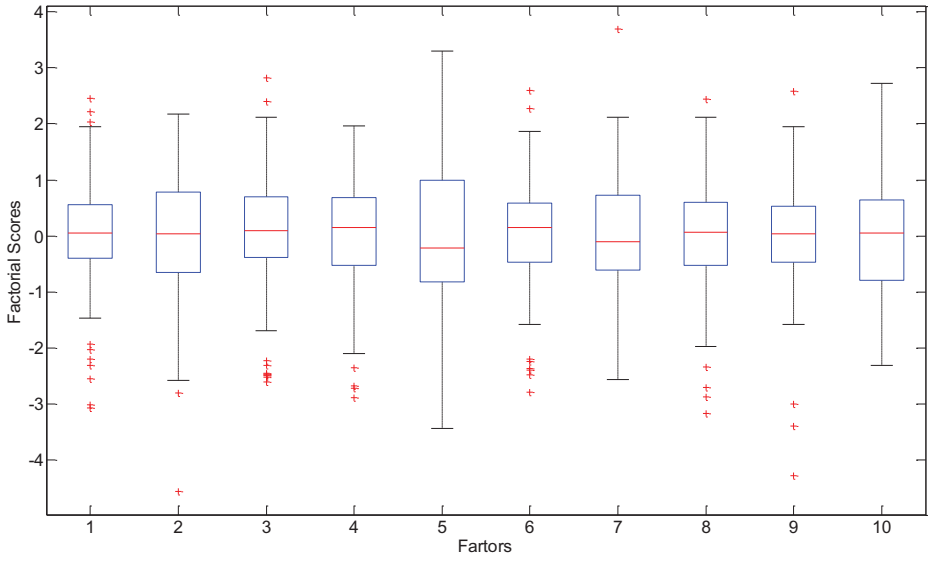


Figure 3 – Distribution of values in each factor (case II).

The results obtained using the STCL method were compared with the *Fuzzy C-Means* (FCM), a well-known method belonging to the C-Means families of batch clustering models [28, 29], suitable for clustering objects represented by time series [30]. The FCM method requires an initial guess for the number of clusters and, in this case, the same number of clusters provided

by the STCL method was considered. The confidence level adopted in the first three stages of the first phase of STCL was 99% in cases I and II.

The application of the STCL method in case I was capable of recognizing the existence of three patterns or demand profiles. On the other hand, two of the three patterns recognized by the FCM were similar attesting the recognition of only two patterns (Fig. 4). The STCL method was capable of recognizing a third pattern of consumption related to 13 curves. This result suggests the ability of STCL to handle a sample of objects with a higher level of heterogeneity (before the replacement of refrigerators). Furthermore, the third pattern represents a profile with lower energy consumption even considering the use of old refrigerators. The quality of clustering obtained with the FCM method was slightly lower according to the Silhouette Index (GSI equal to 0.25 and 0.28 for FCM and STCL methods respectively).

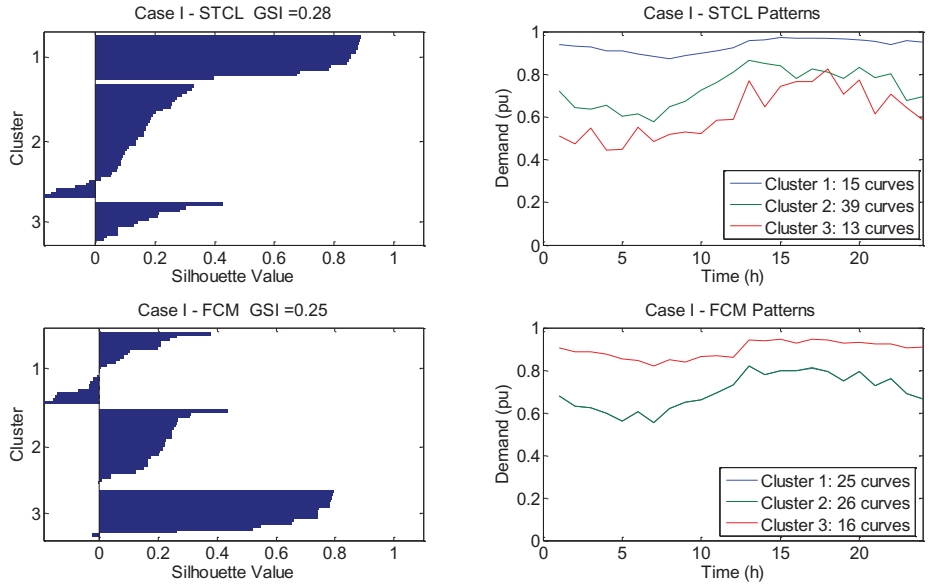


Figure 4 – Silhouette Indices and patterns recognized via the STCL and FCM methods without outlier curves in the sample (case I).

For the sample of load curves after refrigerator replacement (case II), both STCL and FCM recognized the existence of only one cluster and similar patterns (Fig. 5). This shows that the electric energy demand profiles became more similar after the refrigerator replacement, indicating an increase in uniformity among consumers.

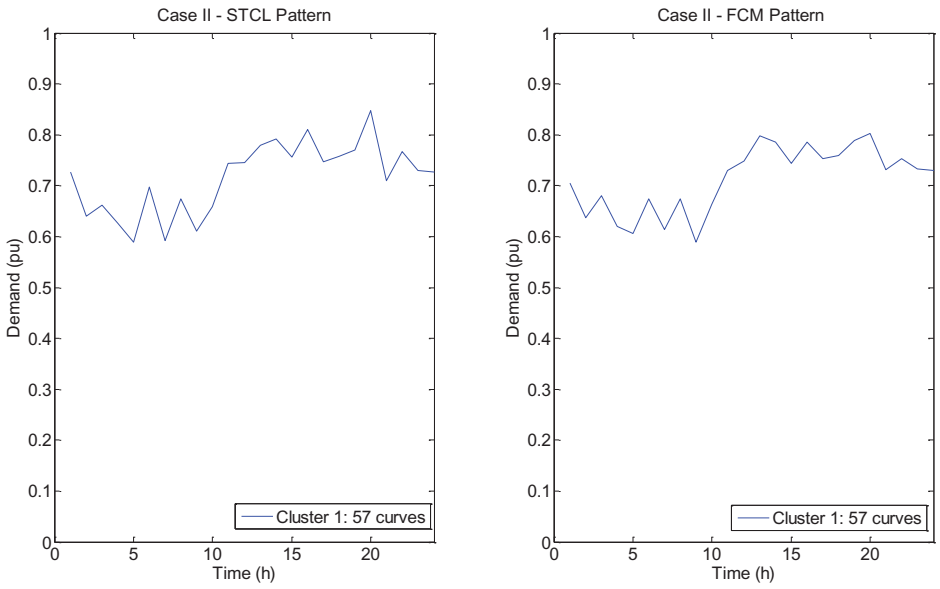


Figure 5 –Patterns recognized via the STCL and FCM methods (case II).

Each cluster recognized may be regarded as a sample of possible trajectories that can also be considered as a stochastic process (sample space of possible trajectories [30-32]). From this perspective, the ensemble of load curves of each cluster is a known sample of the underlying stochastic process. In this case, one can check the seasonal variations in each cluster through autocorrelation analysis [31-33]. The existence of two levels of trend in the patterns recognized (specially in relation to case II) suggests, in this case, the application of autocorrelation analysis on the first order differences of the original series in order to mitigate non-stationary effects [31-33]. Fig. 6 presents the correlogram associated to the modal cluster (highest number of load curves) of case I. Only the first value of the autocorrelation (at lag 1) is significant indicating that seasonal variations are due to random factors. In case II (Fig. 7) some autocorrelation values at lags higher than 2 are close to zero and, at the same time, some non-successive autocorrelations are significant, confirming the existence of two more pronounced levels of consumption and the increase in seasonal variations, also confirmed in the preliminary analysis of the curves (factorial analysis). In this case, the increase in the seasonal variations is predictable and consistent because the new refrigerator has better thermal insulation meaning that the motor consumes less power.

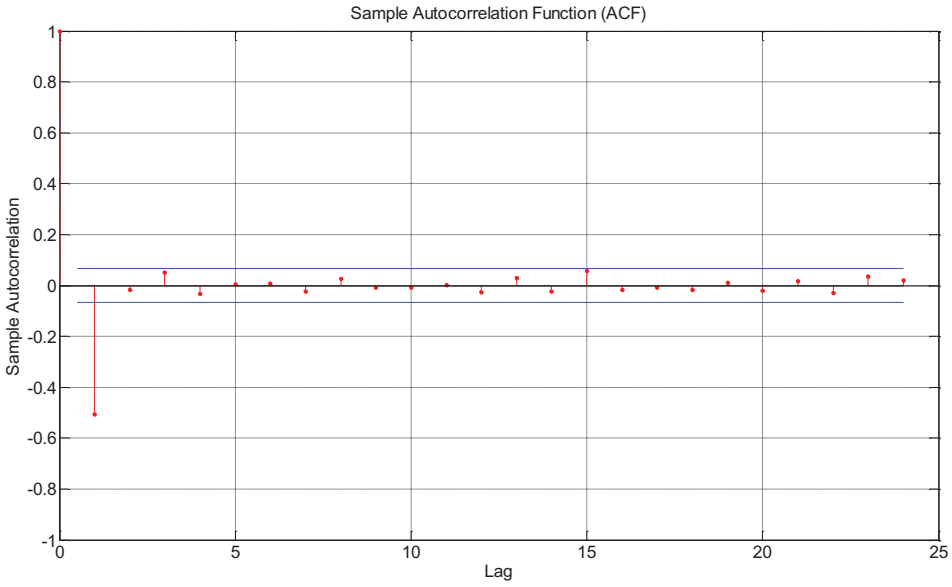


Figure 6 - Autocorrelation values of the first order difference (modal cluster - case I).

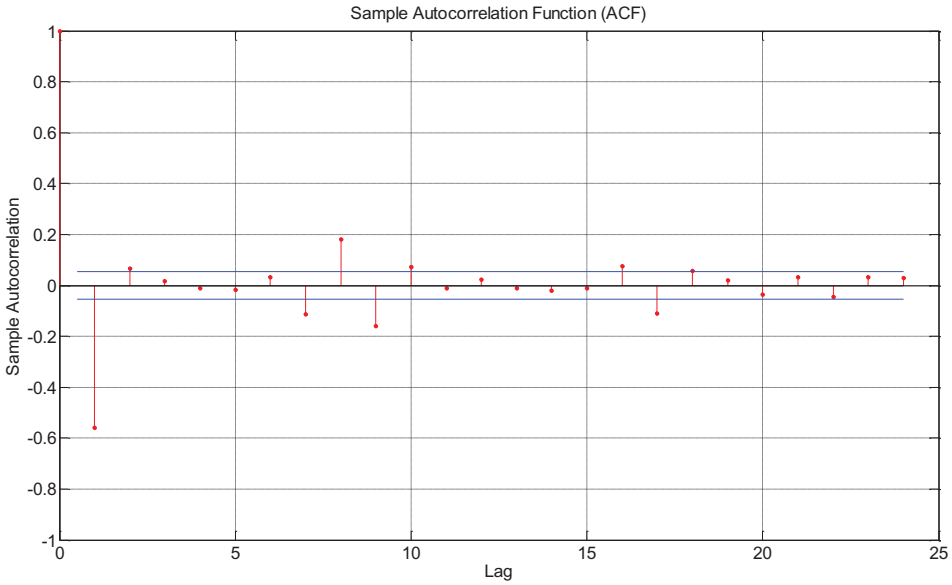


Figure 7 – Autocorrelation values of the first order difference (case II).

An additional analysis comprised the distribution of load factor in the clusters recognized. In case I before optimization (Fig. 8), the STCL method provided clusters whose median load factors is close to the load factor of the respective pattern recognized (0.65, 0.75 and 0.95 pu for the clusters 3, 2, and 1, respectively – Fig. 4). This does not occur with FCM method revealing an inconsistency in the recognition of the patterns (typical curves) in this case. According to STCL method (and also FCM), the distribution of load factors in case II shows lower values (Fig. 9). The reasons are related to the same factors that increase seasonality, i.e.

new refrigerators make use of more advanced technology such as better insulation and a motor with a lower energy demand.

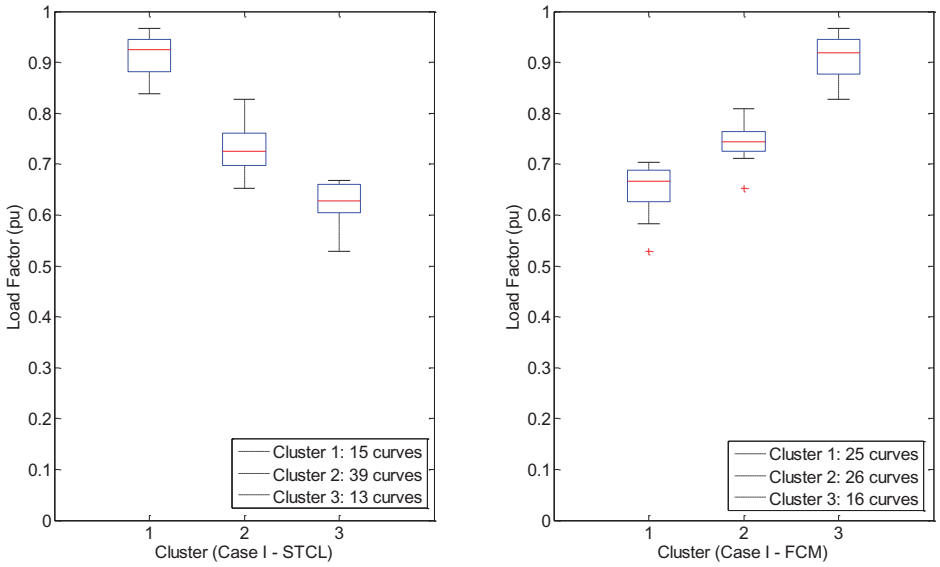


Figure 8 – Distribution of Load Factor in the clusters recognized (case I).

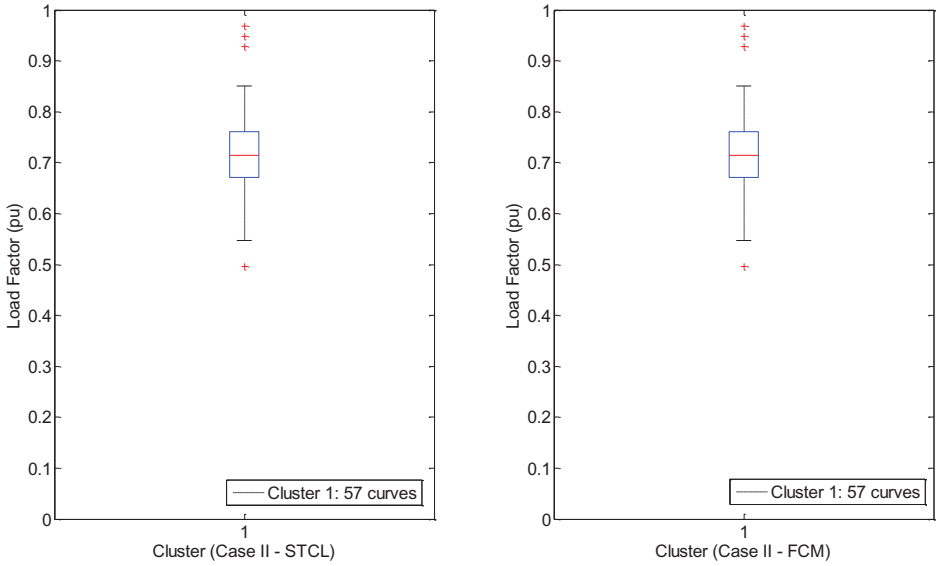


Figure 9 – Distribution of Load Factor in the clusters recognized (case II).

4 CONCLUSION

This study presented a new method of selection, typification, and clustering of load curves (STCL) in which the extraction of features is based on indicators and parameters intrinsic to the electricity sector. The STCL method is suitable for unlabeled data and its adherence to the electric sector makes this a potential tool for applications in this area. The results demonstrate its good performance in recognizing patterns in samples with heterogeneous data (common situation in the electric sector). Unlike C-means models of clustering, the number of clusters is also a result obtained by STCL method.

The case studied looked at an energy efficiency program carried out by the Energy Company of Maranhão (CEMAR-Brazil) which analyzed the impact of replacing 5,250 old refrigerators with new ones for low-income consumers. The results obtained by STCL, compared to a well-known method of clustering (*Fuzzy C-Means*, FCM), reveal the viability and potential of the former in recognizing patterns and in generating conclusions coherent with the reality of the electric power sector. This supports the implementation of efficiency actions based on real features within the consumer market and can also support decision-making at management level.

REFERENCES

[1] Nabeel I., Tawalbeh A. Daily load profile and monthly power peaks evaluation of the urban substation of the capital of Jordan Amman. Int J Electr Power Energy Syst vol. 37, 1, p. 95–102.

[2] Tsekouras G. J,, Tsaroucha M. A., Tsirekis C. D., Salis A. D., Dialynas E. N., Hatziargyriou N. D. A database system for power systems customers and energy efficiency programs. Int J Electr Power Energy Syst 2011;33(6):1220–8.

[3] Jota Patricia R. S., Silva Valéria R. B., Jota Fábio G. Building load management using cluster and statistical analyses. Int J Electr Power Energy Syst 2011;33(8):1498–505.

[4] Monedero Iñigo, Biscarri Félix, León Carlos, Guerrero Juan I., Biscarri Jesús, Millán Rocío. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. Int J Electr Power Energy Syst 2012;34(1):90–9.

[5] Lin J. K., Tso S. K., Ho H. K., Mak C. M., Yung K. M., Ho Y. K. Study of climatic effects on peak load and regional similarity of load profiles following disturbances based on data mining. Int J Electr Power Energy Syst 2006;28(3):177–85.

[6] Piatetsky G. Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from ‘‘universty’’ to ‘‘business’’ and ‘‘analystics’’. Data Min Knowled Discov 2007;15:99–105.

[7] Gerbec D., Gasperic S., Smon I., Gubina F. A methodology to classify distribution load profiles. Presente dat the IEEE 2002:848–51.

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

[9] Geminagnani M. M. F., Oliveira C. C. B., Tahan C. M. V. Proposition and comparative analysis of alternative selection and classification of load curve for defining types for tariff studies. Décimo Tercer Encuentro Reginal Iberoamericano de Cigré – XIII ERIAC; 2009. p. 1–6.

[10] Zalewski W. Aplication of fuzzy inference to electric load clustering. New Delhi: IEEE International Conference on Power Systems; 2006. p. 1–5.

[11] Nizar A. H., Dong Z. Y., Zhao J. H. Load profiling and data mining techniques in electricity deregulated market. In: Presented at the IEEE power engineering society (PES) general meeting 2006, Montreal, Quebec, Canada; June 2006. p. 1–7.

[12] Han J., Pei J., Yiwen Y. Mining frequent patterns without candidate generation. In: Proceedings ACM-SIGMOD international conference on management of data. ACM Press; 2000. p. 1–12.

[13] Silva Daswin, Yu Xinghuo. A data mining framework for electricity consumption analysis from meter data. IEEE Trans Indust Inf 2011;7(3):399–407.

[14] Gerbec D., Gasperic S., Smon I., Gubina F. Determining the load profiles of consumers based on fuzzy logic and probability neural networks. IEEE Proc Gener Transm Distrib 2004;151(3):395–400.

[15] Zuhaina Zakaria, Lo K. L., Hadi Mohamad Sohod. Application of fuzzy clustering to determine electricity consumers' load profiles first international power and energy conference. Putrajaya, Malaysia; 2006. p. 99–103.

[16] Anuar N., Zakaria Z. Cluster validity analysis for electricity load profiling. In: IEEE international conference on power and energy. Kuala Lumpur Malaysia; 2010. p. 35–8.

[17] Cursino dos Santos Arthur Henrique, Werneck Fagá Murilo Tadeu, Moutinho dos Santos Edmilson. The risks of an energy efficiency policy for buildings based solely on the consumption evaluation of final energy. Int J Electr Power Energy Syst 2013;44(1):70–7.

[18] Anssi Seppãlã. Statistical distribution of customer load profile. IEEE, CATALOGUE No. 95TH8130;1995. p. 696–701.

[19] Joanicjusz Nazarko, Styczynski Zbigniew A. Application of statistical and neural approaches to the daily load profiles modeling in power distribution systems. IEEE 1999:320–5.

[20] Joseph Janes. Categorical relationships: chi-square. Library Hi Technol 2001;19(3):296–8.

[21] O'Gorman T. W. A comparison of an adaptive two-sample test to the t-test, rank- sum, and log-rank tests. Commun Statis – Simul Comput 1997;26:1393–411.

[22] Motomasa DAIGO. Factor analysis and pattern decomposition method. SPIE 2005;6043(604317):1–8. 53–65.

[23] Yu Daren, Yu Xiao, Hu Qinghua, Liu Jinfu, Anqi Wu. Dynamic time warping constraint learning for large margin nearest neighbor classification. Inf Sci 2011;181:2787–96.[24] Chiu S. A cluster estimation method with extension to fuzzy model identification. In: Proceedings of the third IEEE conference on fuzzy systems, vol. 2, Orlando – Florida, USA; 1994. p. 1240–5.

[25] Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.

[26] International performance measurement & verification protocol – IPMV; 2007.

[27] Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal 2008;52:5186–201.

[28] Bezdek James C., Keller James, Krisnapuram Raghu, Pal Nikhil R. Fuzzy models and algorithms for pattern recognition and image processing. Springer Science+Businees Media, Inc.; 2005.

[29] Bensaid A., Hall L.O., Bezdek James C., Clarke L. P. Partially supervised clustering for image segmentation. Patt Recogn 1996;29(5):859–87.

[30] Warren Liao T. Clustering of time series data-a survey. Patt Recogn 2005;38(11):1857–74.

[31] Abdel-Aal R. E. Modeling and forecasting electric daily peak loads using abductive networks. Int J Electr Power Energy Syst 2006;28(2):133–41.

[32] Saini L. M., Soni M. K. Artificial neural network-based peak load forecast tingusing conjugate gradient methods. IEEE Trans Power Syst 2002;17:907–12.

[33] Aggarwal Sanjeev Kumar, Saini Lalit Mohan, Ashwani Kumar. Electricity price forecasting in deregulated markets: a review and evaluation. Electr Power Energy System 2009;31:13–22.

ANEXO B – SEGUNDO ARTIGO (PUBLICADO)

FERREIRA, ADONIAS M. S.; CAVALCANTE, CARLOS A. M. T.; FONTES, CRISTIANO H. O. ; MARAMBIO, JORGE E. S. Pattern recognition as a tool to support decision making in the management of the electric sector. Part II: a new method based on clustering of multivariate time series. **International Journal of Electrical Power & Energy Systems** JCR, v. 67, p. 613-626, 2015.

ABSTRACT

This work presents a new method for the clustering and pattern recognition of multivariate time series (CPT-M) based on multivariate statistics. The algorithm comprises four steps that extract essential features of multivariate time series of residential users with emphasis on seasonal and temporal profile, among others. The method was successfully implemented and tested in the context of an energy efficiency program carried out by the Electric Company of Alagoas (Brazil) that considers, among others, the analysis of the impact of replacing refrigerators in low-income consumers' homes in several towns located within the state of Alagoas (Brazil). The results were compared with a well-known method of time series clustering already established in the literature, the Fuzzy C-Means (FCM). Unlike C-means models of clustering, the CPT-M method is also capable to obtain directly the number of clusters. The analysis confirmed that the CPT-M method was capable to identify a greater diversity of patterns, showing the potential of this method in better recognition of consumption patterns considering simultaneously the effect of other variables in additional to load curves. This represents an important aspect to the process of decision making in the energy distribution sector.

KEYWORDS: Clustering, Pattern, Electricity Distribution, Multivariate Time Series.

1 INTRODUÇÃO

In the current model of regulation of the Brazilian electric sector, programs are often launched by the government to improve the performance of the electric energy distribution sector. These are related to energy efficiency in particular, which in turn is encouraged through tax

benefits. One such program comprises the exchange of old refrigerators for new ones in low-income communities. The most common way to evaluate the achievement of goals of energy efficiency is to compare the behavior of the power consumption (load curve) including possible displacement of the peak hours before and after the replacement of refrigerators. In order to improve the quality of the final evaluation of these programs, the load curve must be analyzed together with other time series, increasing the level of knowledge about demand behavior in the electric system [1].

In the management of demand in the electrical system, some variables influence the consumption patterns of electricity [2, 3]. Silk and Joutz [4] present a list of factors (price of electricity, price of home appliances, charging, dependence on energy, geographic location, ambient temperature, among others) that affect the electricity demand. On the other hand, the characterization of power consumption in the Brazilian electric sector is based on the charges imposed by the government regulatory agency. In the tariff class of residential consumers, in particular low-income consumers, refrigerators have the highest impact on energy consumption. The ambient temperature in turn has a strong effect on the thermal efficiency of the refrigerator, especially with respect to the use mode (frequency of opening the refrigerator door) which contributes to internal temperature deviations from the set point [5-7]. The thermal efficiency in this case can be quantified through the inverse of the Coefficient of Performance (COP^{-1}) of Carnot, which expresses the ratio between the work for cooling and the heat absorbed from the cold source [8]. Other meteorological variables (such as air temperature, humidity and rainfall) affect the energy consumption in homes as a whole [9] but, for the case study (consumption in refrigerators), just the room temperature should be considered.

Despite many works on cluster analysis in univariate time series [10, 11]), the standard approaches [10] and the feasibility to treat this kind of problem using point-prototype clustering models [12], the pattern recognition in multivariate time series represents a more complex problem (non-point prototyping problem) with intrinsic features [13]. This kind of problem cannot be solved directly using classic models of clustering point-prototype such as Fuzzy C-Means (as can be seen in Section 1.1), requiring special methods in this situation [14-20]. Furthermore, additional challenges must be considered such as the extraction of features from each data/object (set of time series), the similarity metrics related to the domain adopted and the approaches (featured-based or model-based, [12])

Despite the characterization of energy consumption in the electric sector, there is a lack of works in the pattern recognition using multivariate time series. Chicco [21] present a review about clustering methods (adaptive vector quantization [22], entropy-based, Renyi [23], follow-the-leader [24-29], fuzzy logic [30], fuzzy and ARIMA [31], fuzzy k-means [32, 33, 22], hierarchical clustering [26, 29, 22, 33], iterative refinement clustering [35], k-means [32, 34, 26, 29, 22], min-max neuro-fuzzy [36], multivariate statistics (MANOVA) [37], probabilistic neural network [30, 38], self organizing map [32, 25, 34, 26, 37, 30, 39-41], support vector clustering [42] and weighted evidence accumulation clustering [33]) all applied to the univariate case (only load curves). One can divide these methods into two major groups, namely, hierarchical and non hierarchical methods [43-46] and methods based on artificial intelligence [47-49]. A common feature of these methods is that the number of clusters (number of patterns) need to be previously defined. Recent works argue that methods based on artificial intelligence, specifically on C-means models [12], provide better quality in the clustering of load curves [50-56]. All the methods and works cited cope with the electric power consumption separately from other variables resulting in a point-prototype clustering.

Pattern recognition associated to the electric power consumption carried out together with other variables comprises a non point-prototype problem where each object is a set of time series. This work presents a new method of selection, pattern and clustering of multivariate time series (CPT-M) based on a systematic extraction of features from the objects. The case study analyzed comprises an energy efficiency program carried out by the Electric Company of Alagoas (Brazil), that considers, among others, the analysis of the impact of replacing refrigerators in low-income consumers' homes. The Energy Efficiency Program (EEP) was established by the National Agency of Electrical Energy (NAEE) in order to mitigate electricity losses. The main actions of this Program comprise the replacement of low efficiency end-use equipment (refrigerators) with new ones the developing of strategies to raise awareness of local population with regard to the rational and safe use of electricity. The eligible users must meet the following requirements: have a single-phase residential connection, have no irregular power connections, live in the region covered by the Program, have no debts with the electricity utility and have an average consumption in the last three months of more than 59 kWh (specific requirement for the replacement of refrigerators). Furthermore, the refrigerators removed from the customers' homes are recycled and do not return to the consumer market. In this case, the method is able to recognize patterns of energy consumption conjugated to the dynamic profiles of outside and fridge temperatures also

providing a consistent way to evaluate the efficiency of the program. The proposed method incorporates multiple criteria in the clustering and pattern recognition in load curves unlike traditional approaches that essentially use the criterion of distance between load curves [34]. Section 'The CPT-M method' presents a new method for the clustering and pattern recognition of multivariate time series (CPT-M) and the evaluation metrics adopted [57]. Section 'Case study and results' presents the case study and results obtained from a new version of Fuzzy C-Means - FCM (with algorithm adapted to the non point-prototype problem [58]) and CPT-M methods showing the ability and superiority of the latter to the problem analyzed.

1.1 Multivariate time series and objects

In Data Mining Theory, Clustering is the task of grouping data (or objects) into clusters according to the principles of homogeneity (data or objects belong to the same cluster should be as similar as possible) and heterogeneity (data or objects belong to different clusters should be as different as possible) [59, 60]. The clustering problem comprises unsupervised learning as there are no pre-labeled objects (there is no previous information to distinguish the objects from each other) [17, 45, 60].

A sample with n objects can be represented by the set $X = \{x_1, x_2, \ldots, x_n\}$. If each object $x_i \quad (i = 1, \ldots, n)$ is a feature vector in the space $\Re^p$ (p is the dimensionality of data set) there is the problem of point-prototype clustering [12].

A time series is a series of observations (measurements) made sequentially through time and associated to a specific process variable [60]. Two kinds of objects must be previously considered and represent different problems of clustering and pattern recognition of time series, namely, the Univariate Time Series (UTS) and Multivariate Time Series (MTS) [15, 61, 62]. Considering a general series of observations over time $z_i(t) \quad (i = 1, \ldots, k; \; t = 1, \ldots, m)$ where k is the number of variables (number of sensors), m is the number of observations and i indexes the measurements made at each time instant, a UTS object comprises the case in which k=1. Otherwise ($k \geq 2$) there is a MTS object. Univariate time series has been broadly explored and is often regarded as a point-prototype clustering problem in multidimensional space ($\Re^m$). Traditional metrics of similarity (Euclidian distance) and traditional clustering methods applied to clustering static data (such as Fuzzy *C*-

Means and *K*-Means) can also be used in this case (raw-data-based approach [63]). On the other hand, MTS objects are common in various areas of knowledge. This approach is mandatory when there is need to consider more than one variable and all of them in an integrated way. An MTS should be treated as a whole and may not be transformed into one long univariate time series [61]. The clustering of MTS comprises a non-point prototyping problem and traditional metrics of similarity, appropriate for the univariate case, cannot be applied in this case. Additional challenges should be considered in the clustering of MTS objects such as feature extraction, similarity metrics and the selection of appropriate variables (reduction in the size of the problem) [14, 16; 13, 17; 15; 58]. Despite feature extraction and similarity metrics, and the fact that traditional metrics cannot be used, specific alternatives based on Principal Component Analysis (PCA) and Wavelets Transform have been proposed [15, 64]. The selection of appropriate variables can be performed either through the use of specific techniques (such as PCA) or also by an analysis of system and profiles of the time series.

A MTS object can be represented by the following $m \times k$ matrix:

$$Z_i = \begin{bmatrix} z_{i1}(1) & \cdots & z_{ik}(1) \\ \vdots & \ddots & \vdots \\ z_{i1}(m) & \cdots & z_{ik}(m) \end{bmatrix}$$

where Z_i is the object, z_{ij}(t) is the measurement of variable j (j=1,…,k) at time instant t (t=1,…,m) in the object Z_i (i=1,…,n objects). The column j contains the time series related to the variable j.

In this work, each object is a MTS with three variables (three time series), namely, the energy consumption (represented by load curves), external and fridge temperatures. The Figs. 1 and 2 present examples of objects (load curves together with each one of the temperatures) extracted from the samples associated to the old refrigerators (under subnormal conditions of use) and new (same technical specifications) refrigerators.

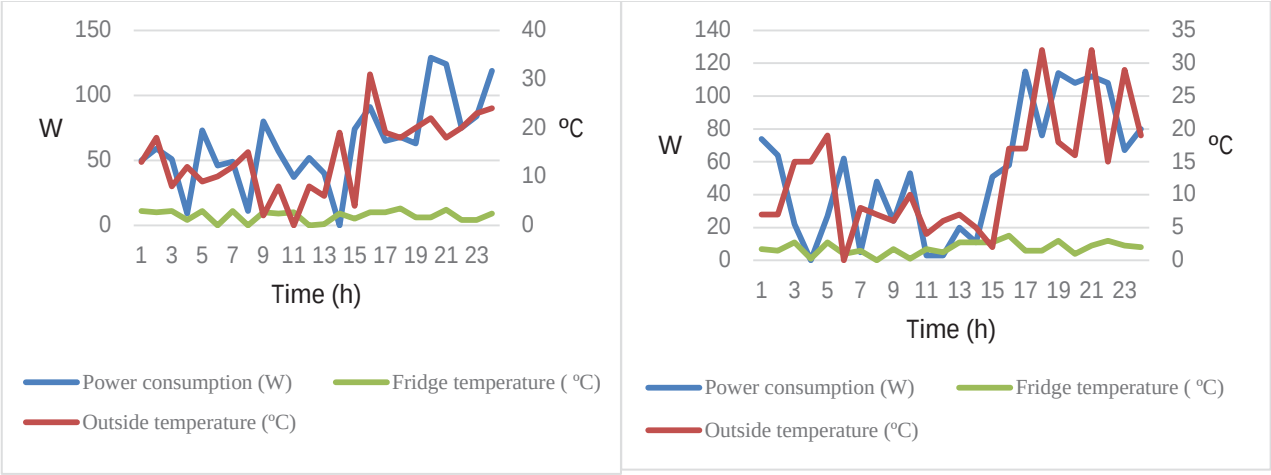


Figure 1 – Examples of objects (set of time series) – old refrigerators.

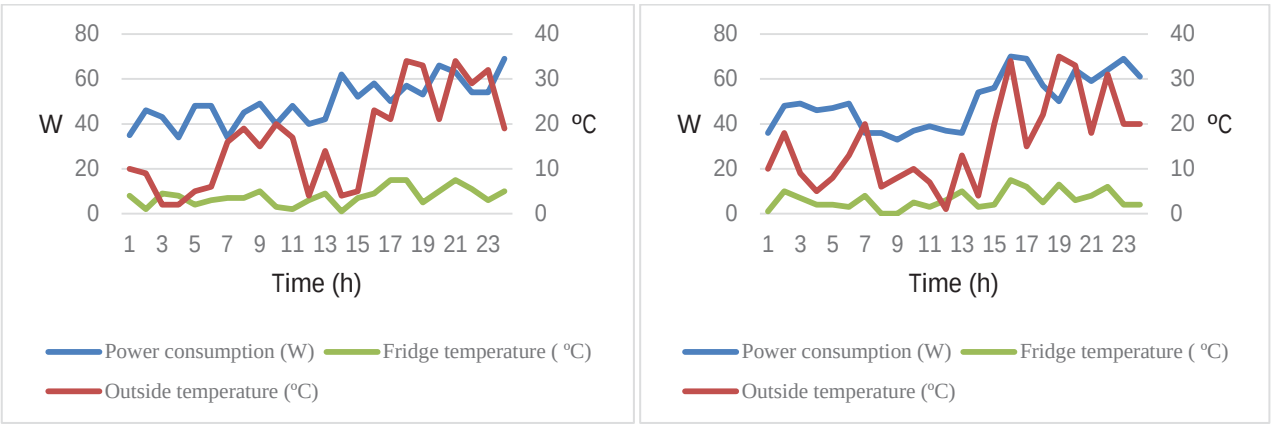


Figure 2 – Examples of objects (set of time series) – new refrigerators.

The clustering procedure also enables the recognition of one pattern (prototype or center) for each cluster. Each pattern also comprises a MTS with three time series and is not associated to a specific piece of equipment but it is associated to a cluster of refrigerators representing its thermal and consumption behavior.

2 THE CPT-M METHOD

The CPT-M method (Fig. 3 and Table 1) is an extension of its initial version (STCL [65, 66]) that applies only to univariate time series (load curve, point-prototype problem). The CPT-M method is capable of recognizing patterns in multivariate time series. In this work each object comprises three time series related to the three variables, namely, energy consumption (load), external and internal fridge temperatures.

The CPT-M method comprises two phases. The first carries out pattern recognition through successive iterations. In turn, each iteration comprises four stages and in each of these the clustering of the objects is performed through a similarity analysis that considers specific criteria. Only in the first iteration are all the objects presented in the original sample clustered according to the statistical test applied at each stage (Table 1). After the first iteration, the method assumes that the median of each group keeps its own features and the same tests are applied successively considering the medians. Thus, from the second iteration only the medians of each cluster obtained in the previous iteration are considered. The median (pattern) of each cluster is also a set of three time series (non-point prototype) comprising the respective medians of the variables considered (load, external and internal fridge temperatures). Two or more patterns (medians) may be joined according to the degree of similarity between them, identified through the statistical tests applied. The union of patterns (MTS) results in the recognition of a new pattern formed by the medians between the respective time series. At the end of the first phase (after convergence of successive iterations), the final patterns are recognized. The second phase defines the final clusters associating each object in the original sample to one of the patterns recognized in the first phase.

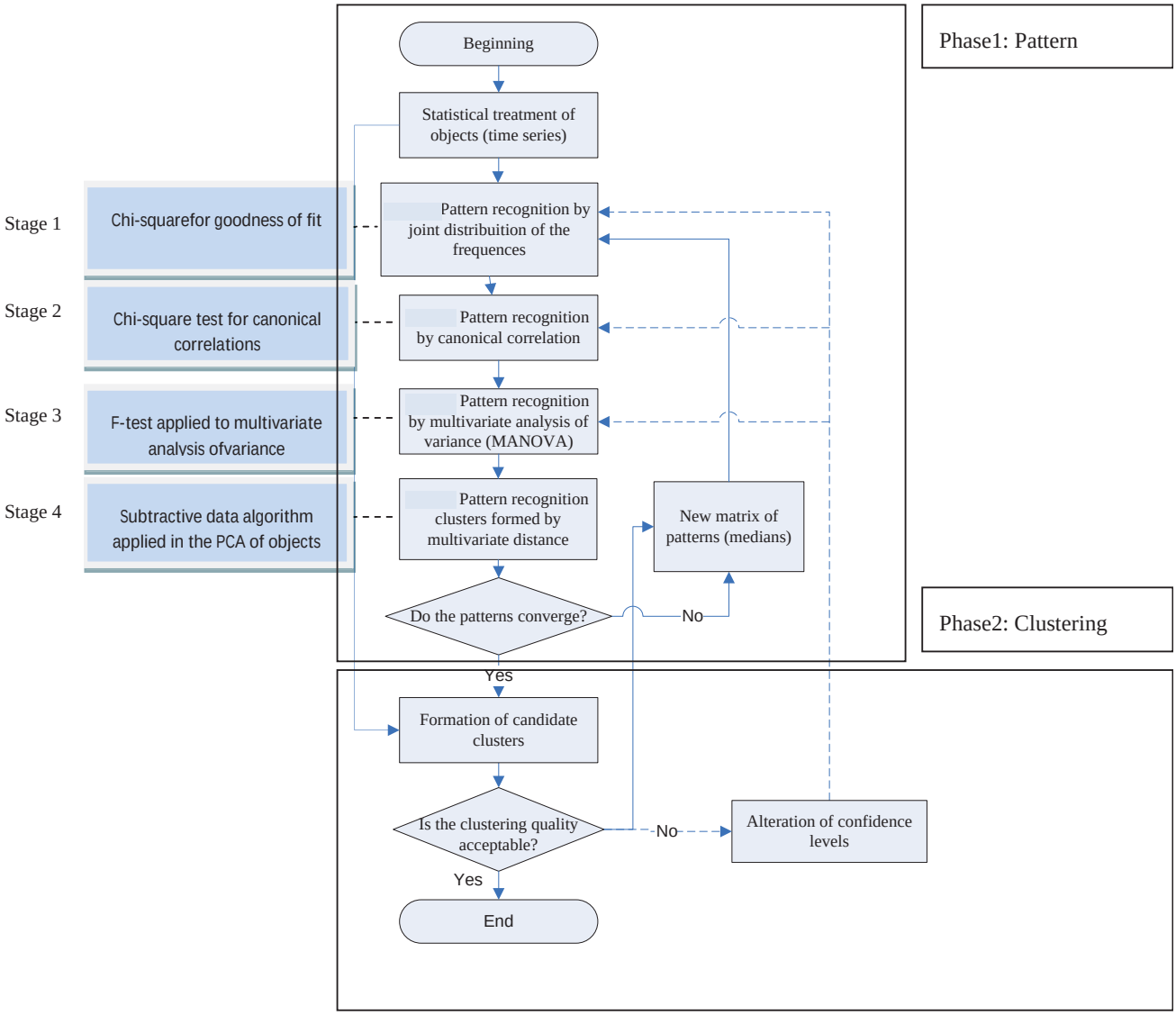


Figure 3–The CPT-M method.

STCL (univariate time serie)			CPT-M (multivariate time series)		
Clustering criterion	Statistical test	Assumptions	Clustering criterion	Statistical test	Assumptions
Distribution of probability of hourly demand (Stage 1)	Chi-square for goodness of fit [67]	• Sample random and independent • Frequency of class with at least five occurrences • Uniform distribution of the deviations between observed and expected values	Joint distribution of the frequency of the multivariate time series (objects)	Chi-square for goodness of fit [68]	The same as the STCL method
Selection by correlation level between load curve (Stage 2).	Independent two-sample Student's *t* test significance for Pearson correlation coefficients [69]	• Normal distribution • Equality between the correlation coefficients • Homogeneity of variance (homoscedasticity)	Canonical correlation between two objects.	Chi-square test for canonical correlations [70]	• Linearity • Multivariate normality • Homoscedasticity • Absence of multicolinearity
Selection by mean consumption (Stage 3)	Independent two-sample Student's *t* test for significance difference in mean [69]	• Normal distribution • Equality between means • Homogeneity of variance	Multivariate analysis of variance (MANOVA) between pairs of objects.	F-test applied to multivariate analysis of variance [71]	• Multivariate normal distribution • Equality between vectors of means

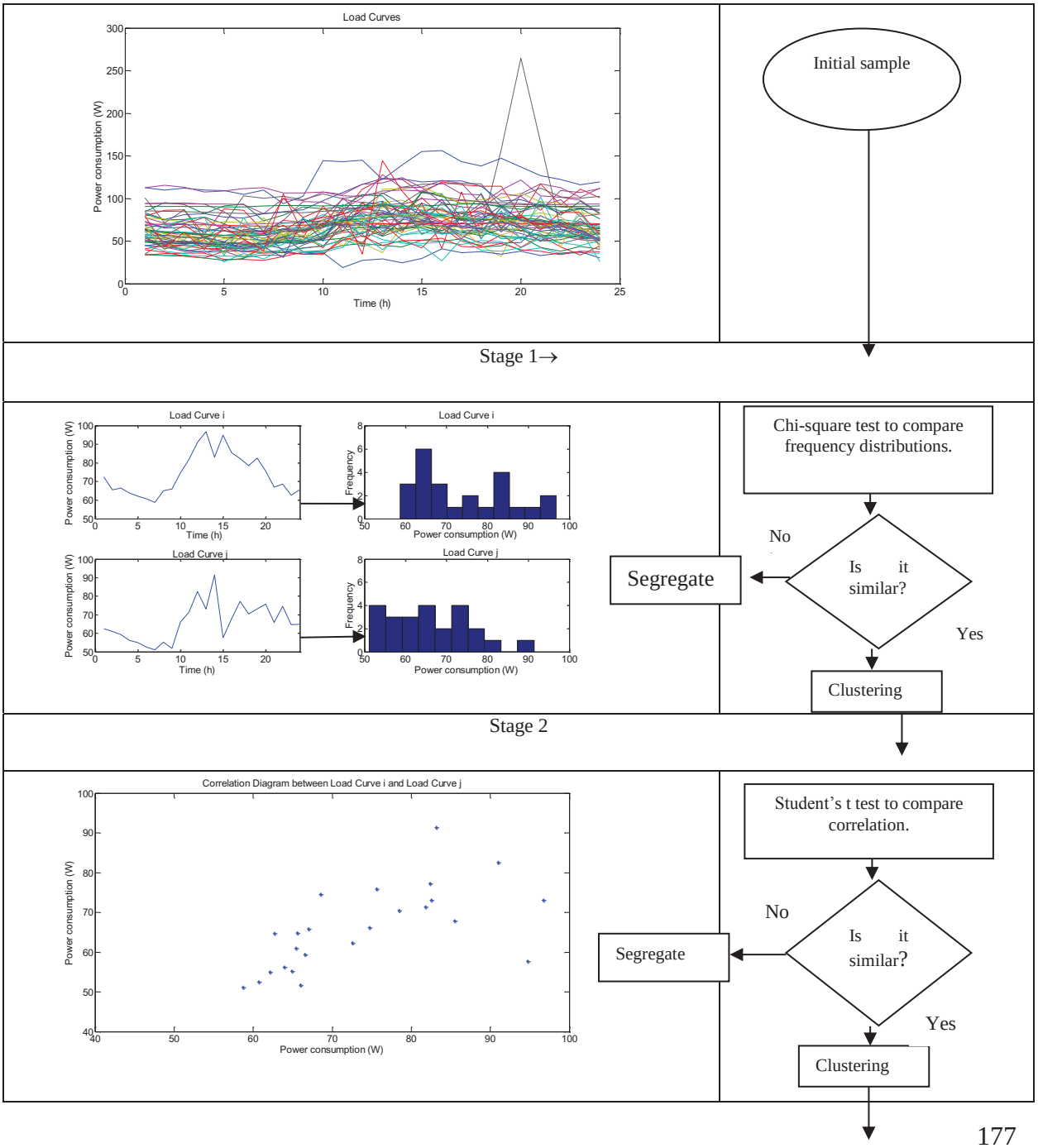
Table 1 – Statistical tests used in CPT-M method.

Initially, each time series belonging to each object is normalized within the interval [0;1] dividing the hourly measurements by the peak value of each. The resulting magnitude is called a score. Table 1 presents the assumptions and statistical tests considered in the first phase, as an extension of the univariate case (STCL, [66]).

Fig. 4 presents the conceptual modeling carried by the sequence of statistical tests performed in the first stage of the proposed method. In order to enable the graphic presentation of the four stages, the illustration is based only on univariate time series (load curves). The extension to the CPT-M method is straight forward through the use of an appropriate statistical test in a multivariate approach (Table 1).

In the first stage a frequency distribution considering the power consumption values for each load curve is represented through a histogram. The similarity between the histograms is checked through the Chi-square test (Table 1) and two load curves (objects) can be grouped together or not.

In the second stage the coefficients of binary correlations between the load curves belonging to the same cluster recognized in stage 1 are obtained. The similarity between these coefficients are checked through the Student's *t* test and new clusters are recognized hierarchically.



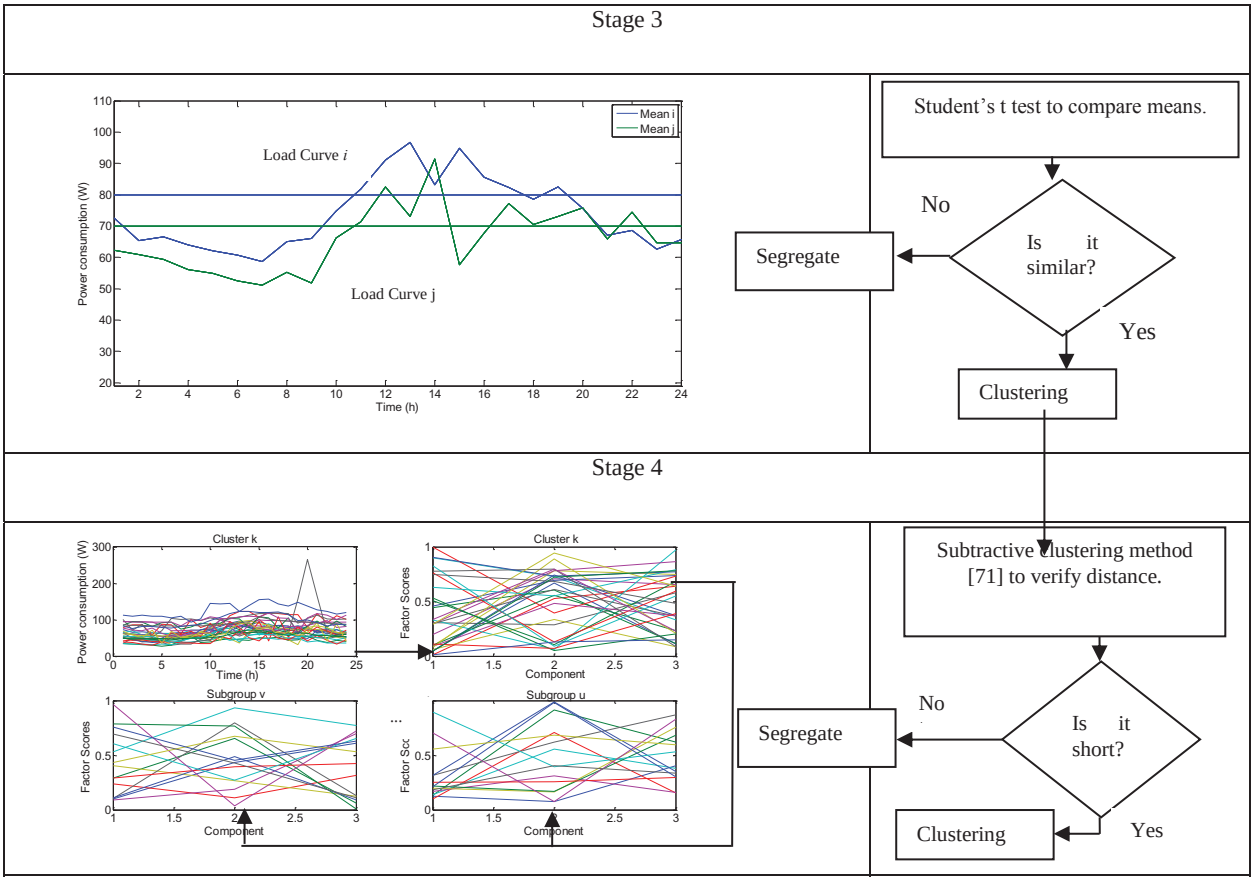


Figure 4 - Phase 1 of the method – illustration considering the univariate case.

In the third stage the level of similarity between the arithmetic means of the load curves belonging to the same cluster recognized in the stage 2 is also checked by the Student's *t* test.

An additional stage (4th stage) checks the similarity between the seasonal behavior of the load curves. The procedure was based on Principal Component Analysis (PCA) through the use of score and loading vectors [62]. The loading vectors are the principal component themselves and each score vector is the projection of the data sample onto the principal component axis. This decomposition is capable of describing the major variances in a data set. In the CPT-M method (multivariate approach) the same procedure may be applied through the reduction of the three time series presented in each object by a single vector of factor scores.

A general data matrix $Y \in \Re^{v \times w}$ (v rows sample points and w measured variables) can be broken down as follows:

$$Y = t_1 \times p_1^T + t_2 \times p_2^T + \ldots + t_\pi \times p_\pi^T + E \quad (1)$$

where t_i is the score vectors ($t_i \in \Re^v$), p_i is the loading vectors (principal component, $p_i \in \Re^w$) and E is the residual matrix [60, 72]. Therefore, the PCA method reduces the original set of

variables to π principal components and the remaining small variances in a residual matrix (E) [62].

The hourly consumptions of load curves belonging to the same cluster recognized in the previous stage are placed in a $v \times 24$ matrix where v is the number of load curves in each cluster. The three principal components of the matrix of each cluster (enough to capture at least 70% of the variability of the objects) are obtained and each load curve is represented by a vector of three factor scores. The vectors of factor scores are grouped according to the Subtractive Clustering Method [73] using the Euclidian distance as similarity metrics.

Although the parametric tests require some assumptions such as the normal distribution of data and homogeneity of variances [74], in this work there was no need to check these assumptions because the values used in the tests are means of samples (means of clusters of variables obtained at each iteration of the algorithm) and the use of parametric tests is supported by the central limit theorem.

In the first three stages, the testing of multivariate statistical hypotheses between the objects is carried out according to Table 1 (CPT-M). There is a fourth stage (Fig. 5) that comprises a multivariate quantification of dissimilarity between the objects generated in the third stage according to their seasonal behaviors. This last stage is carried out in four sub stages (Fig. 5). The first comprises the reduction of dimensionality of the three time series through the achievement of two principal components (enough to capture at least 95% of the variability of the sample data) associated to each object. These principal components are derived from the original variables (energy consumption, outside temperature and fridge temperature) through the decomposition of the covariance matrix of the original variables into two matrices (one with eigenvalues arranged in descending order of impact and the other with their respective eigenvectors or principal components [75]). In the second sub stage, the similarity between each pair of objects is obtained through the PCA (Principal Component Analysis) similarity metrics (*SPCA* [13, 75, 62, 76]) that is obtained using the two largest principal components (PCs) of each object. *SPCA* is a measure of similarity between multivariate time series (Eq. (1)).

$$SPCA_{pq} = 1/4 \sum_{j=1}^{2} \sum_{i=1}^{2} \cos^2 \theta_{ji} \tag{2}$$

θ_{ji} is the angle formed between the "j^{th}" principal component of the "p^{th}" object and the "i^{th}" principal component of the "q^{th}" object$(p,q=1,\ldots,n)$. $SPCA_{pq}$ is the measure of similarity between the "p^{th}" and "q^{th}" objects. Considering a sample with *n* objects SPCA is a *nxn* matrix. For the sake of clearness, Fig. 2 presents the SPCA matrix with hypothetical values.

The dissimilarity matrix (*D*) consists of the dissimilarity metrics between each pair of objects (D_{pq}):

$$D_{pq} = 1 - SPCA_{pq} \tag{3}$$

In the third sub stage, a multidimensional scaling (MDS) representation based on the dissimilarity matrix (D) is performed (Bécavin et al [77]). This analysis begins with the following transformation on each element of dissimilarity matrix:

$$K_{pq} = (-1/2).\left(D_{pq}^2 - D_{.q}^2 - D_{p.}^2 + D_{..}^2\right) \tag{4}$$

where

$$D_{p.}^2 = \sum_{q=1}^{n} D_{pq}^2 / n\,. \tag{5}$$

$$D_{.q}^2 = \sum_{p=1}^{n} D_{pq}^2 / n\,. \tag{6}$$

$$D_{..}^2 = \sum_{p=1}^{n}\sum_{q=1}^{n} D_{pq}^2 / n\,. \tag{7}$$

Resulting in a matrix *K* (*nxn*) that can be decomposed according to $K = V \cdot \Lambda \cdot V^T$. The columns of V (*nxn*) are the corresponding eigenvectors of the matrix K and Λ is a diagonal matrix (*nxn*) with the eigenvalues. The two eigenvectors with the highest eigenvalues are the orthogonal axes used for the multidimensional scaling (MDS). The distance between the points in the multidimensional scaling representation (Fig. 5 – sub stage 3) provides the level of dissimilarity between the original objects in the sample [77]. So, MDS enables the evaluation of the distance between multidimensional objects in a two-dimensional plane.

Finally, in the fourth sub stage one clustering method (date subtractive algorithm [73]) is applied to the points presented in the MDS representation (Fig. 5 – sub stage 4).

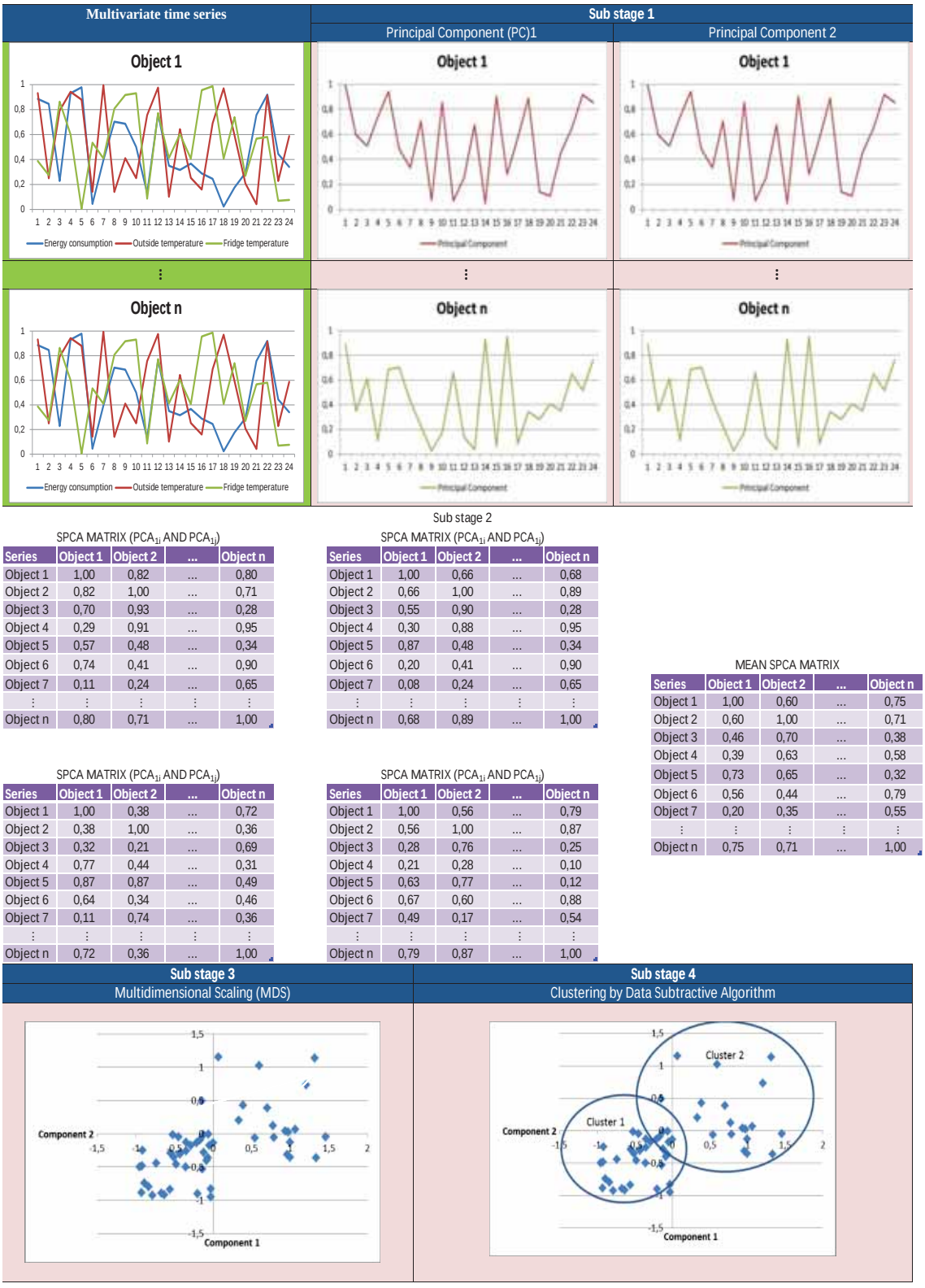


Figure 5 -The fourth stage of CPT-M method.

The first phase is repeated successively in order to verify any possible similarity between some of the object patterns (median of each variable within cluster). This first phase is concluded when there is a convergence in the number of patterns. Thus, the number of patterns is a result of the method itself avoiding the need for an initial estimation.

In the second phase of the CPT-M method (Fig. 3) the objects of the original sample are associated to one of the non point-prototypes (pattern) recognized in the first phase using a non-hierarquical clustering procedure. Each object (matrix with three columns) is represented by its first principal component (enough to capture at least 70% of the variability of the data) and the similarity of the latter from the first principal component of each pattern is measured using the Euclidian distance. Each object is associated with the most similar pattern (smallest Euclidian distance) given the clustering a feature of crisp partition [12]. If only one principal component is not able to capture at least 70% of the variability of the data, the similarity analysis should consider more than one component and a similarity metric suitable for multivariate time series must be employed such as the SPCA (Similarity Principal Component Analysis) (Eq. (2) [75, 62, 76]).

The metrics adopted to measure the clustering quality is the Silhouette Index or Silhouette Value [78]. This index measures the cohesion within and differences between the clusters regardless of the clustering method applied.

Some metrics have been proposed to measure the clustering (partition) quality using a single real value [12, 61]. Some of these are based only on the membership values of the objects and others comprise the object values themselves [79]. The Silhouette Index is a useful and comprehensive metric that has performed well compared to other metrics [80] and is suitable for the analysis of clustering in time series (both univariate and multivariate cases).

The Silhouette Index evaluates the quality of the clusters resulted by an algorithm and the contribution of each object for the overall performance of the clustering.

Considering N_K objects belonging to the K cluster and a total of G clusters ($G \geq 2$), the Silhouette Index for each object (adapted for multivariate time series) is:

$$S_i^L = \frac{b_i^L - a_i^L}{\max\left\{a_i^L, b_i^L\right\}} \qquad i = 1,\ldots,\sum_{k=1}^{G} N_k \tag{8}$$

Where $S_i^L \ \left(-1 \leq S_i^L \leq 1\right)$ is the Silhouette Value of i^{th} object belonging to the L cluster ($1 \leq L \leq G$). a_i^L (Eq. (9)) is the average distance between the i^{th} object and all other objects belonging to the same L cluster. b_i^L (Eq. (10)) is the lowest of the average distances between the i^{th} object and the objects belonging to the other clusters.

$$a_i^L = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{N_L - 1} d_{ij}^{L,L}}{N_L - 1} \qquad i = 1, \ldots, N_L \tag{9}$$

$$b_i^L = \min_{\substack{K=1 \\ K \neq L}}^{G} \left(d_{iK} \right) \quad \text{onde} \quad d_{iK} = \frac{\sum_{j=1}^{N_K} d_{ij}^{L,K}}{N_K} \qquad i = 1, \ldots, N_L \tag{10}$$

In the second phase of the method, considering that each object (three time series) can be represented by only one vector (principal component with highest eigenvalue), $d_{ij}^{L,K}$ is the Euclidian distance between the principal component of the i^{th} object, belonging to the L cluster, and the principal component of the j^{th} object belonging to the K cluster.

The distribution of the Silhouette Indexes of all objects allows us to evaluate the level of overlap between the two clusters and therefore the quality of the clustering. According to Eq. (8), when the membership of an object relative to a given cluster is well defined, the corresponding Silhouette Index approaches unity. On the other hand, negative values imply that there is more homogeneity between clusters and less internal cohesion, which represents the worst scenario. In this work, the mean between the Silhouette Values obtained for all objects (General Silhouette Index – GSI) was also adopted to evaluate the clustering quality.

3 CASE STUDY & RESULTS

This program essentially involved the replacement of 5,000 old refrigerators for new ones in low-income communities. The sampling procedure was performed based on determining the sample size for estimating the mean [74]. Data from a sample of 54 old refrigerators with high electricity consumption (case I, average consumption of 35 kWh) and another sample of 54 new refrigerators (case II, average consumption of 18 kWh) were available. The sample size (54 consumers in both cases I and II) gives an error of 8.7% and 6.0% for the cases I and II, respectively, and confidence levels of 94.3% and 94.3% in the prediction of the population parameter. The International Performance Measurement & Verification Protocol (IPMV) recommends a sampling error of up to ± 10% [81]. Furthermore, the target population (5,000 homes) comprises a subset of the class of residential consumers (low-income consumers) which implies a lower variability in seasonal consumer behavior. The variability of the population could be even smaller in case II because the replacement of the refrigerators (all

consumers with new ones with the same technical specifications) helps to homogenize the consumption profiles.

Initially the data was analyzed to identify and eliminate outlier objects. This procedure comprised the decomposition of original sample (data set) through PCA [82] into a number of score and loading vectors. This decomposition is capable of describing the major variances in the data set and can provide the identification and elimination of outliers from the original sample.

Based on the original samples of each case (54 objects and each one with 24 measurements for each variable, i.e. load, external and internal temperatures), a data matrix $Y \in \Re^{52\times74}$ was decomposed according to Eq. (1) and the original samples were reduced to 4 (Case I) and 6 (case II) score vectors considering in each case the description of at least 50% of the variability in the data. Figs. 6 and 7 show the distribution of scores (elements) of each score vector for the cases I and II. The points identified in box-plots [83] by the sign "+" suggest outlier objects. Therefore, the scores were useful for the identification of outliers according to a multivariate approach. This analysis enabled the exclusion of 3 outlier objects in case I and 20 outlier objects in case II. The increase in the number of factors (number of box-plots) in case II is associated to the greater seasonal variation caused by the lower energy demand of the motors of the new refrigerators (as will be shown later). Even with the removal of outliers these errors increased to 9.0% and 7.5% respectively.

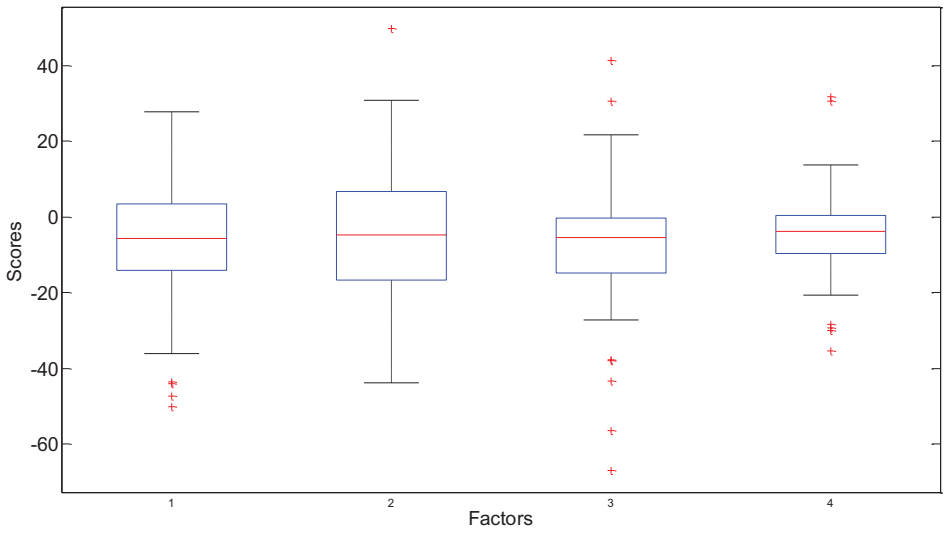


Figure 6 – Distribution of values in each factor (case I).

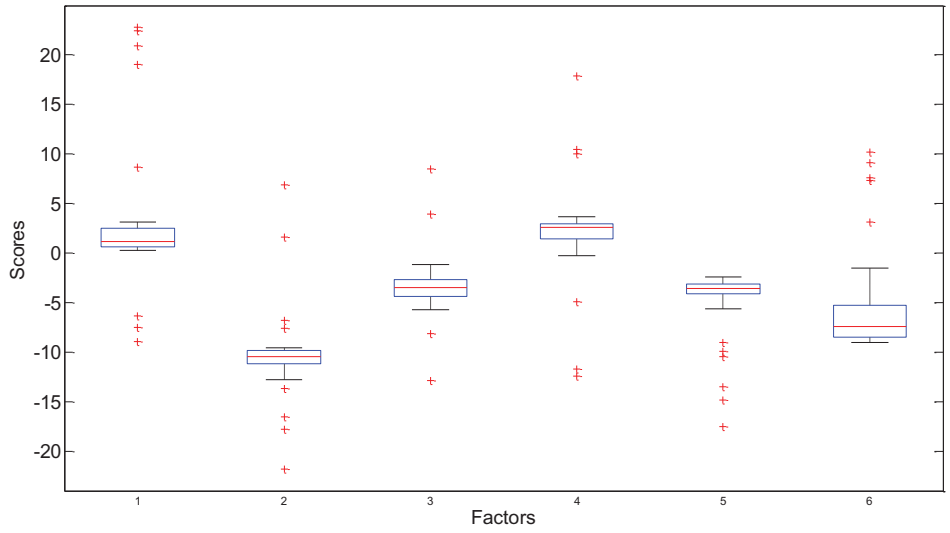


Figure 7 – Distribution of values in each factor (case II).

The results obtained using the CPT-M method were compared with the *Fuzzy C-Means* (FCM), a well-known method belonging to the C-Means families of batch clustering models [84], suitable for clustering objects represented by time series [13]. In this case, the basic formulation of the FCM algorithm [12], suitable for a point-prototype problem, had to be changed in order to use the *SPCA* instead of the Euclidian metric of similarity and a new formulation for the FCM method [58]. Even with this modified version, only one tuning parameter of the optimization problem was considered (degree of fuzzification of the partition, $m \geq 1$). This parameter was set equal to 2 according to what is suggested in the literature [12, 59].

The application of the CPT-M method in both cases was able of recognize the existence of two clusters and two patterns. The Global Silhouette Index (GSI) obtained by the CPT-M method in cases I (GSI = 0.19) and II (GSI = 0.46) and the indexes obtained by the FCM method (case I, GSI = - 0.12 and case II, GSI = 0,21) suggest that this last method presented, in both cases, a clustering quality lower than the CPT-M (Figs. 8 and 9). Furthermore, the patterns recognized by the FCM method could be converted into a single pattern in both cases I and II. This may be verified through Figs. 10-13 that present, for each cluster, the parity graphic using the two principal components associated to each object and also to the pattern recognized (multidimensional scaling). The proximity between the principal components of the patterns recognized by the FCM method suggests that CPT-M was able to identify clusters more heterogeneous. The arrangement of the principal components in both cases also at tests the best clustering quality obtained by the CPT-M method.

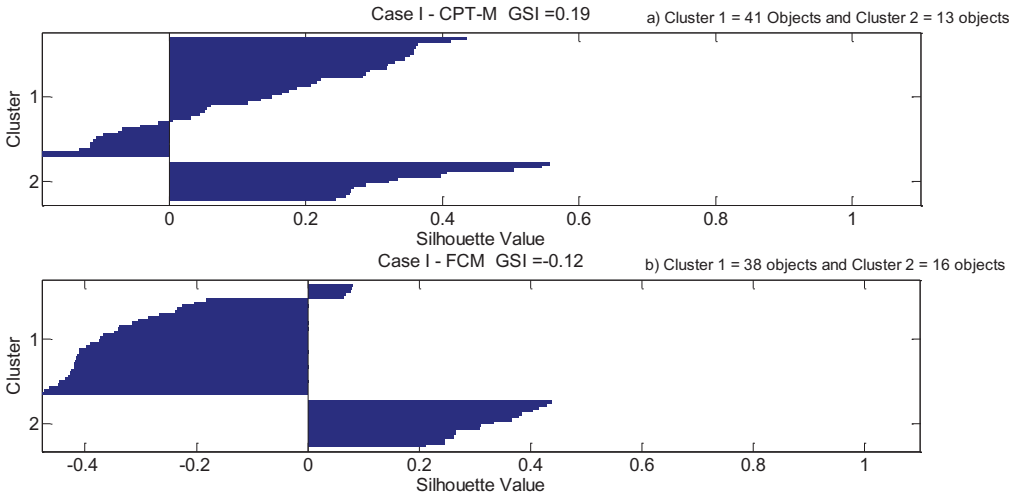


Figure 8 – Silhouette Indices via the CPT-M and FCM methods (case I).

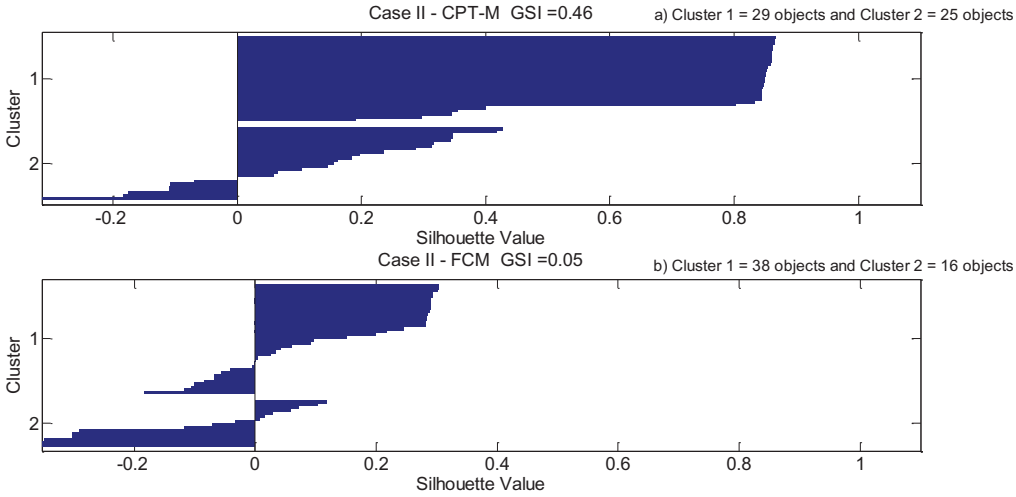


Figure 9 – Silhouette Indices via the CPT-M and FCM methods (case II).

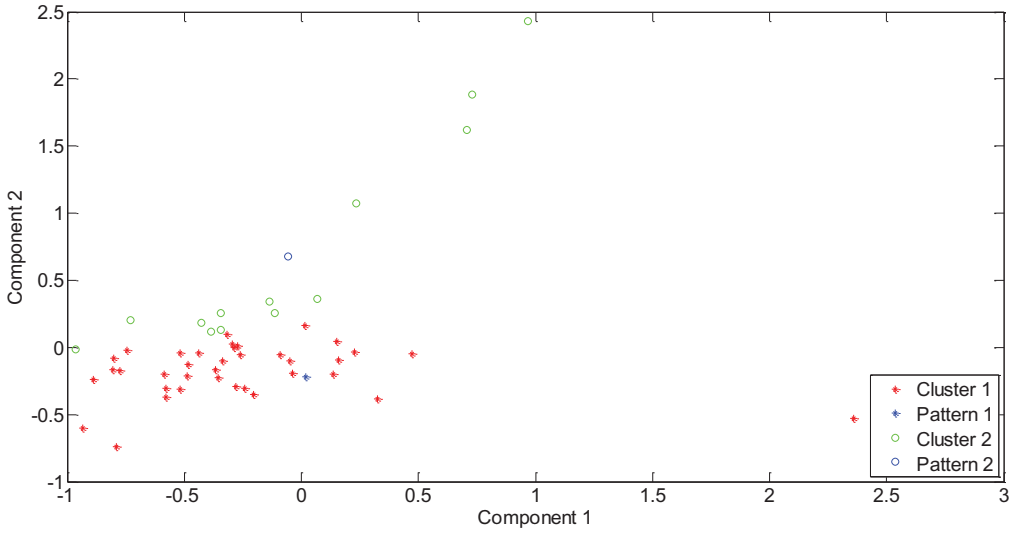


Figure 10 – Objects and Patterns obtained by CPT-M represented by principal components – case I.

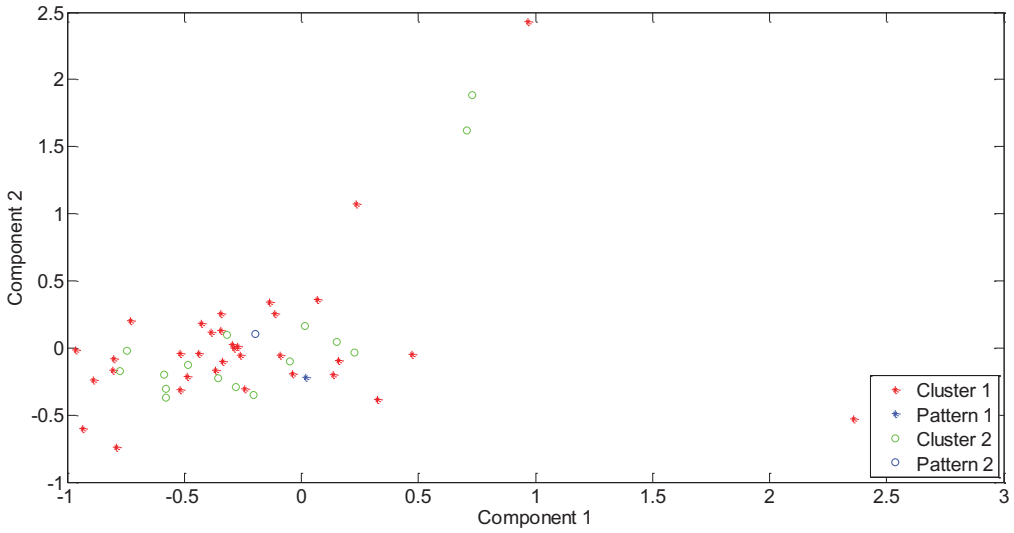


Figure 11 - Objects and Patterns obtained by FCM represented by principal components – case I.

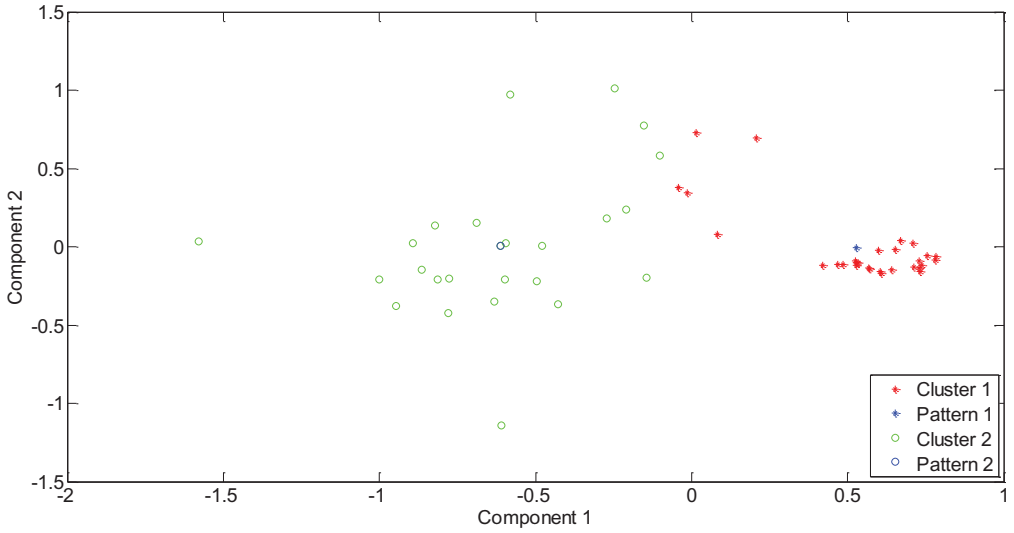


Figure 12 - Objects and Patterns obtained by CPT-M represented by principal components – case II.

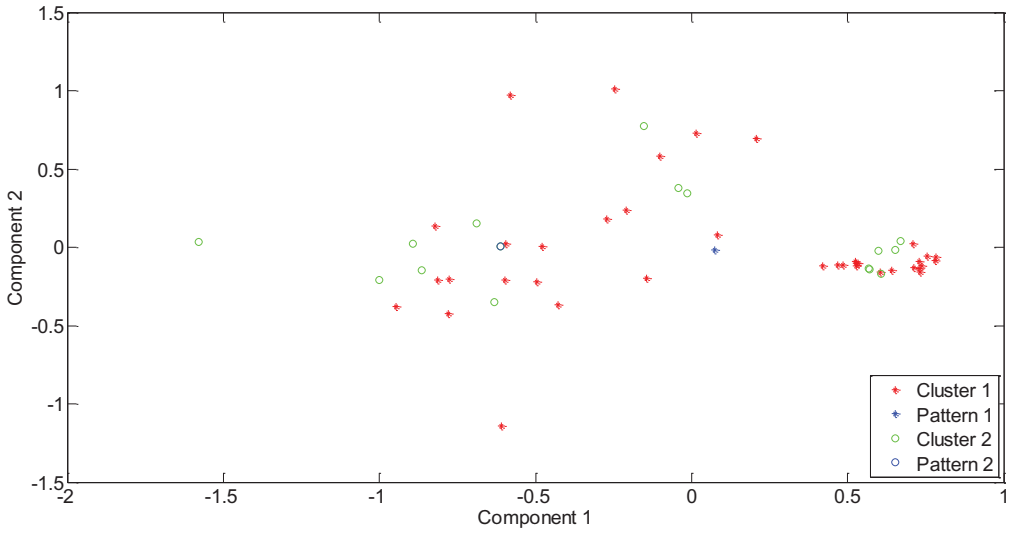


Figure 13 - Objects and Patterns obtained by FCM represented by principal components – case I.

The Figs. 8-13 present a comparison between the results obtained from CPT-M and FCM methods considering different validation criteria associated to the clustering quality.

In Figs. 8 and 9 the clustering quality according to the features of cohesion and separation of the clusters is analyzed through the distribution of the Silhouette index of all the objects in each cluster recognized. The CPT-M results for case I (Fig. 8a) shows that cluster 1 has about 33% of objects with negative Silhouette indexes and cluster 2 (amount of objects less than half of the first cluster) does not present negative Silhouette indexes. Furthermore, the FCM results for case I (Fig. 8b) shows that cluster 1 has about 90% of the objects with negative Silhouette indexes. In case II the cluster 1 (Fig. 9a) recognized by CPT-M does not present

negative silhouette indexes. The results obtained with FCM for case II shows that the CPT-M method provided a better clustering quality and the high difference between the number of objects in the partition process (CPT-M, 29 and 25, FCM, 38 and 16) suggests that the FCM method had difficulty recognizing the existence of two clusters of consumers after the replacement of refrigerators. The Global Silhouette Index (GSI) obtained by the CPT-M method in cases I (GSI = 0.19) and II (GSI = 0.46) and the results obtained by the FCM method (case I, GSI = – 0.12 and case II, GSI = 0.21) also suggest that this last method presented a clustering quality which was lower than the CPT-M in both cases.

Figs. 10-13 are parity graphs which use the two principal components associated to each object in cluster and also to the pattern recognized. Adopting an approach similar to what has been accomplished in identifying outliers, the data matrix $Y \in \Re^{52\times 74}$ was broken down in 2 principal components. This is also a way to analyze the cohesion of each group and the level of overlap between the respective objects. The proximity between the principal components of the patterns recognized by the FCM method suggests that CPT-M was able to identify more heterogeneous clusters. The arrangement of the principal components in both cases also demonstrates the best clustering quality obtained by the CPT-M method. Furthermore, the patterns recognized by the FCM method could be converted into a single pattern in both cases I and II.

The inverse of the Coefficient of Performance (COP^{-1}), used to quantify the thermal efficiency, is the ratio $\frac{T_o - T_f}{T_f}$ (T_o and T_f are external and internal fridge temperatures respectively). It is known that the analysis of COP^{-1} is more appropriate for the electric sector than the analysis of temperatures (internal fridge and external) when dealing with cooling equipment [8, 85]. Based on the patterns (time series) recognized for each variable (load, internal and external temperatures), Figs. 11-12 present a joint analysis involving the thermal and energy efficiencies (the first quantified by the COP^{-1} and the second quantified by the load) of the refrigerators for both cases I and II. The daily profile of the external temperature (T_f) in the region analyzed together with the adoption of a constant set point for the fridge temperature (T_0) establish a typical profile for COP^{-1}, in both cases I and II. Both methods (FCM and CPT-M) show a greater power consumption (lower energy efficiency) before the replacement of refrigerators (case I) in order to ensure the thermal efficiency profile (pretty

much the same in cases I and II). This result attests the success of the energy efficiency program.

In case I in particular, the CPT-M method (Fig. 14) was capable of recognizing two patterns (both associated to demand and coefficient of performance) which were more distinct from each other than in the FCM method, showing the ability of CPT-M to recognize slight differences in consumption profiles. More specifically, the COP^{-1} patterns obtained by the CPT-M method are different during the afternoon whereas the FCM method does not identify this difference which demonstrates the advantage of this method. On the other hand, the proximity between the patterns recognized in case II highlights, as expected, the uniformity in consumption behavior due to the replacement of the old fridge with a new one of equipment. Even in this case, the CPT-M was capable of recognizing two patterns of consumption showing slight differences between these.

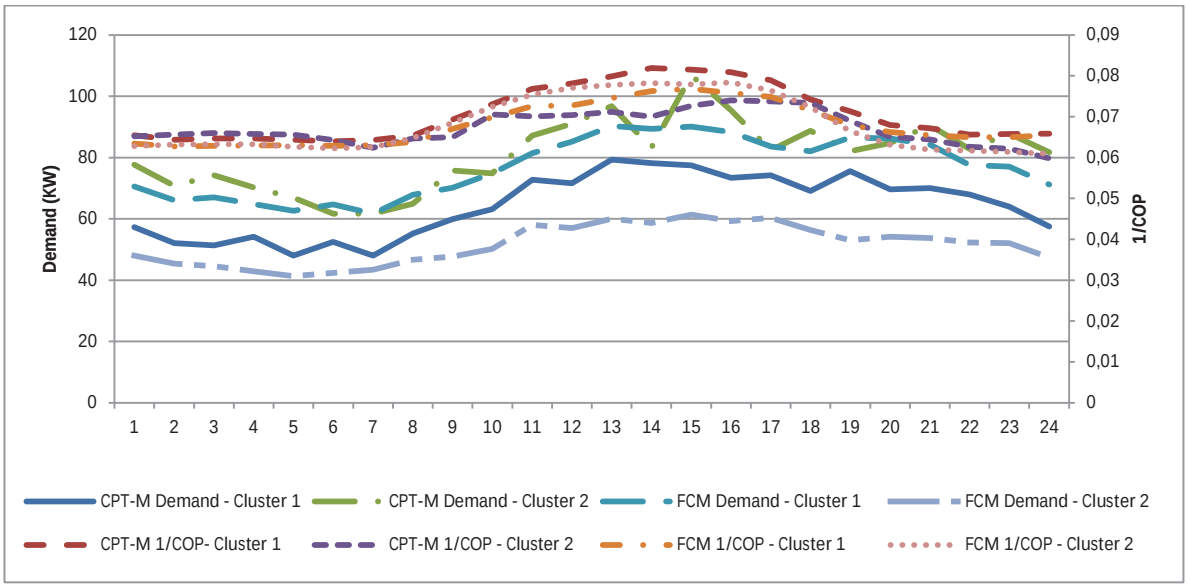


Figure 14–Patterns of the motor efficiency and energy demand - case I.

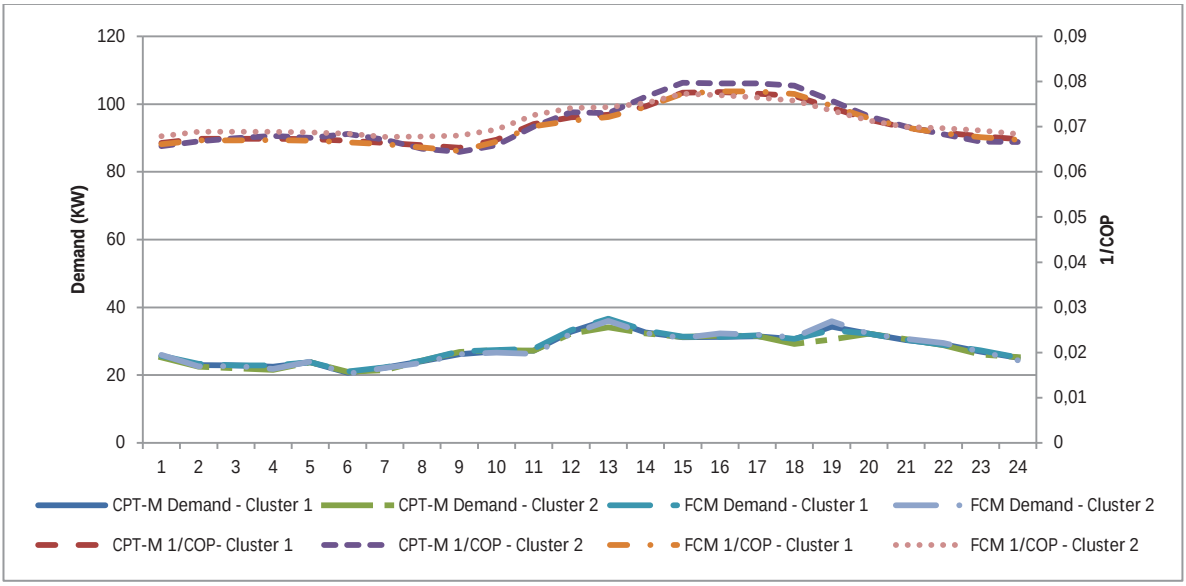


Figure 15–Patterns of the motor efficiency and energy - case II.

The results show that the CPT-M method is capable of identifying a greater diversity in demand patterns and can be used as a potential tool for improved decision making through better classification of heterogeneous consumer profiles.

Figs. 16 and 17 present a joint analysis of seasonality considering only the modal clusters (cluster with the highest number of objects) in each case. Each object in the cluster was transformed into a single vector containing the principal component. Each vector of the principal component represents a possible trajectory and the set of these is a sample of the underlying stochastic process [86]. The seasonality analysis in the cluster can be performed through the autocorrelation analysis of the set of these vectors [87]. The existence of two trends in the patterns recognized (Figs. 14-15) suggests, in this case, the application of autocorrelation analysis on the first order differences on these vectors in order to mitigate non-stationary effects [88]. The first peaks in Figs. 16-17 are associated to the non-stationary conditions and show high autocorrelation that decreases over time reaching levels associated with steady-state conditions. In these conditions the peaks occur intermittently (not consecutively) showing the seasonal behavior of the time series (load and temperatures) analyzed. According to Fig. 13 there are seasonal peaks in lags 5 (early morning) and 17 (late afternoon), among others, which indicate changes in the level of consumption. This coincides with the seasonal peak of the Brazilian electric system, possible associated with consumption habits, suggesting the need of actions to improve the demand profile. Fig. 17 (case II) presents a larger number of peaks due to the greater frequency of engine shutdown. This is consistent

because the new refrigerator has better thermal insulation meaning that the motor consumes less power (also attested by Fig. 15).

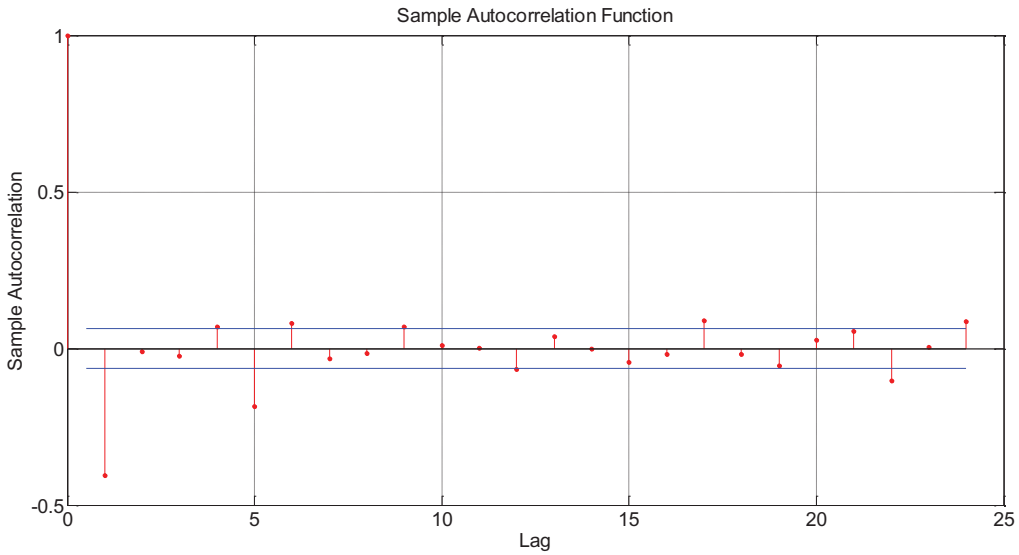


Figure 16 - Autocorrelation values of the first order difference (modal cluster - case I, CPT-M method).

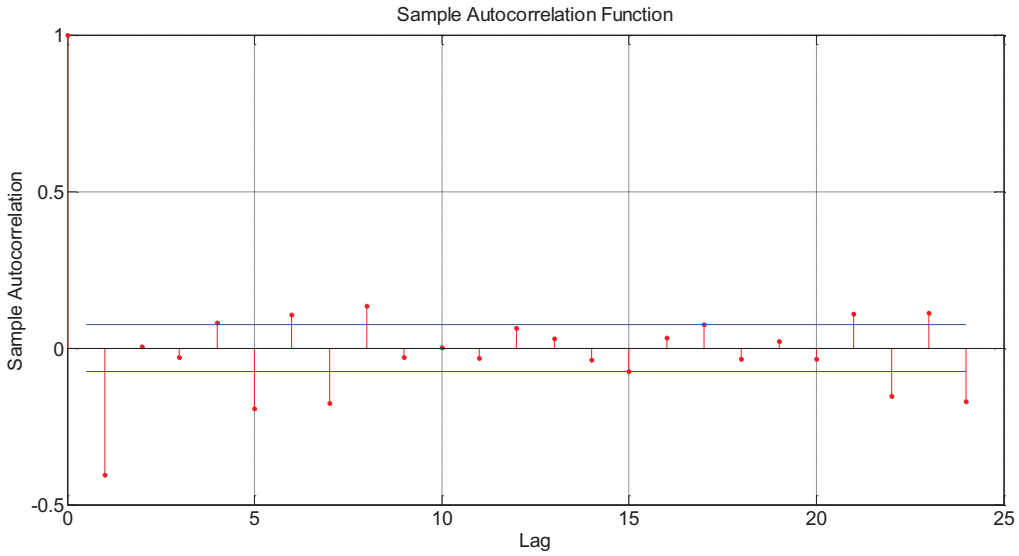


Figure 17 –Autocorrelation values of the first order difference (modal cluster - case II, CPT-M method).

Additional discussion and results should be presented about the contribution of the proposed method to support the decision-making in the energy distribution sector. Indeed, this decision-making process is associated to the performance evaluation of the energy efficiency program carried out by the Electric Company and this evaluation can be performed through a model identification procedure, based on the pattern recognized, in order to estimate the effective gain obtained by the replacement of the equipment.

The pattern (Multivariate Time Series) recognized with the old equipment (Case I, before the replacement of the refrigerators) can be used to identify a dynamic model in which the energy consumption (output variable) is a function of the Coefficient of Performance (COP^{-1}), i.e. the energy consumption is an effect of the thermal efficiency of the equipment. Fig. 18 presents the predicted (projected) consumption for the pattern associated to the modal cluster of Case I (old refrigerators) considering the same Coefficient of Performance (COP^{-1}) (same thermal efficiency condition) of the pattern associated to the modal cluster of Case II (new refrigerators). According to the International Protocol for Measurement and Verification of Performance (IPMVP) [81], the baseline (smoothed consumption) is a useful reference to represent the modal cluster pattern (recognized based on the original load curves), and the adjusted baseline (predicted consumption) provides a comparison with the consumption profile pattern after replacing the refrigerators, contributing to an effective assessment of the energy efficiency program. This analysis also supports decision-making processes at the management level, contributing (or not) to expand and consolidate the program to include other consumers. Furthermore, the gain or energy savings can be useful in decision-making

regarding the postponement of future spending related to the expansion of the energy supply considering generation, transmission and distribution. According to the National Agency of Electrical Energy of Brazil [89], energy saving programs are also important for obtaining subsidies from the government.

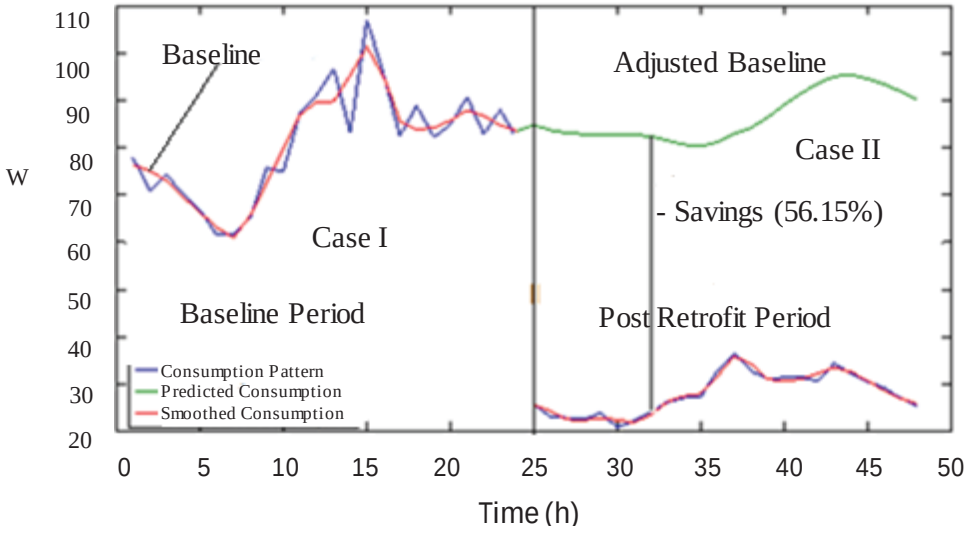


Figure 18 – Consumption /Smoothed Consumption (Patterns of the modal cluster – Case I and Patterns of the modal cluster – Case II) and Prediction Consumption for the old refrigerators based on the thermal conditions of Case II.

The predicted consumption presented in Fig. 3 was performed using an ARX (AutoRegressive with Exogeneous input) model [93-96] according to Eq. (11). The identification procedure also comprised a pre-filtering of the data in order to reconstruct the signal with adequate noise smoothing.

$$A(q)y(k) = B(q)u(k) + v(k) \tag{11}$$

where,

$$A(q) = 1 - a_1 q^{-1} - \ldots - a_{n_y} q^{-n_y}; \tag{12}$$

$$B(q) = b_1 q^{-1} + \ldots + b_{n_u} q^{-n_u}; \tag{13}$$

and q^{-1} is the backward shift operator $\left(q^{-1} \cdot u(k) = u(k-1)\right)$, $v(k)$ is the white noise, u is the input, y is the output signal and k is the discrete time. In this work, the best adjustment was achieved with $n_y = n_u = 1$.

$$\left(1 - q^{-1}\right) \cdot COP^{-1}(k) = q^{-1} \cdot PC(k) + v(k) \tag{14}$$

where PC is the output (power consumption).

Based on the results and on the sampling procedure, the expected reduction in electricity consumption provided by the energy efficiency program carried out by the Electric Company of Alagoas (Brazil) would be around 56 % which represents a gain of 250,000.00 kWh considering the total population analyzed (five thousand residential consumers).

The advantages of the proposed method can be highlighted by the following:

- Identification of a greater diversity of patterns;
- Recognition of seasonality through a multi criteria approach;
- Improving decision making through better classification of heterogeneous consumer profiles;
- Setting the number of clusters through a semi-hierarchical clustering approach.

4 CONCLUSION

This works presents a new method for the selection, pattern and clustering of multivariate time series (CPT-M) suitable for the electric power sector and especially in the analysis of consumption profiles associated to the refrigeration equipment. The algorithm is suitable for unlabeled data and comprises four steps that extract essential features of multivariate time series of residential users with emphasis on seasonal and temporal profile, among others. The results demonstrate its good performance in recognizing patterns in samples with heterogeneous data (common situation in the electric power sector). Unlike the typical C-means models of clustering, the number of clusters is also a result obtained by CPT-M method.

The CPT-M method is capable of recognizing different clusters using multiple criteria to recognize patterns which are difficult to identify using traditional methods and the clustering quality considering the levels of cohesion and separation between clusters. The multiple criteria approach also contributes to recognizing seasonalities in time series which is useful for the identification of prediction models to support the decision-making related to the implementation of energy efficiency programs. On the other hand, the second phase of the CPT-M method comprises the clusters formation and only one criteria (geometric distance) is applied in this phase.

The case studied looked at an energy efficiency program carried out by the Energy Company of Alagoas (CEAL-Brazil) which analyzed the impact of replacing 5,000 old refrigerators with new ones for low-income consumers. The results obtained by CPT-M, compared to a well-known method of clustering (Fuzzy C-Means, FCM), reveal the viability and potential of the former in recognizing patterns and in generating conclusions coherent with the reality of the electric power sector. This supports the implementation of efficiency actions based on real features within the consumer market and can also support decision-making at management level.

The capability of including fridge and outside temperatures together with the load curves provides the possibility to jointly analyze the thermal and energy efficiencies of the equipment and identify abnormal behavior of consumption associated to the environment temperature and the performance of its thermal control.

REFERENCES

[1] Tsekouras G. J., Tsaroucha M. A., Tsirekis C. D., Salis A. D., Dialynas E. N., Hatziargyriou N. D. A database system for power systems customers and energy efficiency programs. Electric Power Energy Syst 2011;33:1220–8.

[2] Lin J. K., Tso S. K., Ho H. K., Mak C. M., Yung K. M., Ho Y. K. Study of climatic effects on peak load and regional similarity of load profiles following disturbances based on data mining. Electric Power Energy Syst 2006;28:177–85.

[3] Chicco G., Napoli R., Postolache P., Scutariu M., Toader C. Customer characterization options for improving the tariff offer. IEEE Power Eng Rev 2002;22(11):60.

[4] Silk J. I., Joutz F. L. Short and long-run elasticity's in US residential electricity demand: a cointegration approach. Energy Econom 1997;19(4):493–513.

[5] Geller H. et al. The efficient use of electricity in Brazil: progress and opportunities. Energy Policy 1998;26(11):859–72.

[6] Goldman Charles A., Hopper Nicole C., Osborn Julie G. Review of US ESCO industry market trends: an empirical analysis of project data. Energy Policy 2005;33:387–405.

[7] Clinch J. P., Healy J. D. Cost-benefit analysis of domestic energy efficiency. Energy Policy 2001;29:113–24

[8] Stoecker Wilbert F. Industrial refrigeration handbook. New York: McGraw-Hill;1998.

[9] Yusri Syam Akil, Hajime Miyauchi. Seasonal peak characteristic comparison analysis by hourly electricity demand model. Int J Energy Power Eng 2014;3(3):132–8.

[10] Lia T. W. Clustering of time series data – a survey. Sci Direct. Pattern Recognit 2005;38:1857–74.

[11] Keogh E. J., Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min Knowl Disc 2003;7(4):349–71.

[12] Bezdek James C., Keller James, Krisnapuram Raghu, Pal Nikhil R. Fuzzy models and algorithms for pattern recognition and image processing. Springer; 2005.

[13] Singhal A., Seborg D. E. Clustering multivariate time-series data. J Chemometr 2005;19:427–38.

[14] Rani Sangeeta, Sikka Geeta. Recent techniques of clustering of time series data: a survey. Int J Comput Appl 2012;52:1–9.

[15] D'Urso Pierpaolo, Maharaj Elizabeth Ann. Walets-base clustering of multivariate time series. Fuzzy Sets Syst 2012;193:33–61.

[16] Coppi Renato, D'Urso Pierpaolo, Giodani Paolo. A fuzzy clustering model multivariate spatial time series. J Classification 2010;27:54–88.

[17] Kavitha V., Punithavalli M. Clustering time series data stream – a literature survey. Int J Comput Sci Inform Secur 2010;8:289–94.

[18] Gao Zhong-Ke, Zhang Xin-Wang, Jin Ning-De, Donner Reik V, Norbert M, et al. Recurrence networks from multivariate signals for uncovering dynamic transitions of horizontal oil–water stratified flows. Europhys Lett 2013;103.50004p1–.50004p6.

[19] Gao Zhong-Ke, Zhang Xin-Wang, Jin Ning-De, Norbert M, Jürgen K. Multivariate recurrence network analysis for characterizing horizontal oil–water two-phase flow. Phys Rev E 2013;88(3). 032910_1–.032910_12.

[20] Gao Zhong-Ke, Fang Peng-Cheng, Ding Mei-Shuang, Jin Ning-De. Multivariate weighted complex network analysis for characterizing nonlinear dynamic behavior in two-phase flow. Exp Thermal Fluid Sci 2015;60:157–64.

[21] Chicco Gianfranco. Overview and performance of the clustering methods for electrical load pattern grouping. Energy 2012;42(1):68–80.

[22] Tsekouras G. J., Hatziargyriou N. D., Dialynas E. N. Two-stage pattern recognition of load curves for classification of electricity customers. IEEE Trans Power Syst 2007;22(3):1120–8.

[23] Chicco G., SumailiAkilimali J. Renyi entropy-based classification of daily electrical load patterns. IET Gener Transm Distrib 2010;4(6):736–45.

[24] Chicco G., Napoli R., Postolache P., Scutariu M., Toader C. Customer characterization options for improving the tariff offer. IEEE Trans Power Syst 2003;18(1):381–7.

[25] Chicco G., Napoli R., Piglione F., Scutariu M., Postolache P., Toader C. Load pattern-based classification of electricity customers. IEEE Trans Power Syst 2004;19(2):1232–9.

[26] Chicco G., Napoli R., Piglione F., Scutariu M., Postolache P., Toader C. Emergent electricity customer classification. IEE Proc Gener Transm Distrib 2005;152(2):164–72.

[27] Yu I. H., Lee J. K., Ko J. M., Kim S. I. A method for classification of electricity demands using load profile data. Proc. Fourth Annual ACIS Intern. Conf Comput Inf Sci; 2005, p. 164–8.

[28] Carpaneto E., Chicco G, Napoli R, Scutariu M. Electricity customer classification using frequency-domain load pattern data. Electric Power Energy Syst 2006;28(1):13–20.

[29] Chicco G., Napoli R., Piglione F. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006;21(2):933–40.

[30] Gerbec D., Gasperic S., Smon I., Gubina F. Determining the load profiles of consumers based on fuzzy logic and probability neural networks. IEE Proc Gener Transm Distrib 2004;151(3):395–400.

[31] Nazarko J., Jurczuk A., Zalewski W. ARIMA models in load modelling withclustering approach. Proc. IEEE power Tech, St. Petersburg, Russia; 2005, p. 27–30.

[32] Marques D. Z., de Almeida K. A., de Deus A. M., da Silva Paulo A., da Silva Lima W. A comparative analysis of neural and fuzzy cluster techniques applied to the characterization of electric load in substations. Proc. IEEE/PES transmission and distribution conference and exposition: Latin America; 2004, p. 908–13.

[33] Ramos S., Vale Z., Santana J., Duarte J. Data mining contributions to characterize MV consumers and to improve the suppliers–consumers settlements. Proc IEEE/PES Gen Meeting; 2007, p. 24–8.

[34] Figueiredo V., Rodrigues F., Vale Z., Gouveia J. B. An electric energy consumer characterization framework based on data mining techniques. IEEE Trans Power System 2005;20(2):596–602.

[35] Batrinu F., Chicco G., Napoli R., Piglione F., Scutariu M., Postolache P., Toader C. Efficient iterative refinement clustering for electricity customer classification. Proc. IEEE power Tech. St. Petersburg (Russia); 2005, p. 1–7.

[36] Lamedica R., Fracassi G., Martinelli G., Prudenzi A., Santolamazza L. A novel methodology based on clustering techniques for automatic processing of MV feeder daily load patterns. Proc. IEEE PES summer meeting 2000. Vol. 1. Seattle (WA); 16–20 July 2000, p. 96–101.

[37] Verdu S. V., Garcia M. O., Senabre C., Marin A. G., Franco F. J. G. Classification, filtering, and identification of electrical customer load patterns through the use of self organizing maps. IEEE Trans Power System 2006;21(4):1672–82.

[38] Gerbec D., Gasperic S., Smon I., Gubina F. Allocation of the load profiles to consumers using probabilistic neural networks. IEEE Trans Power System 2005;20(2):548–55.

[39] Valero S., Ortiz M., Senabre C., Alvarez C., Franco F. J. G., Gabaldon A. Methods for customer and demand response policies selection in new electricity markets. IET Gener, Transm Distrib 2007;1(1):104–10.

[40] Tsekouras G. J., Kotoulas P. B., Tsirekis C. D., Dialynas E. N., Hatziargyriou N. D. A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. Electric Power Syst Res 2008;78(9):1494–510.

[41] Räsänen T., Voukantsis D., Niska H., Karatzas K., Kolehmainen M.. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl Energy 2010;87(11):3538–45.

[42] Chicco G., Ilie I. S. Support vector clustering of electrical load pattern data. IEEE Trans Power Syst 2009;24(3):1619–28.

[43] Nazarko Joanicjusz, Styczynski Zbigniew A. Application of statistical and neural approaches to the daily load profiles modeling in power distribution systems. IEEE; 1999. p. 320–5.

[44] Gerbec D., Gasperic S., Smon I., Gubina F. A methodology to classify distribution load profiles. Presented at the IEEE; 2002, p. 848–51.

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

[46] Geminagnani M. M. F., Oliveira C. C. B., Tahan C. M. V. Proposition and comparative analysis of alternative selection and classification of load curve for defining types for tariff studies. Décimo Tercer EncuentroReginal Iberoamericano de Cigré– XIII ERIAC; 2009, p. 1–6.

[47] Zalewski W. Application of fuzzy inference to electric load lustering. IEEE international conference on power systems. New Delhi; 2006, p. 1–5.

[48] Nizar A. H., Dong Z. Y., Zhao J. H. Load profiling and data mining techniques in electricity deregulated market. Presented at the IEEE power engineering society (PES) General Meeting 2006. Montreal, Quebec (Canada); June 2006, p. 1–7.

[49] Han J., Pei J., Yiwen Y. Mining frequent patterns without candidate generation. In: Proceedings ACM-SIGMOD international conference on management of data. ACM Press; 2000. p. 1–12.

[50] Silva Daswin, Yu Xinghuo. A data mining framework for electricity consumption analysis from meter data. IEEE Trans Industr Inf 011;7(3): 399–407.

[51] Gerbec D., Gasperic S., Smon I., Gubina F. Determining the load profiles of consumers based on fuzzy logic and probability neural networks. IEE Proc- Gener Transm Distrib 2004;151(3):395–400.

[52] Zakaria Zuhaina, Lo K. L., Sohod Hadi Mohamad. Application of fuzzy clustering to determine electricity consumers' load profiles first international power and energy conference. Putrajaya (Malaysia); 2006, p. 99–103.

[53] Nizar A. H., Dong Z. Y., Wang Y. Power utility nontechnical loss analysis with extreme learning machine model. IEEE Trans Power Syst 2008;23(3): 946–55.

[54] Nizar A., Dong Z., Jalaluddin M., Raffles M. Load profiling method in detecting non-technical loss activities in a power utility. In: Proceedings of the IEEE international power and energy conference; 2006, p. 82–7.

[55] Nagi J., Mohammad A., Yap K., Tiong S., Ahmed S. Nontechnical loss analysis for detection of electricity theft using support vector machines. In: Proceedings of the 2nd IEEE international power and energy conference; 2008, p. 907–12.

[56] Monedero I., Biscarri F., Le´ on C., Biscarri J., Mill´an R. Midas: detection of non-technical losses in electrical consumption using neural networks and statistical techniques. In: Proceedings of the international conference on computational science and applications, lecture notes in computer science, vol. 3985. Berlin/Heidelberg: Springer; 2006. p. 725–34 (5).

[57] Anuar N, Zakaria Z. Cluster validity analysis for electricity load profiling. IEEE international conference on power and energy. Kuala Lumpur (Malaysia); 2010, p. 35–8.

[58] Cristiano H. O. Fontes, Carlos A. Cavalcante, Otacílio J. Pereira, Sergio T. Barreto, Luciana Pacheco, Welinton leite. Pattern recognition using multivariable time series for fault detection in a thermoelectric unit. Comput-Aid Chem Eng 2012;31:315–9.

[59] Hoppner Frank, Klawoon Frank, Kruse Rudolf, Runkler Thomas. Fuzzy cluster analysis – methods for classification. Data analysis and image recognition. John Wiley & Sons LTD; 2000.

[60] Theophano Mitsa. Temporal Data Mining. Boca Raton (FL): Chapman & Hall/ CRC Data Mining and Knowledge Discovery Series; 2010.

[61] Kiyoung Yang, Cyrus Shahabi. A PCA-based similarity measure for multivariate timeseries. MMDB ‘04 Proceedings of the 2nd ACM international workshop on multimedia databases; 2004, p. 65–74.

[62] Li Shun, Wen Jin. Application of pattern matching method for detecting faults in air handling unit system. Automat Construct 2014;43:49–58.

[63] Warren Liao T. Clustering of time series data-a survey. Pattern Recogn 2005;38(11):1857–74.

[64] Chaovalit P., Gangopadhyay A., Karabatis G., Chen Z. Discrete wavelet transform-based time series analysis and mining. J ACM Comput Surveys (CSUR) Surveys Homepage Arch 2011;43(2):6:1–6:37.

[65] Adonias M. S. Ferreira, Cristiano H. O. Fontes, Carlos A. M. T. Cavalcante, Jorge E. S. Marambio. A new proposal of typing load profiles to support the decision making in the sector of electricity energy distribution. International conference on industrial engineering and industrial management (ICIEOM); 2012, p. 18.1–18.7.

[66] Adonias M. S. Ferreira, Cristiano H. O. Fontes, Carlos A. M. T. Cavalcante, Jorge E. S. Marambio. A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector. Int J Electric Power Energy Syst 2013;53(C):824–31.

[67] Seppãlã Anssi. Statistical distribution of customer load profile, vol. 95. IEEE; 1995. p. 696–701.

[68] Janes Joseph. Categorical relationships: chi-square. Library Hi Tech 2001;19(3):296–8.

[69] O'Gorman T. W. A comparison of an adaptive two-sample test to the t-test, rank-sum, and log-rank tests. Commun Stat – Simulat Computat 1997;26:1393–411.

[70] Khalil B., Ouarda T. B. M. J., St-Hilaire. Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. J Hydrol 2011;405:277–87.

[71] Zimba Paul V., Mischke Charles C, Brashear Suzanne S. Pond age–water column trophic relationships in channel catfish Ictalurus punctatus production ponds. Aquaculture 2003;219:291–301.

[72] Deng Xiaogang, Tian Xuemin, Chen Sheng. Modified kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis. Chemometric Intell Lab Syst 2013;127: 195–209.

[73] Chiu S. A cluster estimation method with extension to fuzzy model identification. Proceedings of the third IEEE conference on fuzzy systems. vol. 2, Orlando –Florida (USA); 1994, p. 1240–5.

[74] Zhang J., Dobson I., Alvarado F. L. Quantifying transmission reliability margin. Int J Electr Power Energy System 2004;26(9):697–702.

[75] Kiyoung Yang, Cyrus Shahabi. A PCA-based similarity measure for multivariate time MMDB'04; 2004.

[76] Dobos La´ szlo´, Ja´ nos Abonyi. On-line detection of homogeneous operation ranges by dynamic principal component analysis based time-series segment. Chem Eng Sci 2012;75:96–105.

[77] Bécavin Christophe, Tchitchek Nicolas, Mintsa-Eya Colette, Annick Lesne, Arndt Benecke. Improving the efficiency of multidimensional scaling in the analysis of high dimensional data using singular value decomposition. Bioinformatics 2011;27(10):1413–21.

[78] Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.

[79] Halkidi M., Batistakis Y., Vazirgiannis M. On clustering validation techniques. J Intell Inform Syst 2001;17:107–45.

[80] Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J. M., Perona I. An extensive comparative study of cluster validity índices. Pattern Recogn 2013;46:243–56.

[81] International Performance Measurement & Verification Protocol – IPMV; 2007. [82] Motomasa DAIGO. Factor analysis and pattern decomposition method. SPIE 2005;6043:1–8.

[83] Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal 2008;52:5186–201.

[84] Bensaid A., Hall L. O., Bezdek James C., Clarke L. P. Partially supervised clustering for image segmentation. Patt Recog 1996;29(5):859–87.

[85] Vine E. An international survey of the energy service company (ESCO) industry. Energy Policy 2005;33:691–704.

[86] Abdel-Aal R. E. Modeling and forecasting electric daily peak loads using abductive networks. Electric Power Energy Syst 2006;28:133–41.

[87] Saini L. M., Soni M. K. Artificial neural network-based peak load forecasting using conjugate gradient methods. IEEE Trans Power Syst 2002;17:907–12.

[88] Aggarwal Sanjeev Kumar, Saini Lalit Mohan, Kumar Ashwani. Electricity price forecasting in deregulated markets: a review and evaluation. Electric Power Energy System 2009;31:13–22.

[89] Guide for the Preparation of Energy Efficiency Program – MPEE – Version; 2008.

[90] Box G. E. P., Jenkins G. M., Reinsel G. C. Time series analysis, forecasting and control. 4th ed. Hoboken (NJ): Wiley; 2008.

[91] Sadaei H. J., Enayatifar R., Abdullah A. H., Gani A. Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search. Int J Electr Power Energy System 2014;62:118–29.

[92] Kristiansen T. A time series spot price forecast model for the Nord Pool market. Int J Electr Power Energy System 2014;61:20–6.

[93] Xiong T., Bao Y. Interval forecasting of electricity demand: a novel bivariate EMD-based support vector regression modeling framework. International Journal Electrical Power Energy System 2014;63:353–62.