JAIR ROCHA DO PRADO

MÉTODOS ESTATÍSTICOS NO MONITORAMENTO DA POTÊNCIA ELÉTRICA

LAVRAS - MG

2015

JAIR ROCHA DO PRADO

MÉTODOS ESTATÍSTICOS NO MONITORAMENTO DA POTÊNCIAELÉTRICA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientadora

Dra. Thelma Sáfadi

Coorientador

Dr. Joaquim Paulo da Silva

LAVRAS - MG
2015

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Prado, Jair Rocha do.
Métodos estatísticos no monitoramento da potência elétrica / Jair Rocha do Prado. – Lavras : UFLA, 2015.
159 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2015.
Orientador: Thelma Sáfadi.
Bibliografia.

1. ICA. 2. Análise de agrupamento. 3. Índices de Moran e Geary. Potência ativa. D. Energia elétrica. I. Universidade Federal de Lavras. II. Título.

CDD -

JAIR ROCHA DO PRADO

MÉTODOS ESTATÍSTICOS NO MONITORAMENTO DA POTÊNCIA ELÉTRICA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 11 de dezembro de 2014.

Dr. Ednaldo Carvalho Guimarães	UFU
Dr. João Domingos Scalon	UFLA
Dr. Joaquim Paulo da Silva	UFLA
Dr. Renato Ribeiro de Lima	UFLA

Dra. Thelma Sáfadi
Orientadora

LAVRAS - MG

2015

AGRADECIMENTOS

A Deus, por dar-me força em todos os momentos da minha vida.

Aos meus pais, que são exemplos de honestidade. Serei sempre grato a eles pela educação a mim proporcionada.

A minha esposa, que foi parceira em todos os momentos, se mostrando uma base forte na construção desse sonho.

A toda a família da minha esposa, em especial ao meu sogro José Macedo e minha sogra Silvânia, pelo constante apoio nessa caminhada.

Ao meu grande amigo, Padre Otair Cardoso Cruz, que contribuiu para o desenvolvimento da minha inteligência emocional.

À Universidade Federal de Lavras (UFLA), que me proporcionou cursar o doutorado em Estatística e Experimentação Agropecuária.

À CNPQ, pela concessão de bolsa de estudos.

À professora Thelma Sáfadi, pessoa e profissional que admiro muito. De todo o tempo de convivência ficou um grande respeito e uma vasta gratidão.

Ao professor Joaquim, por ser um coorientador dedicado e amigo.

A todos os meus colegas de curso, em especial ao Diogo, Felipe, Danilo, Manoel e Marcelo, que admiro muito e considero grandes amigos.

A todos os professores do Departamento de Ciências Exatas (DEX), que são exemplos para a minha carreira acadêmica.

Aos membros da banca examinadora, pelas contribuições a este trabalho.

A todos os funcionários do DEX, que são pessoas amigas e prestativas.

À Universidade Estadual de Londrina, em especial ao Departamento de Estatística, por me acolher e dar a oportunidade de desenvolver a minha profissão.

A todas as pessoas que, de alguma forma, participaram na realização deste sonho.

RESUMO

O monitoramento de energia elétrica é uma tarefa muito importante nos diversos setores da sociedade, sejam empresas, universidades, indústrias, etc. Geralmente as análises feitas em variáveis energéticas subutilizam ou até mesmo negligenciam métodos estatísticos para solução de problemas. Neste trabalho o objetivo é mostrar que a estatística pode ser mais utilizada em pesquisas na área de eletricidade, e que esta pode contribuir de forma eficaz nesse ramo. Para esse estudo serão utilizados dados de energia elétrica da Universidade Federal de Lavras (UFLA). A partir de uma análise inicial dos dados, teve-se a ideia de usar uma nova abordagem estatística, que considera um conjunto de métodos. As metodologias estatísticas utilizadas foram: análise de componentes independentes (ICA), análise de agrupamento (AA) e índices de Moran e Geary. No contexto de inovação dos métodos estatísticos, o objetivo é considerar os índices de Moran e Geary para escolher o ponto de corte em dendrogramas, ou seja, encontrar o agrupamento de melhor qualidade. A ideia foi basicamente fazer comparações: primeiro, comparar alguns departamentos/setores e os dias da semana da UFLA, considerando a variável potência ativa, segundo, confrontar os meses do ano, a partir das variáveis demanda de potência faturada em horários de ponta e fora de ponta. Realizaram-se comparações considerando tanto os dados originais quanto algumas partições. Foram realizadas também comparações considerando os componentes independentes (CI's), obtidos das séries de potência, tanto para os departamentos/setores, quanto para os dias da semana. Os agrupamentos obtidos para o departamentos/setores amostrados em 2010 mostraram que geralmente o Centro de Informática (CIN) e o Restaurante Universitário (RU) são dissimilares aos demais. Para os dias da semana, os agrupamentos encontrados levaram a conclusões esperadas, com o sábado e domingo ficando em grupos separados dos demais dias. Para o horário de ponta dos anos de 2010 a 2013 o agrupamento de melhor qualidade encontrado para os meses, pelo índice de Moran, foi jan; nov, dez; fev, jul e mar, abr, mai, jun, ago, set, out. No horário fora de ponta, para o mesmo período, os grupos de meses de melhor qualidade, encontrados pelo índice de Geary, foram: set; abr; jul; out; dez; mai, ago; jan, jun e fev, mar, nov. Por meio dos CI's observou-se, por exemplo, que a segunda e a sexta-feiras são atípicos considerando os dias úteis. Pode-se concluir que os métodos ICA, AA e índices de Moran e Geary, podem em conjunto ser uma maneira eficiente de monitorar energia elétrica em uma universidade ou numa empresa qualquer.

Palavras-chave: ICA. Análise de agrupamento. Índices de Moran e Geary. Potência ativa. Energia elétrica.

ABSTRACT

The monitoring of electric energy is a very important task in many sectors of society, be them companies, universities, industries, etc. Generally the analyses performed on energy variables underuse or even neglect statistical methods for solving issues. In this work, the objective is to show that statistics can be more used in researches on the area of electricity, and that this may effectively contribute in this field. For this study, we will use electric energy data from the Universidade Federal de Lavras (UFLA). With an initial data analysis, we had the idea of using a new statistical approach, considering a set of methods. The statistical methodologies used were: independent component analysis (ICA), grouping analysis (GA) and Moran and Geary indexes. In the context of innovation of statistical methods, the objective is considering the Moran and Geary indexes, which are measured originating in spatial statistics, to choose the cutting point in dendrograms, that is, find the best quality grouping. The idea was basically to compare: first, compare a few departments/sectors and the days of the week of UFLA, considering the active potency variable, second, confront the months of the year, from the variables of demanded invoiced potency in peak hours and outside peak hours. Comparisons were performed considering both the original data and some partitions. We also performed comparisons considering the independent components (ICs), obtained from the potency series, for both department/sectors and days of the week. The groupings obtained for the departments/sectors sampled in 2010 showed that, generally, the Informatics Center (INC) and the University Restaurant (UR) are dissimilar to the others. For the days of the week, the groupings found led us to expected conclusions, with Saturday and Sunday remaining in separate groups from the remaining days. For peak hours the years 2010-2013 the best quality grouping found for the months, by the Moran index, was (Jan), (Nov, Dec), (Feb, Jul) and (Mar, Apr, May, Jun, Aug, Sep, Oct). In the outside the peak, for the same period, the groups of best quality months, found by the Geary index, were: (Set) (Apr), (Jul), (Oct) (Dec), (May, Aug), (Jan, Jun) and (Feb, Mar, Nov). By means of the ICs, we observed that, for example, Monday and Friday are atypical considering busyness days. We can conclude that the ICA, GA and Moran and Geary indexes might, together, be an efficient way of monitoring electric energy in a university or in any given enterprise.

Keywords: ICA. Grouping analysis. Moran and Geary Indexes. Active Potency. Electric Energy.

SUMÁRIO

1 INTRODUÇÃO

A evolução da humanidade sempre esteve associada à energia em suas diversas formas. A dependência é tão significativa que o consumo de energia pode ser usado para medir o poder aquisitivo da sociedade. Assim como na sociedade, a energia elétrica tem papel relevante também nas universidades e, devido a esse fato, as universidades precisam também monitorar e controlar melhor alguns parâmetros relacionados a essa fonte de energia.

O gerenciamento contínuo de variáveis energéticas pode levar a decisões mais precisas, contribuindo tanto na economia de energia quanto no seu uso eficiente, sem desperdícios. Além disso, em certas ocasiões o acompanhamento da base de dados de energia elétrica pode proporcionar contratos mais vantajosos com as empresas de distribuição, pois multas que poderiam ocorrer por excesso de consumo podem ser evitados por meio de um planejamento.

Observa-se que estudos relacionados ao monitoramento da energia elétrica muitas vezes subutilizam a estatística, ou até mesmo negligenciam o seu uso. O mesmo acontece com análises em empresas, universidades e nos diversos setores da sociedade. Assim, nesta tese um dos objetivos é preencher essa lacuna e mostrar que a estatística pode ser melhor utilizada e contribuir de forma significativa no processo de investigação de variáveis energéticas. Para mostrar essa contribuição, foi considerado como cenário de estudo a Universidade Federal de Lavras (UFLA).

Na UFLA, o monitoramento e gerenciamento da energia elétrica sempre foi realizado pelos profissionais (engenheiros e administradores) envolvidos no processo, porém sem uma contribuição efetiva de profissionais da área de estatística, que por meio do seu conhecimento podem levar a resultados satisfatórios, considerando modelos probabilísticos.

Somente a partir do ano de 2011, surgiram trabalhos utilizando métodos

estatísticos para análise de variáveis energéticas na universidade. Duas dissertações foram feitas: uma no programa de pós-graduação em engenharia agrícola, e outra no programa de estatística e experimentação agropecuária. Os trabalhos citados são de Jesus (2011) e Prado (2011), respectivamente. Ambos dão ênfase a importância de se estudar variáveis energéticas na universidade. Prado (2011) afirma que análises estatísticas de dados energéticos nunca haviam sido feitas na universidade.

Os métodos estatísticos envolvidos no trabalho de Jesus (2011) foram somente de Estatística Descritiva, por meio de gráficos, medidas de posição e dispersão. Já Prado (2011) apresentou modelos de séries temporais para as variáveis demanda de potência registrada e consumo de energia elétrica, ambas em horário de ponta e fora de ponta.

Para se ter uma ideia da importância de estudos estatísticos nesta área, fazendo um paralelo entre os períodos antes e depois dos estudos feitos por Prado (2011) e Jesus (2011), tem-se que até 2010 em cerca de 48% dos meses houve superação da demanda registrada em relação à contratada, enquanto que de 2011 a 2013 este índice caiu para um valor próximo de 8%. Considerando o período fora de ponta as porcentagens caíram de 59% para 22% aproximadamente. Esses resultados não foram consequências apenas dos dois trabalhos expostos. A melhoria apresentada pôde ser obtida em virtude das análises estatísticas e de medidas tomadas por profissionais da área em estudo.

Esses valores indicam que realmente a estatística pode contribuir com as decisões no momento da contratação de demandas de potência da universidade. Observa-se também que, apesar dos índices indicarem melhorias, ainda pode-se avançar mais. Portanto, a proposta nesta tese é utilizar uma combinação de métodos estatísticos presentes na literatura que não foram anteriormente explorados na

área energética.

A UFLA nos últimos anos vem passando por mudanças importantes, tanto em sua estrutura física, quanto no número de funcionários e alunos. Alguns cursos novos fazem parte dessas mudanças, são eles: engenharia, direito, medicina, sendo este último com previsão para início do funcionamento no primeiro semestre de 2015. Tudo isso torna ainda mais importante a realização de estudos sobre o comportamento de variáveis associadas à energia elétrica no campus.

Para se avaliar o comportamento da energia elétrica algumas variáveis relacionadas à tal energia são importantes, dentre elas tem-se: potência ativa, intensidade da corrente elétrica, potência reativa indutiva, entre outras.

Assim, usando as técnicas de Análise de Componentes Independentes (ICA), Análise de Agrupamento (AA) e índices de Moran e Geary, deseja-se neste trabalho avaliar na Universidade Federal de Lavras algumas situações que envolvam a variável potência ativa.

Uma primeira análise desta avaliação baseia-se na classificação de alguns departamentos/setores da universidade, dos dias da semana e dos meses do ano, de acordo com as similaridades obtidas pela AA, considerando a variável de interesse. Para os meses, fez-se a classificação considerando a demanda de potência em horário de ponta e fora de ponta.

Para se obter informações em períodos importantes para a universidade, fez-se as análises em subpartes das séries originais, para cada uma delas realizou-se o mesmo procedimento, ou seja, aplicou-se AA para encontrar os agrupamentos, visualizados por meio de um dendrograma.

Após a construção dos dendrogramas, há a necessidade de se encontrar o ponto de corte neste gráfico, ou seja, obter o número de grupos que gera o melhor agrupamento. Na literatura, existem alguns critérios, porém, segundo Mingoti

(2005), nenhum deles emite a resposta de forma exata, ou seja, os métodos existentes somente ajudam na decisão final. Neste ponto, entra a inovação desta tese no contexto estatístico. A proposta aqui foi considerar para escolha do número de grupos os índices de Moran e Geary, que são medidas com origem na estatística espacial. Para o cálculo desses índices, foi proposta uma nova maneira de se construir a matriz de vizinhanças, em que o elemento para ser vizinho não necessariamente deve estar próximo fisicamente.

A combinação dos índices de Moran e Geary com a AA tem a intenção de mostrar uma nova maneira de se determinar o número "ótimo" de grupos, diferente do que é feito normalmente pelos métodos presentes na literatura, os quais indicam somente o melhor agrupamento. A ideia aqui é obter além do agrupamento "ótimo", outros agrupamentos pertinentes, considerando os índices de Moran e Geary significativos, proporcionando assim ao pesquisador a possibilidade de escolha, considerando a natureza de cada pesquisa.

Uma segunda análise consiste em aplicar ICA aos dados de potência dos departamentos/setores e dias da semana, com o intuito de encontrar particularidades e, em seguida, obter agrupamentos considerando os coeficientes da combinação linear dos componentes independentes (particularidades) obtidos a partir da ICA. O objetivo dessa classificação é obter grupos de departamentos e dias que sejam homogêneos dentro de si e heterogêneos entre si de acordo com cada uma das particularidades encontradas. Nessa análise, utilizou-se também os índices de Moran e Geary para encontrar o ponto de corte nos dendrogramas, e assim obter o número de grupos.

Considerando o contexto prático deste estudo, tem-se que os vários departamentos podem potencialmente ser usados para avaliações energéticas no campus, mas por meio dos agrupamentos obtidos torna-se viável usar alguns deles,

seja para monitoramento ou outra finalidade, como por exemplo, planejamento de futuros departamentos/setores que porventura venham a ser construídos. Esse raciocínio vale também para os dias da semana, a partir do agrupamento obtido destes, ações previstas podem ser tomadas com maior segurança.

Considerando a classificação dos meses do ano, medidas podem ser tomadas no sentido de que futuras contratações de demandas de potência, tanto no horário de ponta quanto no horário fora de ponta, possam ser melhor quantificadas. Isso será possível, pois considerando os grupos formados pode-se ver junto à concessionária de energia elétrica um contrato com diferentes demandas de acordo com os grupos formados.

Com relação às particularidades encontradas pela ICA e os agrupamentos oriundos desta, espera-se obter informações mais refinadas que não tenham sido exploradas com as outras técnicas aqui apresentadas, reforçando ainda mais a ideia de eficiência da estatística na avaliação de variáveis energéticas.

Assim, com esta tese, a intenção é mostrar que a estatística pode contribuir de forma eficiente na resolução de problemas associados ao consumo de energia elétrica, seja numa universidade ou uma empresa qualquer, pública ou privada.

No contexto do monitoramento de energia elétrica o conjunto de metodologias aqui propostas, deverá responder perguntas como:

a) Quais grupos são formados considerando as variáveis de interesse, os períodos em estudo, os departamentos, os dias da semana e os meses?

b) Considerando os agrupamentos obtidos no item anterior, quais são significativos e qual é o melhor para cada análise a partir dos índices de Moran e Geary?

c) Quais são os componentes independentes obtidos para os departamen-

tos/setores e os dias da semana?

d) Quais interpretações práticas podem ser retiradas de cada componente independente (CI)?

e) A partir dos CI's, quais grupos podem ser formados? Segundo os índices de Moran e Geary, quais desses grupos são significativos e qual é o melhor para os departamentos, os dias da semana e os meses?

f) Quais as informações relevantes encontradas nos periodogramas para cada componente independente?

Em resumo, espera-se que com o conjunto de métodos estatísticos propostos nesta tese o monitoramento da energia elétrica seja realizado de forma mais eficiente na UFLA, ocasionando economia nas contas de energia em decorrência da redução do número de meses em que ocorrem multas.

2 REVISÃO BIBLIOGRÁFICA

2.1 Trabalhos associados à energia elétrica

Alguns trabalhos científicos já publicados servem de suporte e motivação para o estudo proposto nesta tese. Serão expostos a seguir alguns trabalhos que usam ou não métodos estatísticos para monitoramento de energia elétrica. O intuito é observar que nos trabalhos relacionados à energia elétrica a combinação de métodos propostos nesta tese não foi utilizado anteriormente por nenhum outro autor, tornando assim este trabalho uma grande contribuição na área. A seguir são apresentados alguns desses trabalhos.

Em seu estudo, Ferreira (2006) propõe modelos para séries temporais de demanda de energia elétrica baseados em métodos estatísticos, redes neurais artificiais, algoritmos de identificação de tendências, ciclos e sazonalidades e análise de componentes independentes. Foram utilizadas as séries de energia das diferentes regiões do Brasil, além da série do Sistema Interligado Nacional (SIN). Nesse trabalho, foi comprovado que a vulnerabilidade presente nos métodos e processos pode ser minimizada com a combinação de métodos de previsão, com o intuito de reduzir a incerteza inerente aos métodos e elevar a qualidade das previsões realizadas.

Martins (2008) apresenta uma metodologia para monitoramento do consumo de energia elétrica. Foram abordados dois estudos de caso: uma universidade e uma indústria de fabricação de rações animais. O primeiro baseou-se nas seguintes metodologias: análise de crescimento demográfico do campus, análise das contas de energia elétrica, análise de estruturas tarifárias do campus e análise da curva de demanda de um dia típico. O segundo estudo de caso foi baseado nas informações fornecidas pelo consumidor e em medições e observações dos parâ-

metros elétricos da instalação. O autor afirma que a metodologia proposta é de fácil implementação e facilita a tomada de decisões.

Outro estudo encontrado na literatura foi de Ferreira (2010). Neste trabalho foram utilizadas técnicas de processamento de sinais e inteligência computacional com o intuito de analisar, detectar e classificar distúrbios elétricos. A análise de componentes independentes foi aplicada aos múltiplos distúrbios, com o objetivo de encontrar os distúrbios isolados, para fins de classificação e análises. Por meio de testes, o autor mostra que o método é promissor e passível de aplicação em tempo real.

Ohtsuka, Oga e Kakamu (2010) estudaram a demanda de energia elétrica no Japão em diferentes regiões. Esses autores propõem um modelo que considera somente a análise temporal e outro que leva em consideração tanto a dependência espacial quanto a dependência temporal. Os resultados obtidos levaram a conclusão de que o modelo espaço-temporal possui melhor desempenho quando o objetivo é fazer previsões.

Jesus (2011) evidencia em seu trabalho a importância de se estudar variáveis energéticas no ambiente universitário. Segundo esse autor, devido principalmente à falta de planejamento a longo prazo, as universidades experimentam situações desfavoráveis, sendo muitas vezes penalizadas por ultrapassagem da demanda contratada. Em seu trabalho é apresentado alternativas de monitoramento de instalações elétricas na UFLA. Neste estudo, as estatísticas envolvidas foram somente descritivas, por meio de gráficos, medidas de posição e dispersão.

Prado (2011) apresenta alguns modelos de séries temporais para as variáveis demanda de potência registrada e consumo de energia elétrica, ambas em horário de ponta e fora de ponta. Os dados foram coletados na Universidade Federal de Lavras. Nesse trabalho, o autor afirma que os modelos propostos ajustaram-se

bem aos dados de energia elétrica, ou seja, podem contribuir de forma bastante significativa com as decisões tomadas nesta instituição de ensino.

Villamagna (2013) analisou séries de consumo e demanda de energia elétrica na Universidade Federal de Lavras (UFLA). A autora considera séries mensais, compreendendo o período de janeiro de 1995 a dezembro de 2011. Foram considerados dois métodos para a modelagem das séries temporais: Box e Jenkins e Redes Neurais Artificiais. O objetivo foi fazer previsões. Os resultados mostraram que as Redes Neurais Artificiais obtiveram melhor desempenho em todas as situações, considerando como critério de comparação o erro quadrático médio de previsão e o erro percentual absoluto médio de previsão.

2.2 Análise de componentes independentes

A análise de componentes independentes (ICA) é um método para encontrar os fatores ou componentes independentes de um conjunto de dados multivariados. O que distingue o ICA dos outros métodos, como por exemplo análise de componentes independentes, é que essa análise considera os componentes como sendo estatisticamente independentes mesmo sendo não gaussianos.

Neste texto, por questões de simplificação de cálculos, todas as variáveis aleatórias serão assumidas ter média zero, salvo indicação contrária. A abordagem de ICA que será apresentada aqui pode ser encontrada com detalhes em Hyvärinen, Karhunen e Oja (2001).

2.2.1 Definição

Pode-se definir ICA como um modelo estatístico de variáveis latentes. Sejam p variáveis aleatórias $X_1, ..., X_p$, que são combinações lineares de p variáveis

$S_1, ..., S_p$. Tem-se daí que:

$$X_i = a_{i1}S_1 + a_{i2}S_2 + ... + a_{ip}S_p, \forall i = 1, ..., p, \tag{2.1}$$

em que $a_{ij}, i, j = 1, ..., p$ são coeficientes reais e, pela definição, S_i são mutuamente independentes.

Na equação 2.1, tem-se o modelo básico da ICA. Observe que este descreve como as variáveis $X_1, ..., X_p$ são geradas por um processo de mistura dos componentes S_i. Os componentes independentes S_i não podem ser diretamente observados, por isso são chamados de variáveis latentes. Os coeficientes das combinações lineares dos S_i, a_{ij}, são desconhecidos. Assim tudo o que se pode observar são as variáveis aleatórias X_i, e tanto os coeficientes a_{ij} quanto os componentes independentes S_i devem ser estimados.

Usando a notação matricial, o modelo pode ser escrito como

$$\mathbf{X} = \mathbf{AS}, \tag{2.2}$$

em que $\mathbf{X}$ é o vetor aleatório das variáveis observáveis $X_1, ..., X_n$, $\mathbf{S}$ é o vetor aleatório dos componentes independentes $S_1, ..., S_n$ e $\mathbf{A}$ é a matriz cujos elementos são os coeficientes a_{ij}, chamada matriz de mistura.

Pode-se definir ICA também como um problema de determinar uma transformação linear dada pela matriz $\mathbf{B}$, tal que

$$\mathbf{S} = \mathbf{BX}, \tag{2.3}$$

em que $\mathbf{B}$ é a inversa da matriz $\mathbf{A}$.

Às vezes tem-se o interesse nas colunas da matriz $\mathbf{A}$. Se estas forem definidas por $\mathbf{a}_j$, o modelo de ICA pode ser escrito como

$$\mathbf{X} = \sum_{i=1}^{n} \mathbf{a}_i S_i. \tag{2.4}$$

Segundo Hyvärinen, Karhunen e Oja (2001), esta definição de modelo é uma das mais básicas, pois omite a existência de ruído e considera o número de componentes igual ao número de sinais observados.

2.2.2 Independência estatística

Um conceito fundamental, que é usado pela análise de componentes independentes, é a independência estatística. Para se entender este conceito considere a princípio duas variáveis aleatórias Y_i e Y_j distintas. A variável Y_i é independente de Y_j se o fato de conhecer o valor de Y_j não traz qualquer informação extra sobre o valor de Y_i. De acordo com Hyvärinen, Karhunen e Oja (2001), o valor de um dado jogado e de uma moeda lançada ou o sinal de fala e um ruído de fundo decorrente de um sistema de ventilação são exemplos de variáveis aleatórias independentes. A seguir serão apresentados conceitos considerando Y_i e Y_j como variáveis contínuas. De maneira análoga os conceitos podem ser obtidos para variáveis discretas.

A independência pode ser definida estatisticamente considerando densidades de probabilidade. As variáveis Y_i e Y_j são ditas independentes, se e somente se,

$$f_{Y_i,Y_j}(y_i, y_j) = f_{Y_i}(y_i)\, f_{Y_j}(y_j),$$

em que f_{Y_i,Y_j} denota a função densidade probabilidade conjunta das variáveis Y_i e Y_j e f_{Y_i} e f_{Y_j} são as densidades marginais de Y_i e Y_j, respectivamente.

Utilizando do conceito de esperança, pode ser demonstrado que a inde-

pendência de variáveis aleatórias satisfaz a seguinte propriedade

$$E\left[h\left(Y_i\right)g\left(Y_j\right)\right] = E\left[h\left(Y_i\right)\right]E\left[g\left(Y_j\right)\right], \tag{2.5}$$

em que $f(Y_i)$ e $g(Y_j)$ são quaisquer funções integráveis de Y_i e Y_j, respectivamente.

O conceito de independência apresentado anteriormente pode ser generalizado para mais de duas variáveis.

Outro conceito importante é o de variáveis não correlacionadas. Duas variáveis Y_i e Y_j distintas são não correlacionadas quando a covariância entre elas é zero

$$Cov\left(Y_i, Y_j\right) = E\left(Y_iY_j\right) - E\left(Y_i\right)E\left(Y_j\right) = 0. \tag{2.6}$$

A partir da equação 2.6, tem-se que

$$E\left(Y_iY_j\right) = E\left(Y_i\right)E\left(Y_j\right). \tag{2.7}$$

Para variáveis de média nula, tem-se que

$$E\left(Y_iY_j\right) = 0,$$

o que torna as variáveis Y_i e Y_j, além de não correlacionadas, ortogonais.

As equações 2.5 e 2.7 revelam que a independência estatística é uma propriedade muito mais forte do que a não correlação. Observe que a equação 2.7, que define a não correlação, é um caso particular da equação 2.5, que caracteriza a independência estatística, quando $h(Y_i)$ e $g(Y_j)$ são funções lineares.

2.2.3 Variáveis não gaussianas

Tem-se que variáveis conjuntamente gaussianas não correlacionadas são necessariamente independentes. Portanto, se as variáveis observadas forem gaussianas é simples encontrar componentes que sejam independentes. Basta que se aplique uma técnica que torne os dados não correlacionados, como análise de componentes principais (PCA). Porém, na realidade, os dados muitas vezes não seguem uma distribuição gaussiana. Assim, procedimentos mais refinados devem ser utilizados, como por exemplo ICA, que utiliza estatísticas de ordem superior nas suas estimativas.

No entanto, uma restrição ao se utilizar a análise de componentes independentes diz respeito exatamente às variáveis gaussianas, pois estatísticas de ordem superior são nulas para estas variáveis, tornando, assim, inútil o efeito da ICA.

2.2.4 Restrições do modelo de ICA

A partir dos conceitos e resultados apresentados anteriormente, para se certificar de que o modelo básico ICA possa ser aplicado, algumas suposições e restrições devem ser feitas. São elas:

a) os componentes são assumidos estatisticamente independentes;

b) os componentes independentes devem possuir distribuições não gaussianas;

c) por simplicidade, deve-se assumir que a matriz de mistura seja quadrada, em outras palavras, o número de componentes independentes é igual ao número de variáveis observadas.

A última restrição apresentada pode ser relaxada. Mais detalhes podem ser encontrados em Hyvärinen, Karhunen e Oja (2001).

2.2.5 Ambiguidades do modelo de ICA

Por meio do modelo ICA da equação 2.2, pode-se mostrar a existência das seguintes ambiguidades ou indeterminações.

a) Não é possível determinar as variâncias (energias) dos componentes independentes.

A razão para esta restrição está no fato de que tanto **S** quanto **A** são desconhecidas. Assim qualquer multiplicação por escalar que se faça em uma das fontes S_i pode ser cancelada pela divisão da correspondente coluna $\mathbf{a}_i$ de **A** pelo mesmo escalar. Se α_i é este escalar, então

$$\mathbf{X} = \sum_{i=1}^{n} \left(\frac{1}{\alpha_i} \mathbf{a}_i \right) (S_i \alpha_i) .$$

O que pode ser feito para fixar as magnitudes dos componentes independentes é assumir que cada um tem variância unitária. Porém, a ambiguidade do sinal ainda continua, ou seja, pode-se multiplicar um componente por -1 sem afetar o modelo, mas isso não causa grandes problemas para a maioria das aplicações. No método de solução para ICA apresentado neste texto será considerada a variância unitária dos componentes.

b) Não é possível determinar a ordem dos componentes independentes.

Aqui também a justificativa está no fato de que, tanto **S** quanto **A** são desconhecidos. Pode-se livremente alterar a ordem dos termos da soma na equação

2.4 e associar, por exemplo, qualquer um dos componentes ao primeiro. Formalizações podem ser encontradas em Hyvärinen e Oja (2000).

2.2.6 Procedimentos de pré-processamento para ICA

Com o intuito de simplificar a teoria de estimação dos componentes independentes, os dados observados devem passar por algumas mudanças antes da aplicação de um algoritmo. Os principais procedimentos realizados são centralização e branqueamento.

2.2.6.1 Centralização

Uma suposição que simplifica consideravelmente os cálculos da ICA é a média nula das variáveis. Porém, quando isto não se verifica, faz-se a centralização das variáveis originais, ou seja, faz-se uma subtração da média nos dados. Assim, se $\widetilde{\mathbf{X}}$ é um vetor de variáveis não nulas, então o vetor submetido ao algoritmo ICA é dado por

$$\mathbf{X} = \widetilde{\mathbf{X}} - E\left(\widetilde{\mathbf{X}}\right).$$

Os componentes independentes também tem média nula, já que

$$E\left(\mathbf{S}\right) = \mathbf{B}E\left(\mathbf{X}\right).$$

Já a matriz de mistura permanece a mesma diante deste pré-processamento. Portanto, a sua estimação não é afetada após a centralização das variáveis observadas. Após encontrar as estimativas da matriz de mistura e dos componentes independentes para os dados de média zero, pode-se obter os componentes das

variáveis originais simplesmente adicionando $\mathbf{BE}\left(\widetilde{\mathbf{X}}\right)$ aos componentes independentes de média zero.

2.2.6.2 Branqueamento

Uma propriedade considerada mais forte que a não correlação é o branqueamento. Essa é uma técnica muito útil, geralmente implementada como pré-processamento para a ICA. O branqueamento tem como objetivo fazer variáveis aleatórias serem não correlacionadas e com variância unitária.

O processo de branqueamento baseia-se na decomposição ortogonal do vetor $\mathbf{X}$, que pode ser escrito como

$$\mathbf{Z} = \mathbf{VX}, \tag{2.8}$$

em que $\mathbf{Z}$ é o vetor de variáveis branqueadas e $\mathbf{V} = \mathbf{ED}^{-1/2}\mathbf{E}^T$ é a matriz de branqueamento.

Tem-se que $\mathbf{E}$ é a matriz de autovetores de $E\left(\mathbf{XX}^T\right)$ e $\mathbf{D}$ é a matriz diagonal de seus autovalores.

O branqueamento transforma a matriz de mistura em uma nova matriz, $\widetilde{\mathbf{A}}$. Segue das equações 2.2 e 2.8 que

$$\mathbf{Z} = \mathbf{VAS} = \widetilde{\mathbf{A}}\mathbf{S}.$$

A importância do branqueamento está no fato de que a nova matriz $\widetilde{\mathbf{A}}$ é ortogonal, pois

$$E\left(\mathbf{ZZ}^T\right) = \widetilde{\mathbf{A}}E\left(\mathbf{SS}^T\right)\widetilde{\mathbf{A}}^T = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T = \mathbf{I}.$$

Desta forma o número de parâmetros a serem estimados passa de n^2 para $n(n-1)/2$. De acordo com Hyvärinen e Oja (2000), pode-se dizer que o branqueamento das variáveis observadas resolve a metade dos problemas de ICA.

A técnica de branqueamento pode ser útil também na redução de dimensão dos dados. Para isso, basta que se descarte os componentes cujos autovalores da matriz **E** sejam considerados pequenos, prática também muitas vezes realizada na análise de componentes principais.

Neste trabalho, em muitas situações os dados serão considerados pré-processados por branqueamento.

2.2.7 Ilustração de ICA

O objetivo aqui é exemplificar o modelo ICA em termos estatísticos. Para isso, considere dois componentes independentes com a seguinte distribuição uniforme:

$$f(S_i) = \begin{cases} \frac{1}{2\sqrt{3}}, & se\ |S_i| \leq \sqrt{3} \\ 0, & para\, outros\, valores \end{cases} \tag{2.9}$$

Na Figura 1, tem-se a densidade conjunta de S_1 e S_2. Observe que esta também é uma distribuição uniforme, pois a densidade conjunta de duas variáveis independentes é apenas o produto de suas densidades marginais. Os pontos que estão na Figura 1 foram obtidos por simulação, a partir dessa distribuição.

Como parte dessa ilustração considere que os componentes independentes foram misturados segundo a seguinte matriz de mistura:

$$\mathbf{A}_0 = \begin{pmatrix} 5 & 10 \\ 10 & 2 \end{pmatrix}.$$

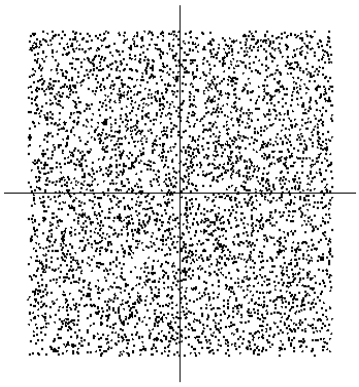

Figura 1 Distribuição conjunta dos componentes independentes S_1 e S_2, cujas distribuições são uniformes. Eixo horizontal: S_1 e eixo vertical: S_2
Fonte: Hyvärinen e Oja (2000)

A partir da mistura realizada tem-se agora duas variáveis X_1 e X_2, cuja distribuição conjunta é uniforme sobre um paralelograma, como pode ser mostrado na Figura 2. Observa-se que existe uma dependência entre estas duas variáveis e isto é devido ao fato de que é possível predizer uma delas, por exemplo X_2, a partir da outra, X_1, o que não ocorria com S_1 e S_2.

A ICA tem o objetivo de encontrar os componentes S_1 e S_2 usando apenas as informações contidas nas misturas X_1 e X_2. Como já visto, uma técnica geralmente utilizada antes de se aplicar os algoritmos da ICA é o branqueamento, uma ilustração gráfica deste pré-processamento pode ser visto na Figura 3, na qual têm-se os dados da Figura 2 branqueados. Estas novas variáveis ainda não são os componentes independentes, o passo seguinte é estimar uma transformação ortogonal, finalizando assim a análise, ou seja, as estimativas dos componentes independentes S_1 e S_2 são enfim encontradas.

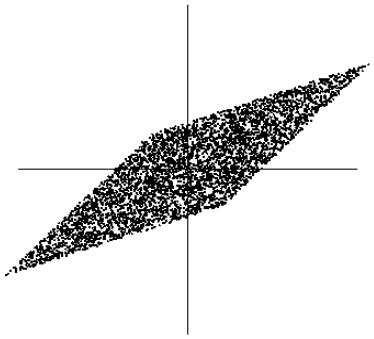

Figura 2 Distribuição conjunta das misturas X_1 e X_2. Eixo horizontal: X_1 e eixo vertical: X_2
Fonte: Hyvärinen e Oja (2000)

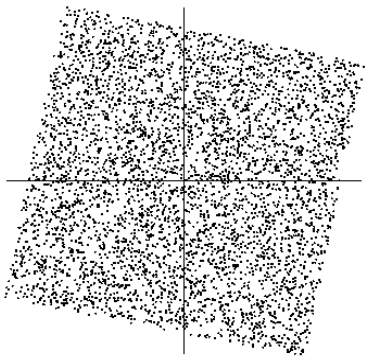

Figura 3 Distribuição conjunta das misturas dos componentes independentes distribuidos uniformemente após o branqueamento
Fonte: Hyvärinen e Oja (2000)

2.2.8 Métodos de estimação na ICA

Os princípios de estimação do modelo ICA utilizam cálculos que geralmente necessitam de uma álgebra linear mais elaborada. Assim, algoritmos numéricos são parte integrante dos métodos de estimação ICA. Os métodos numéricos para estimar o modelo ICA são tipicamente baseados na otimização de alguma função objetivo, seja por meio da sua maximização ou minimização.

Alguns métodos se destacaram ao longo da história da ICA. Um dos mais importantes é o método baseado na maximização da não gaussianidade dos componentes independentes, que pode ser realizado por meio do algoritmo FastICA. Mais detalhes podem ser encontrados em Hyvärinen e Oja (1997). Outro método bastante difundido é o que realiza a maximização da função de verossimilhança. Um princípio que está intimamente relacionado a este método é o Infomax (BELL; SEJNOWSKI, 1995). Uma abordagem também importante para a estimação dos componentes independentes, inspirado pela teoria da informação, é a minimização da informação mútua (COMON, 1994). Têm-se também os métodos tensoriais, que segundo Hyvärinen, Karhunen e Oja (2001), provavelmente, representam a primeira classe de algoritmos que realizaram ICA com sucesso. Nessa categoria é utilizado o algoritmo JADE, proposto em Cardoso e Souloumiac (1993).

Neste trabalho as análises ficarão restritas ao algoritmo FastICA, considerando a maximização da não gaussianidade dos componentes independentes.

2.2.8.1 Maximização da não gaussianidade

Foi visto anteriormente que a característica de não gaussianidade é extremamente importante na estimação do modelo de ICA. Portanto, não é surpreendente que se utilize este princípio como base na estimação dos parâmetros do modelo. A ideia é maximizar a não gaussianidade dos componentes independentes. Uma motivação é o Teorema do Limite Central. Enunciados e demonstrações deste teorema podem ser encontrados em Casella e Berger (2010).

O Teorema do Limite Central (TLC) é um resultado muito importante na teoria da probabilidade e também na ICA. Em linhas gerais, segundo esse teorema, a soma de variáveis aleatórias independentes tende para uma distribuição gaussiana, sob certas condições. Assim, a soma de variáveis independentes geralmente

tem distribuição mais próxima da gaussianidade do que qualquer das variáveis originais.

Por meio da equação 2.2, pode-se intuir, a partir do TLC, que as variáveis do vetor **X** tem distribuição mais próxima da gaussiana do que qualquer um dos componentes independentes do vetor **S**.

Considerando que se queira estimar um dos componentes independentes, tem-se, a partir da equação 2.3, que

$$Y = \mathbf{b}^T\mathbf{X},$$

em que Y é a variável que representa um dos componentes independentes e **b** é o vetor a ser determinado.

O ponto chave da ICA é encontrar o vetor **b** que maximiza a não gaussianidade de $\mathbf{b}^T\mathbf{X}$. Isso porque o componente Y é certamente uma combinação linear do vetor **S**, já que o vetor **X** é também uma combinação linear de **S**. Assim, de acordo com o TLC, o componente Y é geralmente mais gaussiano do que qualquer um dos componentes independentes S_i e, torna-se menos gaussiano quando, de fato, é igual a um dos S_i. Mais detalhes podem ser encontrados em Hyvärinen, Karhunen e Oja (2001).

Visto que neste trabalho será utilizado o algoritmo FastICA, a seguir serão abordadas as medidas de curtose e negentropia, que são as principais medidas para se avaliar não gaussianidade.

2.2.8.1.1 Curtose

Curtose é uma medida muito importante quando se fala de não gaussianidade na estimativa do modelo de ICA.

A curtose de uma variável Y, denotado por $kurt(Y)$, é definida por

$$kurt\,(Y) = E\left(Y^4\right) - 3\left[E\left(Y^2\right)\right]^2.$$

Vale lembrar que nesse contexto todas as variáveis tem média zero. Assim, em casos gerais, a definição de curtose seria diferente da apresentada.

A curtose é uma medida não nula para praticamente todas as variáveis não gaussianas e vale zero para variáveis gaussianas, justificando-se assim a sua importância como uma medida de não gaussianidade de uma variável Y. O valor da curtose também pode ser negativo ou positivo. Variáveis que possuem o valor negativo para curtose são chamadas subgaussianas ou com distribuição platicúrtica e, quando assumem valores positivos, são denominadas de supergaussianas ou com distribuição leptocúrtica.

Normalmente, a não gaussianidade é medida pelo valor absoluto da curtose, podendo também ser medida pelo quadrado da curtose. Essas medidas levam a valores zero para variáveis aleatórias gaussianas e valores maiores do que zero para a maioria das variáveis não gaussianas.

2.2.8.1.2 Negentropia

O maior problema que se enfrenta utilizando a curtose é o fato da medida ser extremamente sensível a valores extremos (*outliers*). Portanto, apesar de ser uma medida simples de não gaussianidade, não é robusta. Assim, outras medidas de não gaussianidade podem ser consideradas, dentre elas a negentropia. A negentropia é uma medida baseada na quantidade de informação teórica da entropia diferencial, ou simplesmente, entropia.

Entropia é uma medida que está relacionada ao quanto de informação uma

variável pode disponibilizar. A entropia será maior quanto mais imprevisível ou aleatório for a variável. A entropia H de uma variável aleatória Y (qualquer variável, não necessariamente um componente independente), com densidade $f_Y(y)$ é definida como

$$H(Y) = -\int_y f_Y(y) \ln f_Y(y)\, dy.$$

Um resultado fundamental relacionado à entropia é que uma variável gaussiana tem a maior entropia entre todas as outras variáveis aleatórias de igual variância. Desta forma, a entropia pode ser usada como uma medida de não gaussianidade.

Como medida de não gaussianidade associada a entropia, utiliza-se frequentemente uma versão normalizada da entropia, denominada de negentropia. Essa medida é sempre não negativa e vale zero para uma variável gaussiana. A negentropia J é, então, definida da seguinte maneira

$$J(Y) = H(Y_{gauss}) - H(Y),$$

em que Y_{gauss} é uma variável aleatória gaussiana que possui a mesma variância de Y.

Negentropia é uma medida de não gaussianidade mais robusta a *outliers* do que a curtose. Porém, pela sua definição, esta medida exige que se conheça a função densidade de probabilidade da variável aleatória em estudo, tornando assim os cálculos de extrema complexidade. Assim, aproximações mais simples de negentropia tornam-se muito úteis na implementação de algoritmos, como por exemplo, o FastICA, que será tratado com mais detalhe a seguir.

As definições apresentadas neste tópico são válidas também para variáveis

aleatórias discretas.

2.2.8.1.3 FastICA

Considerando que as variáveis originais foram branqueadas, um dos componentes independentes pode ser obtido por

$$Y = \mathbf{w}^T\mathbf{Z},$$

em que Y é um dos componentes independentes e $\mathbf{w}$ é o vetor a ser determinado.

Assim, o objetivo da ICA guiada pela máxima não gaussianidade é encontrar o vetor $\mathbf{w}$ que maximiza a não gaussianidade do componente Y, ou seja, é preciso calcular a direção em que o valor absoluto da curtose ou da negentropia de $\mathbf{w}^T\mathbf{Z}$ cresce mais fortemente e, então, mover o vetor $\mathbf{w}$ nessa direção. Isso pode ser feito, por exemplo, por meio dos algoritmos gradiente ou iterativo de ponto fixo (FastICA). Aqui, a opção será trabalhar com algoritmos de ponto fixo usando aproximações da negentropia para medir a não gaussianidade.

O algoritmo FastICA, proposto por Hyvärinen e Oja (1997), é um método computacionalmente mais eficiente que o gradiente, tanto quando se utiliza a curtose, quanto a negentropia. Assim, o FastICA usando negentropia combina propriedades superiores do algoritmo resultante da iteração do ponto fixo com as propriedades estatísticas mais relevantes devido a negentropia.

A seguir, são apresentados brevemente os passos do algoritmo FastICA baseado na maximização de funções objetivo, sendo que estas são aproximações da medida de negentropia.

Antes de se iniciar o algoritmo propriamente dito, é necessária a escolha de uma função para compor a função objetivo e seja considerada uma boa aproxi-

mação da medida de negentropia. Algumas opções são:

$$g_1(y) = \tanh(a_1 y), \tag{2.10}$$

$$g_2(y) = y \exp\left(-y^2/2\right) \quad \text{e} \tag{2.11}$$

$$g_3(y) = y^3, \tag{2.12}$$

em que $1 \leq a_1 \leq 2$ é uma constante, considerada frequentemente igual a 1.

Após a escolha da função g o algoritmo deve passar pelas seguintes etapas:

a) centralizar os dados;

b) branquear os dados, obtendo o vetor **Z**;

c) escolher aleatoriamente valores iniciais para o vetor **w** com norma unitária;

d) calcular $\mathbf{w} \leftarrow E\left[\mathbf{Z}g\left(\mathbf{w}^T\mathbf{Z}\right)\right] - E\left[g'\left(\mathbf{w}^T\mathbf{Z}\right)\right]\mathbf{w}$, em que g é escolhida dentre as opções definidas nas equações 2.10, 2.11 e 2.12 e g' é a sua derivada;

e) normalizar o vetor **w** obtido no passo d;

f) se não convergir, voltar para o passo d.

A convergência é alcançada quando os valores antigo e novo de **w** apontam na mesma direção.

O algoritmo apresentado anteriormente estima apenas um componente independente. Para se estimar mais de um componente, recorre-se a métodos de ortogonalização em que os componentes são estimados um de cada vez ou em paralelo. Esses métodos são chamados de Ortogonalização Deflacionária e Ortogonalização Simétrica, respectivamente. Mais detalhes podem ser encontrados em

Hyvärinen, Karhunen e Oja (2001).

A análise de componentes independentes pode ser aplicada em diversas áreas. A seguir tem-se alguns exemplos bastante distintos com relação a aplicação de ICA, o primeiro relacionado a sinais de eletrocardiograma, o segundo são imagens de sementes, o terceiro são dados do nível do mar, e por fim um exemplo associado a áudios.

Guilhon, Barros e Medeiros (2005) propõem um método de compressão de eletrocardiogramas usando ICA. Os autores comparam o método proposto com outro presente na literatura. Verificaram que o método utilizando ICA gerou erros menores que os do oponente. As diferenças entre os sinais obtidos por ICA e os sinais originais não foram significativas.

Moreto (2008) estudou a ICA para misturas instantâneas aplicado na separação de sinais de áudio. Foram avaliados três algoritmos de separação de misturas: FastICA, PP (Projection Pursuit) e PearsonICA. O autor afirma que, por meio dos experimentos realizados no trabalho pôde-se validar os algoritmos avaliados.

Leite, Sáfadi e Carvalho (2013) utilizaram ICA e análise discriminante com o objetivo de classificar imagens radiográficas de sementes, separando sementes cheias de sementes com algum tipo de dano ou deformação. Nesse trabalho os autores mostraram que a metodologia proposta pode contribuir para uma avaliação rápida e menos subjetiva de imagens radiográficas de sementes.

Em seu artigo, Sáfadi (2014) propõe o uso de ICA para agrupamento de séries temporais. A partir de diferentes números de componentes independentes, e considerando a técnica de análise de agrupamento (AA), pôde-se identificar os grupos com base nos coeficientes da matriz de mistura. O uso da ICA foi considerado importante, pois além de possibilitar a comparação das séries, foi possível obter informações relevantes a partir da análise dos componentes. A metodolo-

gia foi exemplificada por meio de séries temporais do nível do mar, em diferentes países, durante 26 anos. Concluiu-se nesse trabalho, que a ICA revelou as características subjacentes presentes nas séries de dados do nível do mar, e se mostrou uma ferramenta poderosa para agrupamento de séries temporais.

2.3 Análise de agrupamento

A análise de agrupamento (AA) tem o objetivo de classificar objetos, ítens ou indivíduos de acordo com suas similaridades. Os objetos semelhantes são colocados em um mesmo grupo e, consequentemente, os objetos que pertencem a diferentes grupos são considerados dissimilares.

Considerando p variáveis e n objetos, na análise de agrupamento o objetivo é agrupar os n objetos em um número desconhecido de grupos. Uma metodologia também muito conhecida é a análise discriminante, na qual o interesse é agrupar os n objetos, só que considerando um número pré-existente de grupos.

De acordo com Manly (2008), existem muitas razões pelas quais uma análise de agrupamentos pode valer a pena, como por exemplo, encontrar os verdadeiros grupos que se presume existir. A AA também pode ser útil na redução de dados, por exemplo, quando se tem um grande número de cidades que podem ser potencialmente usadas para um teste de mercado, mas é viável usar apenas algumas.

Algumas vezes o interesse da AA não é agrupar os n objetos e sim as p variáveis. Portanto, neste trabalho, para tratar o assunto de forma geral será adotado o termo objeto para designar qualquer elemento, seja ele indivíduo, item, característica, etc.

Segundo Ferreira (2008), os métodos de agrupamento vão além de agrupamentos gráficos. Eles têm a intenção de identificar padrões de agregação dos

objetos de acordo com suas similaridades.

Considerando-se que foram observados p variáveis relativas a n objetos, os dados podem ser representados da seguinte forma

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix} =$$

$$= \begin{pmatrix} \mathbf{y}_1{}^T \\ \mathbf{y}_2{}^T \\ \vdots \\ \mathbf{y}_i{}^T \\ \vdots \\ \mathbf{y}_n{}^T \end{pmatrix} = \begin{pmatrix} \mathbf{y}_{(1)} & \cdots & \mathbf{y}_{(j)} & \cdots & \mathbf{y}_{(p)} \end{pmatrix},$$

em que $\mathbf{y}_i{}^T$ representa o vetor p-dimensional linha de observações do i-ésimo objeto e $\mathbf{y}_{(j)}$ é o vetor n-dimensional coluna de observações da j-ésima variável, $i = 1, 2, ..., n$ e $j = 1, 2, ..., p$.

De maneira geral, os métodos de agrupamento podem ser divididos em métodos hierárquicos e não-hierárquicos. Os métodos hierárquicos, por sua vez, podem ser aglomerativos ou divisivos. O método aglomerativo começa com os objetos dispostos em n grupos com um objeto cada e termina com todos os objetos em apenas um grupo. Já no método divisivo há uma inversão da ordem, ou seja, o método inicia com apenas um grupo contendo todos os objetos e finaliza com n grupos cada qual com um objeto. Nos métodos não-hierárquicos, o número de

grupos é definido inicialmente, daí os n objetos são alocados de maneira otimizada nestes grupos.

A técnica de AA baseia-se em duas etapas, a primeira é a escolha da medida de proximidade e a segunda é a determinação do método que será utilizado para construir os grupos. A seguir serão apresentadas algumas medidas de proximidades e alguns métodos de agrupamento. Neste estudo optou-se por utilizar apenas métodos hierárquicos aglomerativos.

2.3.1 Medidas de proximidades

Segundo Bussab, Miazaki e Andrade (1990), um conceito fundamental para se utilizar as técnicas de AA é a escolha do critério que mensure a distância entre dois objetos, ou que quantifique o quanto eles são próximos, ou seja, as medidas de proximidade. Essas medidas podem ser divididas em duas categorias, medidas de similaridades e de dissimilaridades. Tem-se que as medidas de dissimilaridades são as distâncias, enquanto que as medidas de similaridades são complementares às distâncias.

Os coeficientes obtidos a partir das medidas de proximidade são alocados numa matriz n x n chamada de matriz de proximidade. Na i-ésima linha desta matriz, encontram-se os coeficientes de proximidades entre o i-ésimo objeto e cada um dos demais objetos, inclusive ele mesmo. As medidas de proximidades são calculadas a partir da matriz de dados $\mathbf{Y}$, apresentada anteriormente.

Existem diferentes medidas de proximidades. O que diferencia essas medidas é a forma como os objetos foram medidos e a variância das variáveis. Aqui a discussão se restringirá às medidas de dissimilaridades para variáveis quantitativas.

A distância quadrática entre dois objetos r e s pode ser obtida por

$$d_{rs}{}^2 = |\mathbf{y}_r - \mathbf{y}_s|^2_{\mathbf{\Psi}} = (\mathbf{y}_r - \mathbf{y}_s)^T \mathbf{\Psi} (\mathbf{y}_r - \mathbf{y}_s) \tag{2.13}$$

em que $\mathbf{\Psi}$ representa uma métrica de interesse e $\mathbf{y}_r$ e $\mathbf{y}_s$ são vetores p-dimensionais dos objetos r e s, respectivamente, com $r, s = 1, 2, ..., n$.

De acordo com a equação 2.13 o que diferencia a distância quadrática entre os objetos r e s é a métrica de interesse. Caso a métrica $\mathbf{\Psi}$ seja a matriz identidade, a distância quadrática é a euclidiana. Este tipo de distância é adequada para conjuntos de variáveis que possuem semelhanças nas variabilidades. Se a métrica de interesse for $\mathbf{\Psi} = D^{-1} = diag\,(1/S_{kk})$, $k = 1, 2, ..., p$, então a distância é a euclidiana padronizada quadrática, sendo S_{kk} o estimador da variância da k-ésima variável ao longo da amostra de n objetos. Essa distância é apropriada para variáveis que possuem diferentes escalas, mas que são não-correlacionadas. Por fim, tem-se a distância de Mahalanobis, que leva em consideração tanto a estrutura de correlação quanto as diferenças de escalas entre as variáveis. Nesse caso $\mathbf{\Psi} = \mathbf{S}^{-1}$, em que $\mathbf{S}$ é a matriz de variâncias e covariâncias amostrais.

Uma outra importante família de distâncias é a métrica de Minkowski, da qual a distância euclidiana é um caso particular. Ferreira (2008) expõe mais detalhes com relação a esta métrica.

Na literatura são encontradas outras medidas de dissimilaridades, como por exemplo, as métricas de Gower e de Canberra e o coeficiente de Czekanowski. Veja mais informações em Bussab, Miazaki e Andrade (1990).

2.3.2 Agrupamentos hierárquicos

Quando se pensa em agrupar, a ideia intuitiva é obter todos os agrupamentos possíveis e escolher dentre eles aquele que otimiza algum critério de partição. Porém, essa análise é computacionalmente inviável quando o número de objetos é grande, mesmo com os atuais computadores. Isto mostra a importância dos métodos de agrupamentos, já que estes nos possibilitam encontrar uma solução razoável sem a necessidade de se obter todos os agrupamentos possíveis.

Existem na literatura vários métodos de agrupamentos. Dentre estes têm-se os hierárquicos, que devido a propriedade de hierarquia, possibilita a construção de um gráfico chamado de dendrograma. Segundo Mingoti (2005), o dendrograma é um gráfico de árvore no qual a medida de proximidade é colocada na escala vertical, enquanto que no eixo horizontal são marcados os objetos. As linhas verticais que partem dos objetos possuem altura correspondente ao nível em que os objetos foram considerados semelhantes, isto é, a medida de proximidade.

A seguir serão discutidos alguns métodos hierárquicos aglomerativos, entre eles, os métodos de ligação simples (mínima distância ou vizinho mais próximo), ligação completa (máxima distância ou vizinho mais distante) e ligação média (distância média).

Basicamente, o que diferem os métodos é a forma de recalcular as distâncias entre os grupos recém-formados com os demais grupos obtidos em estágios anteriores do processo iterativo. No método do vizinho mais próximo, para obter a distância entre dois grupos deve-se encontrar a mínima distância entre os seus objetos. No método do vizinho mais distante, a procura é pela máxima distância entre os objetos e no método de ligação média, a distância é a média aritmética das distâncias entre os objetos de dois grupos.

Os passos de um algoritmo geral para realizar o agrupamento de n objetos

utilizando os métodos hierárquicos aglomerativos são:

1) começar a análise com n grupos, considerando a matriz de distâncias $\mathbf{D} = [d_{rs}]$;

2) encontrar na matriz **D** o par de objetos (ou grupos) mais similar, digamos M e N;

3) juntar os grupos M e N e nomeá-lo de MN;

4) recalcular e rearranjar as distâncias na matriz **D** da seguinte forma:

a) eliminar as linhas e as colunas correspondentes aos grupos M e N;

b) acrescentar uma nova linha e uma nova coluna com as distâncias entre o grupo MN e os demais grupos, para o cálculo das distância pode-se utilizar os critérios de distância mínima, máxima e média;

5) repetir os passos 2, 3 e 4 $(n-1)$ vezes até que todos os objetos formem um único grupo.

Outro método bastante difundido é o de Ward (1963). Este método é fundamentado na "mudança de variação" entre os grupos e dentro dos grupos que se formam em cada passo do agrupamento. A ideia básica é aglomerar os grupos que minimizam a soma de quadrados dentro dos grupos. O método de Ward é realizado de acordo com o algoritmo global para os métodos hierárquicos aglomerativos, descrito anteriormente. Mais informações sobre este método podem ser encontradas em Barroso e Artes (2003).

Nascimento, Sáfadi e Silva (2011) estudaram os métodos de agrupamento hierárquico (Ward) e de otimização (Tocher) com o objetivo de formar grupos homogêneos de séries de expressão gênica e realizar previsões. Para o agrupamento

de genes com padrões de expressões gênicas similares, utilizou-se das estimativas dos parâmetros provenientes do modelo autoregressivo de ordem p para dados em painel. Os resultados mostraram que o método de Ward foi o mais apropriado para a formação de grupos de genes com séries homogêneas. Por meio dos grupos obtidos pôde-se ajustar o modelo autoregressivo apropriado e predizer a expressão gênica em um tempo futuro.

Os métodos de agrupamentos possuem características diferentes, cada qual com suas vantagens e desvantagens. De acordo com Rencher (2002), vários estudos mostraram que os métodos de Ward e da ligação média possuem os melhores desempenhos de forma geral. Ferreira (2008) afirma que o desempenho dos métodos é variável e uma boa estratégia é testar vários e considerar aquele que confirmar, de certa forma, algum tipo de agrupamento natural.

Depois de construído o dendrograma, uma questão que surge é como se deve proceder para escolher o número de grupos do conjunto de dados analisado. De acordo com Mingoti (2005), não existe uma resposta exata para esta pergunta, o que se tem são alguns critérios que podem ajudar na decisão final, dentre eles, a análise do comportamento do nível de fusão (distância), análise do comportamento do nível de similaridade, análise da soma de quadrado entre grupos, coeficiente R^2, estatística pseudo F, correlação semiparcial, estatística pseudo T^2, estatística CCC (cluster clustering criterium). A seguir são apresentados alguns trabalhos que foram revisados sobre o tema.

Um critério utilizado para se determinar o número de grupos, tanto nos métodos hierárquicos como nos não-hierárquicos, baseia-se em uma estatística que possui distribuição aproximada F. A estatística envolvida no teste é chamada pseudo F (CALINSKI; HARABASZ, 1974). De acordo com Rencher (2002) a aproximação F é muito imprecisa, e deve ser utilizada somente para uma análise

descritiva.

Mojena (1977) propôs um critério para se determinar o número de grupos que otimiza a qualidade do ajuste do agrupamento aos dados. O objetivo foi buscar a maior amplitude nas distâncias de junção dos grupos formados.

Milligan e Cooper (1985) realizaram a comparação de 30 critérios diferentes para a escolha do número de grupos em métodos não-hierárquicos. Dentre os métodos avaliados, encontram-se pseudo F, pseudo T^2 e *cubic clustering criterion* (CCC). Os resultados apresentados neste trabalho mostram que os dois primeiros métodos aparecem em destaque como bons indicadores do número de grupos.

Peck, Fisher e Ness (1989) introduziram a metodologia de *bootstrap* para encontrar o intervalo de confiança para o número de grupos. Essa técnica é proposta para casos em que o método de agrupamento tem uma função-objetivo bem definida. Nesse artigo, os resultados apresentados são provenientes de simulações de Monte Carlo, considerando como base o procedimento das k-médias e populações normais multivariadas. Para outros métodos não-hierárquicos, como por exemplo Fuzzy c-Médias, a metodologia de *bootstrap* é considerada simples, pois tal método é fundamentado numa função objetivo.

Para métodos não-hierárquicos, Felix (2004) mostra que a metodologia de *bootstrap* pode ser aplicada, desde que se defina uma estratégia de escolha do número de grupos em cada amostra *bootstrap*.

Martins, Pedro e Rosa (2004) apresentam vários critérios para escolha do número de grupos. Os autores afirmam que o critério mais simples utilizado para decidir o número de grupos é a análise subjetiva dos diferentes níveis do dendrograma. Porém esse procedimento é viesado pela opinião do analista. Um outro procedimento é a utilização dos coeficientes de fusão, como por exemplo, Mojena (1977), Mojena revisto, Upper Tail e Médias-móveis. No presente trabalho, ou-

tros critérios são mencionados, os quais são realizados desde que a matriz de proximidades original seja definida pela distância euclidiana ao quadrado, são eles: pseudo-F, pseudo T^2, Beale's, F-ratio, R^2, R^2 semi-parcial, RMSSTD e CCC. Martins, Pedro e Rosa (2004) afirmam que essas medidas não são exatas e devem ser utilizadas em conjunto com outros métodos, que validem o agrupamento. Podem também não conduzir à escolha de uma única solução final mas à detecção de várias candidatas a solução final.

Faria (2009) afirma que a falta de critérios objetivos para se determinar o ponto de corte em um dendrograma ainda é um problema em estudos que utilizam a análise de agrupamentos. Em seu trabalho, a autora compara alguns critérios para determinação do número ótimo de grupos. Os critérios usados foram: o método de Mojema (1977), o método baseado nas trajetórias das curvas dos índices RMSSTD (Root Mean Square Standard Deviation) e RS (R-square), utilizando o método da máxima curvatura modificado. Este último determina o ponto de curvatura máxima das referidas curvas, ponto este que determina o número ótimo buscado. Segundo Sharma (1996), o índice RMSSTD é usado para calcular a homogeneidade dos agrupamentos, enquanto que o índice RS é usado para calcular a dissimilaridade entre agrupamentos.

2.3.3 Validação do agrupamento

Validar o agrupamento significa certificar-se de que os grupos realmente diferem. Nesta etapa da análise podem ser empregados vários testes, dentre eles, análise de variância multivariada (MANAVA), análise discriminante e correlação cofenética. Neste trabalho as análises ficarão restritas à correlação cofenética.

Segundo Barroso e Artes (2003), a correlação cofenética é uma medida utilizada principalmente nos métodos de agrupamentos hierárquicos. O objetivo

dessa medida é comparar as distâncias observadas entre os objetos e as distâncias recuperadas pela análise de agrupamento. Para isso, calcula-se a correlação entre essas distâncias. Assim, se o valor da correlação cofenética for relativamente alto, pode-se afirmar que o agrupamento teve boa qualidade.

2.4 Medidas de similaridades

Em estudos relacionados à análise de dados georeferenciados, o grau de similaridade ou dependência espacial é avaliada por meio de indicadores de autocorrelação espacial.

Dois indicadores de dependência espacial bastante utilizados são: o índice de Moran e o índice de Geary. Esses índices resumem a dependência espacial de toda a região em estudo, em um único valor, por isso também são chamados de índices globais. Esses indicadores são encontrados a partir da comparação entre os valores da amostra em uma certa área e de seus vizinhos.

Assim, em estatística espacial, os índices apresentados são utilizados em um contexto em que a posição geográfica dos objetos em estudo é indispensável, e o esperado é que localizações próximas tendem a ser mais similares do que locais mais distantes.

Nesta tese, os índices de Moran e Geary serão utilizados em um novo contexto, pois a ideia é que a similaridade esteja vinculada às respostas obtidas na análise de agrupamento. Assim, locais considerados similares oriundos na AA não necessariamente estão próximos geograficamente, o mesmo ocorrendo com locais distantes que não necessariamente são dissimilares.

Para o cálculo dos índices de Moran e Geary, há a necessidade da construção da matriz de proximidade, que será abordada a seguir considerando o conceito na sua essência.

2.4.1 Matriz de proximidade

Dado um conjunto de n áreas ($A_1, A_2, ..., A_n$), cada um dos elementos da matriz de proximidade **W**, representado por w_{ij}, tem como valor uma medida de proximidade entre A_i e A_j. Quando $i = j$ o elemento w_{ij} recebe o valor 0. A matriz **W** é uma matriz quadrada, que possui n linhas e n colunas.

Bailey e Gatrell (1995) abordam a construção da matriz **W** considerando diferentes medidas de adjacências, tais como:

a) $w_{ij} = 1$, se o ponto de referência de A_j é um dos k pontos mais próximos ao ponto de referência de A_i, e $w_{ij} = 0$, caso contrário;

b) $w_{ij} = 1$, se o ponto de referência de A_j está dentro de uma distância estipulada do ponto de referência de A_i, e $w_{ij} = 0$, caso contrário;

c) $w_{ij} = 1$, se A_i tem fronteira comum com A_j, e $w_{ij} = 0$, caso contrário;

d) $w_{ij} = c_{ij}/c_i$, em que c_{ij} é o comprimento da fronteira entre A_i e A_j e c_i é o perímetro de A_i;

e) $w_{ij} = 1/(d_{ij})^p$, em que $p \in \mathbb{N}^*$ e d_{ij} é a distância entre os centroides de A_i e A_j.

Neste trabalho, será utilizado o critério de padronização das linhas da matriz **W**, recomendado por Waller e Gotway (2004) e Druck et al. (2004). A normalização refere-se à divisão de cada elemento da matriz pelo total da linha e assim $w_{i.} = \sum_{j=1}^{n} w_{ij} = 1$. Desta maneira, os pesos w_{ij} relacionados com a área A_i somam 1.

Daqui pra frente, nesta tese, quando se falar em áreas vizinhas, entenda áreas que pertencem a um mesmo grupo de acordo com a AA. Assim, a depen-

dência espacial estará relacionada à similaridade existente entre os membros dos grupos obtidos pela AA.

2.4.2 Índice de Moran

O índice de Moran (MORAN, 1950) é uma medida de autocorrelação entre um valor numa certa área e os valores de seus vizinhos, possibilitando obter padrões significativos de associação.

O índice de Moran é calculado por:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2},$$

em que:

n é o número de áreas ou de observações;

Y_i é a variável aleatória na área i;

Y_j é a variável aleatória na área j;

$\bar{Y}$ é a média amostral da variável aleatória em toda a região; w_{ij} são os elementos da matriz de proximidade.

O índice de Moran, na grande maioria da vezes, está entre -1 e 1, raramente é maior do que 1 ou menor do que -1. Seu valor se afasta de zero à medida que aumenta o grau de correlação positiva ou negativa. Quando o valor se aproxima de zero há indícios de independência entre as áreas, caso o índice apresente valores maiores do que 0 as áreas vizinhas tendem a ser similares entre si, e por fim, valores negativos indicam dissimilaridade entre as vizinhanças.

Após o cálculo do índice de Moran, é importante que se faça algum teste para avaliar a sua significância. A hipótese nula para tal teste tem como afirmação

a independência entre as áreas, ou seja, o valor atribuído ao índice seria estatisticamente igual a zero, enquanto que a hipótese alternativa afirma que há dependência entre as vizinhanças (positiva ou negativa), ou seja, valores do índice estatisticamente maiores que 0 ou menores que 0, respectivamente.

Segundo Bailey e Gatrell (1995), duas principais abordagens podem ser consideradas para testar os valores observados de I, considerando a hipótese de que não há autocorrelação entre as observações vizinhas. Uma delas leva em consideração a distribuição aproximada de I, e a outra a distribuição empírica para os possíveis valores de I.

Cliff e Ord (1981) afirmam que o índice de Moran, sob a hipótese de independência entre as observações, segue assintoticamente uma distribuição normal, com média e variância definidas. Porém, a normalidade atribuída ao índice de Moran, depende da existência de muitas áreas, o que nem sempre ocorre. Uma alternativa é o teste de permutação aleatória, que será abordado a seguir.

Um fato importante em relação ao índice de Moran, é que para a sua aplicação o processo estocástico deve apresentar estacionariedade de segunda ordem. Segundo Cressie (1993), um processo estocástico é dito estacionário de segunda ordem se três critérios forem atendidos, são eles: primeiro, a média do processo em qualquer ponto da região não deve depender da posição, segundo, a variância do processo tem que ser constante em toda a região de estudo, e terceiro, a covariância entre dois pontos quaisquer da região, é uma função da distância entre estes pontos.

2.4.2.1 Gráfico de espalhamento de Moran

Segundo Anselin (1996), o objetivo do gráfico de espalhamento de Moran é comparar os desvios $\left(Z_i = Y_i - \bar{Y}\right)$ da variável numa área A_i, com a média dos

desvios de seus vizinhos A_j ponderada pela matriz de proximidade **W** padronizada pelas linhas $\left(WZ_i = \sum_{j=1}^{n} w_{ij} \left(Y_j - \bar{Y} \right) \right)$.

O gráfico apresentado, é portanto um gráfico bidimensional, em que no eixo das abscissas encontra-se Z, enquanto que no eixo das ordenadas, tem-se WZ. Na literatura é também conhecido como diagrama de espalhamento de Moran. Na Figura 4, tem-se uma ilustração do referido diagrama, com os quatro quadrantes (Q1, Q2, Q3 e Q4) separados por linhas pontilhadas.

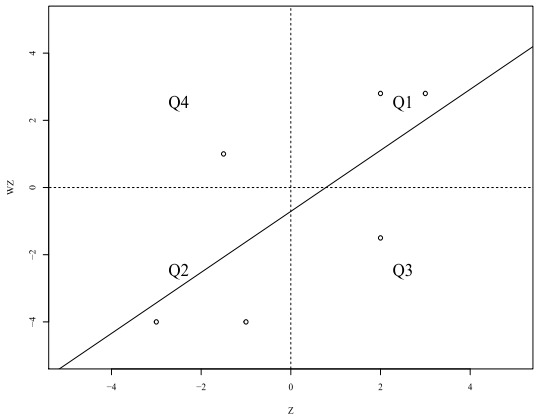

Figura 4 Gráfico de espalhamento de Moran

Os quatro quadrantes da Figura 4 têm as seguintes interpretações:

a) o primeiro quadrante apresenta áreas com altos valores para a variável em análise, cercadas por áreas cuja média também é alta para a variável em análise;

b) o segundo quadrante contém áreas com baixos valores para a variável em análise, e que apresentam vizinhanças cuja média também é baixa em relação à mesma variável;

c) o terceiro quadrante mostra locais com altas medidas para a variável em

análise, envolvidos por locais com comportamento médio baixo para a variável em análise;

d) o quarto quadrante indica locais com baixas medidas para a variável em análise, circundados por locais cuja média é alta para a mesma variável.

Considerando essas interpretações, têm-se que os quadrantes 1 e 2, apresentam áreas com autocorrelação espacial positiva, indicando assim, áreas que possuem vizinhos com valores semelhantes da variável em estudo. Já os quadrantes 3 e 4, apresentam áreas com autocorrelação espacial negativa, ou seja, há dissimilaridade entre as vizinhanças destes quadrantes, com relação à variável analisada.

Tem-se que a inclinação da reta de regressão de WZ versus Z também pode ser utilizada para avaliar a autocorrelação espacial. Caso a inclinação da reta seja positiva, há indícios de autocorrelação espacial positiva, enquanto que se a inclinação da reta for negativa, pode-se supor a existência de autocorrelação espacial negativa.

2.4.2.2 Teste de permutação aleatória de Moran

Dados n valores da variável aleatória Y associados às áreas A_i do mapa, é possível obter $n!$ permutações das observações y_i entre as áreas. Para cada uma das permutações obtidas, existe um valor para o índice I, e um deles corresponde ao valor de I associado ao arranjo original dos dados no mapa. A partir de todas as permutações possíveis, obtém-se uma distribuição empírica dos valores de I.

Pode-se a partir daí construir um teste de hipótese para verificar a significância do valor de I. Esse teste tem como hipótese nula a independência entre as observações, ou seja, independência entre as áreas A_i do mapa, segundo a variável

aleatória Y.

Uma maneira de garantir a realização desse teste, é justamente realizando as $n!$ permutações das observações y_i entre as áreas, pois este fato contribui com a ideia fundamental do teste, que considera as variáveis aleatórias $Y_1, Y_2, ..., Y_n$ independentes e identicamente distribuídas.

O teste descrito é chamado de teste de permutação aleatória. Se I correspondente ao valor do índice associado ao arranjo original dos dados no mapa estiver em um dos extremos da distribuição empírica, há evidência então de autocorrelação espacial.

O algoritmo a seguir sintetiza o teste de permutação aleatória para o índice de Moran.

1) Calcule o índice de Moran para o arranjo original dos dados no mapa, obtendo-se $I_{(0)}$.

2) Calcule o índice $I_{(k)}$ a partir das permutações das observações y_i, em que k é um valor que varia de 0 a N, com N menor ou igual a $n!$. A única limitação para N é a viabilidade computacional.

3) Devido ao fato das variáveis aleatórias Y_i serem independentes e identicamente distribuídas, todas a permutações das observações entre as áreas tem a mesma probabilidade de ocorrer. Dessa maneira, o $valor - p$ para o teste considerando autocorrelação positiva é dado por:

$$valor - p = \frac{número\, de\, I_{(j)} \; > \; I_{(0)}}{N+1}, \; j = 1, ..., N.$$

Caso $valor - p < \alpha$, rejeita-se H_0, ao nível de significância de α.

Quando a autocorrelação é negativa, o $p - valor$ do teste é:

$$valor - p = \frac{número\, de\, I_{(j)} \ < \ I_{(0)}}{N+1},\ j = 1, ..., N.$$

De forma análoga, se $valor - p < \alpha$, rejeita-se H_0, ao nível α.

De acordo com Cliff e Ord (1981), a distribuição empírica de I a partir das permutações dos n valores da variável aleatória Y, continua sendo normal sob hipóteses bem gerais. Segundo estes autores, a média e a variância de I são dadas sob a hipótese de normalidade por:

$$E\left[I\right] = -\frac{1}{n-1}$$

e

$$Var\left[I\right] = \frac{n\left[\left(n^2 - 3n + 3\right) S_1 - nS_2 + 3{S_0}^2\right] - k\left[\left(n^2 - n\right) S_1 - 2nS_2 + 6{S_0}^2\right]}{\left(n-1\right)\left(n-2\right)\left(n-3\right){S_0}^2} - \left(-\frac{1}{n-1}\right)^2,$$

em que: $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}, S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2$, $S_2 = \sum_{i=1}^{n}\left(w_{i.} + w_{.i}\right)$ e $k = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^4 \Bigg/ \left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right]^2$. Tem-se que: $w_{i.} = \sum_{j=1}^{n} w_{ij}$ é a soma da linha i da matriz **W** e $w_{.i} = \sum_{j=1}^{n} w_{ji}$ é a soma da coluna i da matriz **W**.

2.4.3 Índice de Geary

O índice de Geary (GEARY, 1954) difere do I de Moran por utilizar a diferença entre os pares de áreas, e não a diferença entre cada ponto e a média global, como é o caso do índice de Moran. Segue daí que o índice de Geary é uma medida sensível às diferenças em pequenas distâncias, enquanto o índice de Moran é mais sensível a valores extremos.

O índice de Geary é dado por:

$$C = \frac{(n-1)}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - Y_j)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \tag{2.14}$$

em que:

n é o número de áreas ou de observações;

Y_i é a variável aleatória na área i;

Y_j é a variável aleatória na área j;

$\bar{Y}$ é a média amostral da variável aleatória em toda a região e

w_{ij} são os elementos da matriz de proximidade espacial.

O valor do índice de Geary C, na maioria das vezes varia entre 0 a 2. A autocorrelação positiva é encontrada quando o valor de C está entre 0 e 1, e negativa quando o valor de C está entre 1 e 2. O índice de Geary pode assumir valores maiores do que 2, porém, em raras ocasiões (GRIFFITH, 1987).

Segue daí que: caso o índice de Geary apresente valores menores do que 1, as áreas vizinhas tendem a ser similares entre si, e por fim, valores maiores do que 1 indicam dissimilaridade entre as vizinhanças. Têm-se também que valores de Geary que se aproximam de 1, mostram indícios de que exista independência espacial entre as áreas estudadas.

A seguir, como ilustração, tem-se uma breve interpretação do índice de Geary. Para isto será reescrito a equação 2.14, de uma maneira que fique mais simples a sua explicação (ARANHA, 1999).

$$C = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - Y_j)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \Bigg/ \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{(n-1)} \tag{2.15}$$

Na equação 2.15, tem-se, no lado esquerdo da razão, a variabilidade de vizinhança. Nessa variabilidade, caso fosse retirado w_{ij}, o valor obtido representaria a soma das diferenças ao quadrado entre cada particular resultado da variável Y e todos os demais. No entanto, w_{ij} assume o valor 0 quando as áreas comparadas não são vizinhas, anulando assim o resultado da comparação. Assim, acabam só entrando na soma os desvios entre os vizinhos, já que são multiplicados por 1. Observa-se que no denominador $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, corresponde ao número total de parcelas somadas, caso seja utilizado o recurso de ponderação por linhas da matriz W, este valor corresponde a 1. O valor 2 ajusta o fato das diferenças atribuídas ao numerador da expressão variabilidade da vizinhança serem contadas em dobro.

Portanto, a parte da equação 2.15 intitulada variabilidade de vizinhança, corresponde à variância das observações definidas por w_{ij} como vizinhas entre si.

O lado direito do quociente na equação 2.15, corresponde a variabilidade geral, ou seja, a variância dos dados, que é uma medida de dispersão da variável Y em relação a média geral.

A partir da equação 2.15 e considerando uma variável qualquer, tem-se que, se a variabilidade de vizinhança for próxima da variabilidade geral, o índice C deve ser próximo de 1. Porém, se áreas vizinhas têm a tendência de serem similares, a variância de vizinhança será menor do que a variância geral, segue daí que C será menor que 1. Ao contrário, se áreas com grandes valores para a variável

analisada tiverem em seu entorno áreas com pequenos valores (ou vice-versa), a variabilidade entre os vizinhos será maior que a variabilidade geral, e o índice de Geary será maior que 1.

Da mesma forma como foi feito para o índice de Moran, após o cálculo do índice de Geary, é necessário que se faça um teste para avaliar se o valor encontrado é ou não significativo. A hipótese nula para o teste de Geary também considera a afirmação de independência espacial, enquanto que a hipótese alternativa afirma dependência espacial entre as áreas consideradas vizinhas.

Segundo Sokal e Oden (1978), o teste de significância do índice de Geary, pode ser realizado, considerando tanto a distribuição aproximada de C, quanto a distribuição empírica para os possíveis valores de C, da mesma maneira como foi feito para o índice de Moran.

De acordo com Cliff e Ord (1981), o índice de Geary segue assintoticamente uma distribuição normal, com média e variância definidas. Porém, há a necessidade de uma amostra grande para se supor a normalidade dos índices. Portanto, optou-se também por utilizar a distribuição empírica de C. O procedimento é idêntico ao apresentado para o índice de Moran. A seguir têm-se mais detalhes sobre tal teste.

2.4.3.1 Teste de permutação aleatória de Geary

De acordo com Cliff e Ord (1981), a média e a variância de C sob permutação dos índices e suposição de normalidade são dadas por:

$$E\left[C\right] = 1$$

e

$$Var\left(C\right) = \frac{1}{n\left(n-2\right)\left(n-3\right){S_0}^2}\left\{\left(n-1\right)S_1\left[n^2-3n+3-\left(n-1\right)k\right]+\right.$$
$$\left.{S_0}^2\left[n^2-3-\left(n-1\right)^2k\right]-\frac{1}{4}\left(n-1\right)S_2\left[n^2+3n-6-\left(n^2-n+2\right)k\right]\right\}$$

Todas as variáveis de $E\left[C\right]$ e $Var\left[C\right]$ são iguais às definidas para o teste de significância de I de Moran.

Segundo Sokal e Oden (1978), os resultados empíricos obtidos para o I e C são similares, mas não são idênticos.

O objetivo do teste aqui apresentado é o mesmo do teste para o I de Moran, ou seja, a intenção é verificar a significância do valor de C. Na hipótese nula, têm-se que as áreas do mapa são independentes, segundo a variável analisada, enquanto que na hipótese alternativa, esta afirmação é refutada, ou seja, existe dependência entre as áreas (A_i) do mapa. Quando há independência, o valor de C é considerado estatisticamente igual a 1, caso contrário C é estatisticamente diferente de 1, indicando assim dependência negativa ou positiva, respectivamente.

O algoritmo para o teste de permutação aleatória do índice de Geary é idêntico ao apresentado para o índice de Moran.

Em geral, Moran e Geary apresentam conclusões semelhantes. No entanto, Moran é o preferido na maioria dos casos. Isso se deve ao fato de que, segundo Cliff e Ord (1981) Moran é mais poderoso que Geary.

2.4.4 Aplicações das medidas de proximidade

A seguir tem-se algumas referências presentes na literatura na área de estatística espacial. Em todos os trabalhos apresentados a matriz de proximidade sempre é obtida a partir de vizinhanças geográficas.

Bailey e Gatrell (1995) abordaram a construção da matriz de proximidade espacial considerando diferentes medidas de adjacências, como por exemplo, se o ponto de referência de uma área qualquer está dentro de uma distância estipulada do ponto de referência de outra área qualquer, ou se uma área tem fronteira comum com outra determinada área, entre outras.

Aranha (1999) afirma que nos modelos para estimação de potencial de mercado, por meio de regressão linear simples e múltipla, com certa frequência são obtidos resultados insatisfatórios em virtude da ocorrência da autocorrelação espacial. O autor propõe o uso do índice de Geary para avaliar a presença e a intensidade de tal fenômeno. Nesse trabalho foi construído um modelo para a estimação de área de loja de supermercados em municípios paulistas, que incorpora o parâmetro de autocorrelação espacial. Os resultados mostraram que a incorporação desse parâmetro melhora consideravelmente o desempenho do modelo

Segundo Assunção (2001), a matriz de proximidade espacial pode ser assimétrica em algumas situações. Por exemplo, quando o elemento na linha i e coluna j da matriz de proximidade for proporcional ao fluxo de pessoas residentes em i que trabalham em j ou proporcional ao fluxo de telefonemas partindo de residências em i e comunicando-se com residências em j. Uma das escolhas mais comuns para a matriz de proximidade é tomar apenas uma matriz binária com o valor 1 se as áreas i e j compartilham fronteiras e 0, caso contrário, daí tem-se uma matriz simétrica. Para Assunção (2001), uma matriz assimétrica sempre pode ser transformada numa simétrica, por exemplo, redefinindo novos pesos.

Manuel (2011) estudou a distribuição espacial da ocorrência de mortalidade infantil na cidade de Alfenas, MG. A dependência entre as observações de mortalidade infantil foi avaliada por meio dos índices locais e globais de Moran. Nesse trabalho, fez-se também a modelagem dos dados por meio do modelo de

regressão clássico, modelo espacial autoregressivo (SAR) e modelo de erro espacial (CAR). As variáveis independentes utilizadas nos modelos foram o número de mulheres em idade fértil, o número de mulheres em idade de risco gestacional, entre outras. Como variável dependente, usou-se o número de óbitos com menos de um ano de idade. Pelo critério de Akaike (AIC), o modelo SAR foi considerado o melhor modelo, sendo que o parâmetro para dependência espacial foi negativo e estatisticamente diferente de zero, e somente as variáveis número de mulheres em idade fértil e a renda mensal da mulher exercem influência no modelo.

Ferreira (2012) empregou metodologias associadas à estatística espacial e estatística espaço-temporal para avaliar o número de casos de dengue na cidade de Lavras - MG no período de 2007 à 2010. Os resultados encontraram as áreas de maior risco de surto epidêmico.

3 MATERIAL E MÉTODOS

3.1 Materiais

Os dados utilizados nesta tese foram obtidos junto à Universidade Federal de Lavras, por meio das contas de energia elétrica emitidas pela CEMIG (Centrais Elétricas de Minas Gerais) e por meio de medidores da própria universidade, instalados na cabine de medição da UFLA e em alguns departamentos/setores do campus.

Os medidores de energia elétrica utilizados são do modelo Spectrum SX, de fabricação da Nansen S. A. Segundo Jesus (2011), este tipo de equipamento é uma evolução da linha de medidores eletrônicos, capaz de atender as concessionárias de energia elétrica em leituras de alta performance.

Os valores obtidos na cabine de medição da UFLA são dados gerais da universidade, ou seja, englobam todos os departamentos/setores. Já os medidores instalados nos departamentos/setores geraram dados referentes a cada local separadamente.

Além da cabine de medição da UFLA, os locais que receberam medidores foram os seguintes: Cantina Central (CC), Centro de Informática (CIN/UFLA), Departamento de Administração e Economia (DAE), Departamento de Ciências da Computação - Pavimento Térreo (DCC-T), Departamento de Ciência da Computação - Primeiro Pavimento (DCC-PP), Departamento de Ciências Exatas (DEX), Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG), Restaurante Universitário (RU), Departamento de Química (DQI), Fitotecnia e Setor de Sementes.

As medições obtidas pelos medidores da universidade possuem um intervalo de 15 minutos. A variável energética mensurada foi potência ativa. Os dados foram coletados no dia 12/08/2010 (quinta-feira) na CC, CIN/UFLA, DAE, DCC-

T, DCC-PP, DEX, EPAMIG e RU, no dia 28/10/2014 (terça-feira) no DQI, Fitotecnia e Setor de Sementes, e no período de 10/06 a 16/06/2013 na cabine de medição da UFLA. Essas datas foram escolhidas por serem uma amostra representativa da realidade da universidade.

A coleta de dados feita no dia 28/10/2014 no DQI, Fitotecnia e Setor de Sementes, teve o objetivo de enriquecer o trabalho com dados mais atualizados.

Os dados coletados no dia 12/08/2010 e 28/10/2014 foram utilizados para avaliar os departamentos/setores. Já os dados obtidos de 10/06 a 16/06/2013 foram utilizados para comparar os diferentes dias da semana. Esses dias foram selecionados pois são uma amostra representativa da população estudada.

Os dados obtidos das contas de energia são medidas mensais e incluem as variáveis demanda de potência registrada em horário de ponta e demanda de potência registrada em horário fora de ponta. As séries são de janeiro de 1995 a dezembro de 2013, sendo que os valores referem-se exatamente ao mês em que ocorreram, não ao mês em que foi faturada a conta, da forma como foi tratado em Prado (2011).

Segundo Villamgna (2013), a UFLA, atualmente, se enquadra nos consumidores do grupo A4 (2,3 a 13,8 kV), na tarifa horo-sazonal azul. O horário de ponta definido entre a concessionária e a instituição é das 19 às 22 horas e, no horário de verão, das 20 às 23 horas.

A seguir são apresentados os passos para a realização das análises referentes às séries de dados originais.

3.2 Métodos

Considerando os dados originais da variável potência ativa, para comparar os departamentos/setores (CC, CIN/UFLA, DAE, DCC-T, DCC-PP, DEX, EPA-

MIG e RU) e os dias da semana, das variáveis demanda de potência registrada em horário de ponta e demanda de potência registrada em horário fora de ponta, para comparar os meses, têm-se que os passos foram os seguintes:

a) realização de uma análise exploratória dos dados por meio de *boxplots* e gráficos dos dados ao longo do tempo, para verificar possíveis agrupamentos naturais;

b) aplicação de AA, sendo que os objetos de comparação são os departamentos/setores, os dias da semana e os meses. Foram consideradas como variáveis as várias observações ao longo do tempo. Portanto, não são várias variáveis, e sim uma mesma variável medida várias vezes. A análise baseou-se na escolha da medida de proximidade e na determinação do método hierárquico aglomerativo para construção dos grupos. As medidas de proximidade utilizadas foram: distâncias euclidiana, euclidiana padronizada quadrática e generalizada de Mahalanobis. Os métodos empregados foram: métodos de ligação simples, ligação média, ligação completa e de Ward. O agrupamento escolhido foi aquele que maximiza o valor do coeficiente de correlação cofenética, considerando sempre aquele resultado que corrobora, de certa forma, algum tipo de agrupamento natural. Para representar os resultados finais dos agrupamentos, fez-se a construção do dendrograma;

c) construção de todos os possíveis agrupamentos distintos, considerando pontos de corte feitos no dendrograma;

d) construção da matriz de proximidade, considerando cada corte obtido na análise de agrupamento. Tem-se uma ilustração desse procedimento, a seguir.

Se o corte no dendrograma gerou os grupos A, C, E e B, D, tem-se que a matriz de proximidade obtida é dada por:

$$\mathbf{W} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Nas linhas 1, 2, 3, 4, e 5 da matriz **W**, têm-se os locais A, B, C, D e E, respectivamente. Nas colunas 1, 2, 3, 4 e 5, têm-se a mesma sequência de localidades;

e) normalização das linhas da matriz **W**, dividindo cada elemento da matriz pelo total da linha, assim a soma de cada linha da matriz resulta no valor 1. Portanto, os pesos associados com cada área (departamento/setor, dias da semana ou meses) somam um. Para o exemplo anterior, a matriz padronizada é:

$$\mathbf{W}^* = \begin{bmatrix} 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \end{bmatrix}.$$

Neste trabalho, utilizou-se nas análises a matriz normalizada nas linhas, portanto, para simplificação de notação, a matriz $\mathbf{W}^*$, que é a matriz de proximidade padronizada de **W**, será denotada pelo mesmo símbolo **W**, e seus elementos por w_{ij};

f) cálculo dos índices de Moran e Geary, a partir das vizinhanças propostas em cada corte no dendrograma, ou seja, a partir de cada matriz de proximidade obtida no item e;

g) realização dos testes de permutação aleatória de Moran e Geary para verificar a significância dos índices propostos. Para se fazer os testes, considerou-se o nível de significância de 5%. Optou-se pelos testes usando permutação aleatória em virtude do baixo número de observações. Usou-se 7! simulações para os dias da semana e 8! para os demais casos;

h) a partir dos resultados obtidos pelos testes, partir para a verificação de quais cortes são significativos, considerando tanto o índice de Moran, quanto o índice de Geary. Para que um corte seja significativo, o valor do índice de Moran deve ser positivo e estatisticamente diferente de zero, e para o índice de Geary deve ser um valor menor que 1 e estatisticamente diferente de 1, indicando assim similaridade entre as vizinhanças formadas por meio do corte realizado no dendrograma. Assim, esses índices são alternativas para se obter cortes plausíveis no dendrograma;

i) considerando os cortes significativos, seguir com a escolha daqueles que possuam a máxima autocorrelação, obtida a partir dos índices de Moran e Geary. Esses cortes serão considerados os dois agrupamentos de melhor qualidade. Esta metodologia propõe um novo critério para se determinar o número de grupos "ótimo", oriundos de um dendrograma;

j) como ilustração da metodologia proposta, formar vizinhanças que não poderiam ser encontradas de acordo com o dendrograma. Construir a matriz de proximidade, fazer as etapas e, f g e h, e emitir conclusões sobre os agrupamentos analisados;

k) construção do diagrama de espalhamento de Moran, como ferramenta complementar, cujo objetivo é visualizar similaridades e dissimilaridades presentes nos dados. Este item será executado em apenas algumas situações, visto que a intenção é realizar uma ilustração complementar;

l) com a intenção de obter informações em períodos importantes para a universidade, fez-se uma análise dos departamentos/setores em dois períodos específicos. O primeiro período das 06 às 11h45min e o segundo das 12 às 17h45min. Da mesma maneira realizou-se esta análise nos dias 10/06 a 16/06/2013, considerando os mesmos períodos enumerados anteriormente. E por fim, fez-se o estudo dos anos de 2010 a 2013;

m) considerando os períodos do item l, pode-se comparar: os departamentos/setores e os dias da semana para cada período indicado, e os meses para os anos de 2010 a 2013. As análises desse item também englobam os itens de a a k.

A seguir são apresentados com detalhes os passos para a realização das análises associadas aos componentes independentes.

Considerando os dados originais da variável potência ativa, para analisar os departamentos/setores (CC, CIN/UFLA, DAE, DCC-T, DCC-PP, DEX, EPAMIG e RU) e os dias da semana, têm-se que os passos foram os seguintes:

a) aplicação da ICA. Considerando-se que as variáveis do vetor $\mathbf{X}$ da equação 2.2 foram observadas, tem-se então que $\mathbf{X}$ e $\mathbf{S}$ passam a ser matrizes e não mais vetores. Assim, para a análise dos departamentos/setores, cada linha da matriz $\mathbf{X}$ representa os dados de cada departamento no dia 12/08/2010, na análise dos dias da semana, as linhas da matriz são amostras das variáveis dos dias 10/06 a 16/06/2013, portanto de segunda a domingo. A partir dos dados da matriz $\mathbf{X}$, faz-

se a obtenção dos componentes independentes e da matriz de pesos, isto considerando os dois casos do estudo separadamente. Para obtenção dos CI's realizou-se o branqueamento como pré-processamento para a ICA, com o objetivo de reduzir a dimensão dos dados;

b) construção dos periodogramas para cada componente independente obtido no item a;

c) a partir da matriz $\mathbf{A}$ estimada ($\hat{\mathbf{A}}$) e dos componentes independentes, ou seja, da matriz $\mathbf{S}$ estimada ($\hat{\mathbf{S}}$), foi realizado uma comparação dos departamentos/setores por meio dos coeficientes ou pesos de cada componente separadamente, aplicando AA. Além disso, foi realizado a comparação também para os dias da semana. No caso dos departamentos/setores, tem-se que em cada linha da matriz $\hat{\mathbf{S}}$ encontra-se um dos componentes independentes, e em cada casela de uma certa coluna da matriz $\hat{\mathbf{A}}$ tem-se um peso referente a cada departamento/setor, associado a um determinado componente. O objetivo é comparar os departamentos considerando estas caselas para cada componente independente. Para os dias da semana a ideia é análoga;

d) construção dos dendrogramas, considerando cada componente independente obtido em a;

e) construção de todos os possíveis agrupamentos distintos, considerando os pontos de corte feitos nos dendrogramas do item d;

f) construção da matriz de proximidade, considerando cada corte obtido na análise de agrupamento;

g) cálculo dos índices de Moran e Geary, e posteriormente realização dos respectivos testes, considerando o nível de significância de 5%, 7! simulações

para os dias da semana e 8! para os departamentos/setores. Considerando o valor significativo mais próximo de 1 para o índice de Moran, e o valor significativo mais próximo de 0 para o índice de Geary, seguiu-se com a obtenção dos agrupamentos de melhor qualidade;

h) interpretação dos componentes obtidos, e dos respectivos agrupamentos.

Todos os passos mostrados anteriormente foram realizados sempre considerando cada variável separadamente. Com relação aos cálculos dos componentes independentes, foi utilizado o algoritmo FastICA usando negentropia. Como aproximação dessa medida foi usado a função apresentada na equação 2.12, e para se estimar mais de um componente, foi adotado o método de ortogonalização deflacionária.

Para os dados de potência da Fitotecnia, Química e Sementes foi realizada somente a análise de componentes independentes. O objetivo foi complementar as análises anteriores, considerando dados mais atualizados. Como foram coletados dados de apenas três locais não se fez análise de agrupamento nem análise de similaridade por meio dos índices de Moran e Geary. A ideia foi a partir dos componentes encontrados por ICA, obter particularidades dos três departamentos/setores. Para contribuir com a obtenção de resultados relevantes fez-se também a construção de periodogramas para cada componente encontrado.

As análises foram feitas nos softwares R (R CORE TEAM, 2014) e MATLAB (MATLAB..., 2011). As análises descritivas, análises de agrupamento, índices de Moran e Geary e seus respectivos testes, foram feitos no software R. Enquanto que a análise de componentes independentes foi realizada no MATLAB.

4 RESULTADOS E DISCUSSÃO

A difusão de tecnologias e das informações geradas por esta tese ficará a disposição de pesquisadores, técnicos, e público em geral, por meio das publicações previstas, passíveis de acesso a arquivos impressos e virtuais, por meio dos meios de comunicação envolvidos.

A seguir têm-se os resultados obtidos a partir dos dados disponibilizados pela UFLA e das metodologias propostas.

4.1 Análise de agrupamento e medidas de similaridades

Neste tópico são apresentadas as análises associadas as séries originais de potência ativa de alguns departamentos/setores, dias da semana e meses da UFLA.

4.1.1 Departamentos/setores

A Figura 5 mostra as séries temporais (a) e os *boxplots* (b) da potência ativa em alguns setores da UFLA no dia 12/08/2010. Observa-se que o CIN/UFLA e o RU são, dentre os locais estudados, os que mais se diferenciam. O RU tem um acréscimo significativo no valor da potência ativa no período das 7 às 14 horas aproximadamente, que corresponde exatamente o horário de preparação do almoço e realização do almoço em si. Neste período, os funcionários do RU fazem uso de equipamentos de alta potência, como por exemplo, fritadeiras elétricas, máquinas de lavar louça e de higienização do ambiente. Como nessa data a universidade não servia o jantar, durante o período noturno não existia esse acréscimo. Já o CIN/UFLA tem um comportamento bastante homogêneo em praticamente todas as 24 horas do dia, sendo notado apenas um leve acréscimo na potência ativa

a partir das 7 horas aproximadamente, mantendo-se constante neste patamar ao longo do dia e início da noite. Nesse setor, a constância na potência ativa deve-se ao uso permanente de servidores e computadores. O aumento a partir das 7 horas é em virtude da chegada de funcionários para trabalhar, assim computadores e ar-condicionados são ligados.

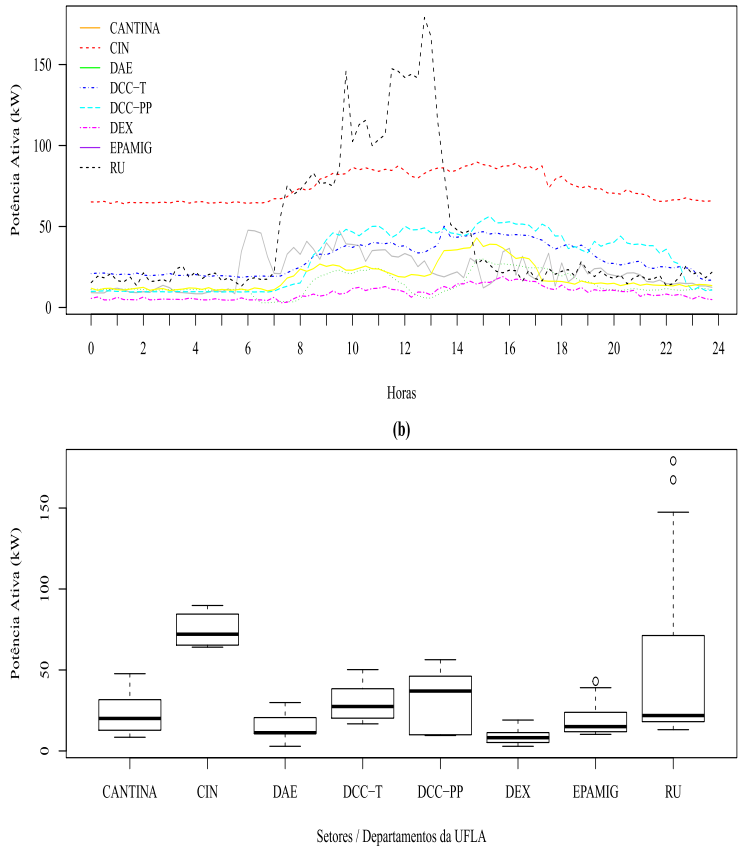

Figura 5 Séries temporais (a) e *boxplots* (b) da potência ativa em alguns setores/departamentos da UFLA, medida em kW, de 15 em 15 minutos no dia 12/08/2010

Observa-se pela Figura 5 (b) que o RU possui a maior variabilidade com

relação à potência. Enquanto que o DEX, DAE e a EPAMIG são os menos dispersos, sendo o DEX o de menor variância. Observa-se também a existência de um e dois *outliers* nos *boxplots* que representam a EPAMIG e o RU, respectivamente. Esses valores mostram que durante certos momentos do dia a potência atinge, nesse caso, altas medidas, podendo ocasionar multas para a universidade por excesso de demanda.

O DEX, DAE e a EPAMIG são locais que têm baixo consumo de energia elétrica, pois possuem apenas salas de professores, laboratórios e salas de atividades administrativas. Assim como o CIN, esses departamentos/setores possuem um crescimento da potência no período da manhã, cuja explicação está associada a chegada de alunos e funcionários. A partir de um certo momento os valores de potência se estabilizam, permanecendo assim ao longo do dia e início da noite.

Na CANTINA, observa-se no período das 6 às 7 horas valores de potência em uma patamar mais alto. Porém, a partir das 7 horas, volta a valores mais baixos. Esse comportamento é esperado, pois nesse horário estão preparando o café da manhã para servir aos alunos e funcionários do turno da manhã. A partir das 7 horas, até o fim da noite, o que pode ser observado de mais relevante são alguns pequenos picos na potência, os quais são explicados pela ocorrência de intervalos nas aulas. À tarde não ocorre um pico maior de potência, pois o que é servido no café da tarde já foi preparado no período das 6 às 7 horas.

Os dois departamentos restantes, DCC-T e DCC-PP, possuem comportamentos semelhantes. Observa-se um crescimento da potência a partir das 8 horas, também decorrente do início das atividades acadêmicas e administrativas. Logo em seguida, têm o comportamento da potência estabilizado. O maior crescimento pode ser observado no DCC-PP, local que tem alguns laboratórios, salas de aula e de professores.

A Figura 6 mostra o dendrograma selecionado para os setores/departamentos. Observa-se que os locais com menor distância entre si são o DEX e o DAE, que tem como particularidade o baixo valor da potência ativa e um comportamento homogêneo praticamente todo o tempo.

O dendrograma apresentado aqui resultou em um coeficiente de correlação cofenética de 0,9632, mostrando uma boa qualidade do agrupamento. A distância utilizada na construção desse dendrograma foi a distância euclidiana, pois dentre as medidas propostas, esta foi a que obteve os agrupamentos mais coerentes com a realidade. Esse fato foi observado em todas as situações desta tese. Teve-se também que em todos os casos o método utilizado na construção do dendrograma foi o de ligação média, pois esse obteve dentre todos os métodos o maior coeficiente de correlação cofenética.

Figura 6 Dendrograma para agrupamento de alguns setores/departamentos da UFLA, obtido a partir de dados da potência ativa da UFLA, medidas em kW, de 15 em 15 minutos no dia 12/08/2010

Alguns pontos de corte podem ser feitos no dendrograma apresentado na

Figura 6. Considerando os cortes de cima para baixo, têm-se:

a) Corte 1: CIN e CANTINA, DAE, DCC-T, DCC-PP, DEX, EPAMIG, RU.

b) Corte 2: RU; CIN e CANTINA, DAE, DCC-T, DCC-PP, DEX, EPAMIG.

c) Corte 3: RU; CIN; DCC-T, DCC-PP e CANTINA, DAE, DEX, EPAMIG.

d) Corte 4: RU; CIN; CANTINA; DCC-T, DCC-PP e DAE, DEX, EPAMIG.

e) Corte 5: RU; CIN; CANTINA; EPAMIG; DCC-T, DCC-PP e DAE, DEX.

f) Corte 6: RU; CIN; CANTINA; EPAMIG; DCC-T; DCC-PP e DAE, DEX.

Com o intuito de verificar quais desses cortes seriam plausíveis de serem utilizados, procedeu-se com a análise dos índices de Moran e de Geary, sendo que cada departamento/setor é representado pelo valor médio e as vizinhanças obtidas em cada corte são utilizadas na construção da matriz de vizinhanças.

Na Tabela 1, verifica-se que, considerando o teste de permutação aleatória para o índice de Moran, os cortes 1, 2 e 3 foram considerados pertinentes, pois os índices foram positivos e estatisticamente diferentes de zero, a 5% de significância. Para o índice de Geary estes cortes foram: 2, 3, 4 e 5. Assim, para os dois índices, estes foram os cortes em que existe similaridade entre os vizinhos.

Tabela 1 Índices de Moran e Geary e seus respectivos testes de permutação aleatória (p.a.), para cada um dos cortes feitos no dendrograma, e dois outros cortes que não são obtidos no mesmo

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,0488	0,0411	0,3291	0,0626
Segundo	0,2128	0,0157	0,1700	0,0180
Terceiro	0,3485	0,0487	0,0513	0,0013
Quarto	0,4144	0,0647	0,0339	0,0025
Quinto	0,4570	0,1068	0,0172	0,0129
Sexto	0,9142	0,0537	0,0340	0,1261

Considerando o índice de Moran como critério para determinar o número de grupos, tem-se que o ponto de corte seria o referente ao corte 3, pois dentre os cortes feitos, este obteve o maior valor positivo significativo do índice de Moran $(0,3485)$. Nesse caso os grupos encontrados são: RU; CIN; DCC-T, DCC-PP e CANTINA, DAE, DEX, EPAMIG. Já para o índice de Geary, o ponto de corte com valor significativo mais próximo de 0 (0,0172) seria o corte 5, formando seis grupos. Os grupos formados são: RU; CIN; CANTINA; EPAMIG; DCC-T, DCC-PP e DAE, DEX.

Observa-se que a diferença entre os critérios foi a separação pelo índice de Geary, da CANTINA e da EPAMIG do DAE e do DEX. Tem-se que os dois agrupamentos escolhidos foram coerentes com as observações apresentadas pela Figura 5. O CIN e o RU, realmente têm comportamentos bastante singulares. O DCC-T e DCC-PP têm comportamentos bastante parecidos. A união entre o DAE e o DEX a CANTINA e a EPAMIG fica a cargo do pesquisador/administrador, dependendo somente da questão do rigor quanto a discriminação dos grupos.

A seguir é apresentado o diagrama de espalhamento de Moran. Esse gráfico representa uma forma adicional de visualizar similaridades e dissimilaridades presentes nos dados.

Na Figura 7, encontra-se o diagrama de espalhamento de Moran referente ao terceiro corte no dendrograma da Figura 6. Observa-se que a maioria dos departamentos/setores (75%) encontram-se no segundo quadrante, que é constituído pelas áreas com valores baixos para a variável em análise cercadas por vizinhos que também apresentam baixos valores. Esse resultado está de acordo com o índice de Moran e de Geary apresentado na Tabela 1, que indicaram similaridade significativa entre os vizinhos obtidos pelo terceiro corte no dendrograma. Ademais, a inclinação positiva da reta também comprova a existência de autocorrelação positiva. Além disso, esse foi considerado o melhor corte pelo índice de Moran.

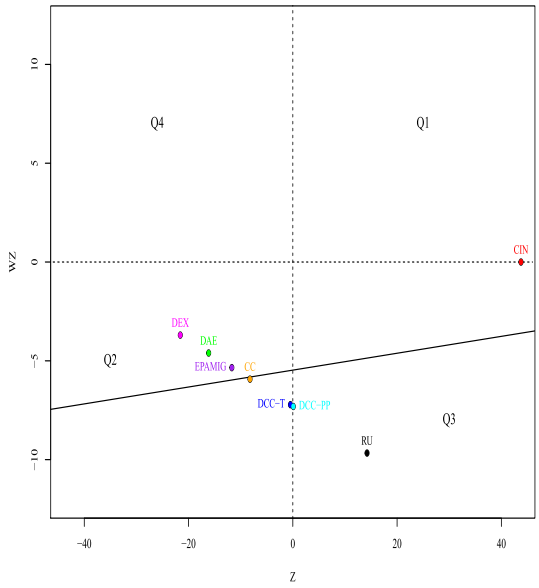

Figura 7 Diagrama de espalhamento de Moran referente ao terceiro corte no dendrograma de alguns setores/departamentos da UFLA

Como ilustração da metodologia proposta, formou-se vizinhanças que não poderiam ser encontradas de acordo com o dendrograma apresentado na Figura 6.

A ideia foi mostrar que grupos muito diferentes dos indicados pelo dendrograma levam a dissimilaridade dos departamentos/setores indicados como vizinhos, ou ausência de dependência. Os agrupamentos propostos foram:

a) Agrupamento 1: CANTINA, DCC-T, DCC-PP, EPAMIG e CIN, DAE, DEX, RU.

b) Agrupamento 2: CANTINA, CIN, DAE, DCC-T e DEX, EPAMIG, DCC-PP, RU.

Observa-se no primeiro agrupamento que o CIN e o RU estão juntos com o DAE e o DEX. Essa formação é incompatível com as características identificadas na Figura 5, enquanto o CIN e o RU possuem altas potências, o primeiro em todo o período analisado, e o segundo em alguns momentos, principalmente no período da manhã, o DAE e o DEX são pequenos consumidores de energia elétrica.

No segundo agrupamento, a incoerência está presente no fato do RU se juntar a locais como DEX, EPAMIG e DCC-PP, que possuem valores da variável analisada distintos ao apresentado por esse setor, em quase todo o período estudado. O mesmo ocorre com a junção do CIN a departamentos/setores como CANTINA, DAE e DCC-T.

Na Tabela 2, encontram-se os índices de Moran e Geary e seus respectivos testes de significância, considerando os grupos propostos.

Tabela 2 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,2435	0,4176	1,0880	0,4102
Agrupamento 2	-0,2548	0,3581	1,0979	0,3667

Observa-se na Tabela 2 que para ambos os casos o índice de Moran foi negativo e não significativamente diferente de zero a 5% de significância, ou seja, existe dissimilaridade entre as vizinhanças, porém esta não foi significativa. Portanto, as vizinhanças estipuladas não levaram a agrupamentos satisfatórios, visto que os vizinhos são independentes. Os valores do índice de Geary, apesar de indicarem dependência negativa, também foram não-significativos, indicando que há uma independência entre as observações vizinhas. Portanto não há uma tendência de áreas vizinhas formarem grupos com seus vizinhos. Esses resultados confirmam o que era esperado para esses agrupamentos.

Em todas as situações analisadas nessa tese, em que se formaram vizinhanças que não poderiam ser encontradas de acordo com o dendrograma, os resultados foram semelhantes aos aqui apresentados, ou seja, existe dissimilaridade entre as vizinhanças, porém esta não foi significativa. Portanto, os resultados seguintes não serão explorados com detalhes. Os agrupamentos, os índices de Moran, Geary e seus respectivos testes foram feitos e estão no (Anexo A).

A seguir, é apresentado na Figura 8 o diagrama de espalhamento de Moran para o agrupamento 1. O diagrama mostra que 50% dos departamentos/setores amostrados encontram-se nos quadrantes três e quatro, indicando dissimilaridade entre as áreas vizinhas com relação à variável em estudo. Veja, por exemplo, que no terceiro quadrante está o CIN, que é uma área com alto valor de potência cercada por áreas com baixos valores, como o DAE e o DEX. Já no caso do quarto quadrante estão os departamentos DAE e DEX, locais de baixos valores de potência cercados por valores altos, como os do CIN e RU. A inclinação negativa da reta também induz à possível existência de autocorrelação negativa.

Observa-se, por meio desses resultados, que os índices de Moran e Geary podem ser uma alternativa para se obter cortes plausíveis no dendrograma, e dentre

estes o que possui melhor qualidade de agrupamento.

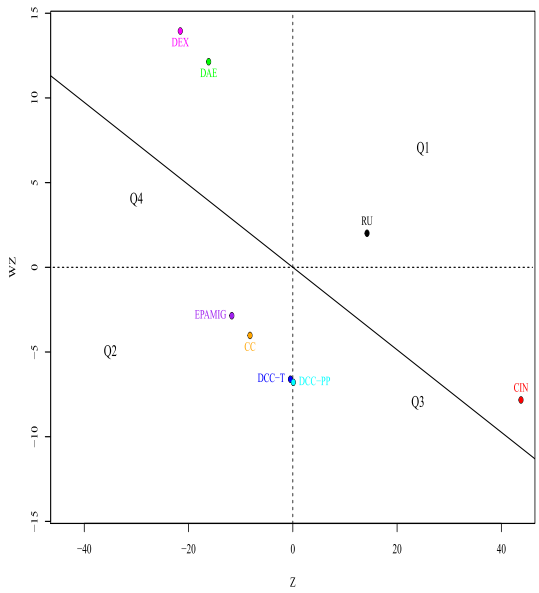

Figura 8 Diagrama de espalhamento de Moran referente ao corte no dendrograma de alguns setores/departamentos da UFLA que originou o Agrupamento 1

Com a intenção de obter informações em períodos importantes para a universidade, fez-se o estudo dos seguintes horários: das 6 às 11:45 horas e das 12 às 17:45 horas. A seguir os resultados dessas análises.

4.1.1.1 Período das 6 às 11h45min

Na Figura 9, estão as séries temporais e os *boxplots* da potência ativa de alguns setores/departamentos da UFLA no período das 6 às 11h45min.

A Figura 9 (b) mostra que no período analisado os departamentos/setores de menor dispersão quanto à potência são a CANTINA, o DEX e EPAMIG, sendo

o DEX a menor variância. As maiores variabilidades são para o DCC-PP e RU, sendo o RU o que possui maior inconstância na medidas de potência.

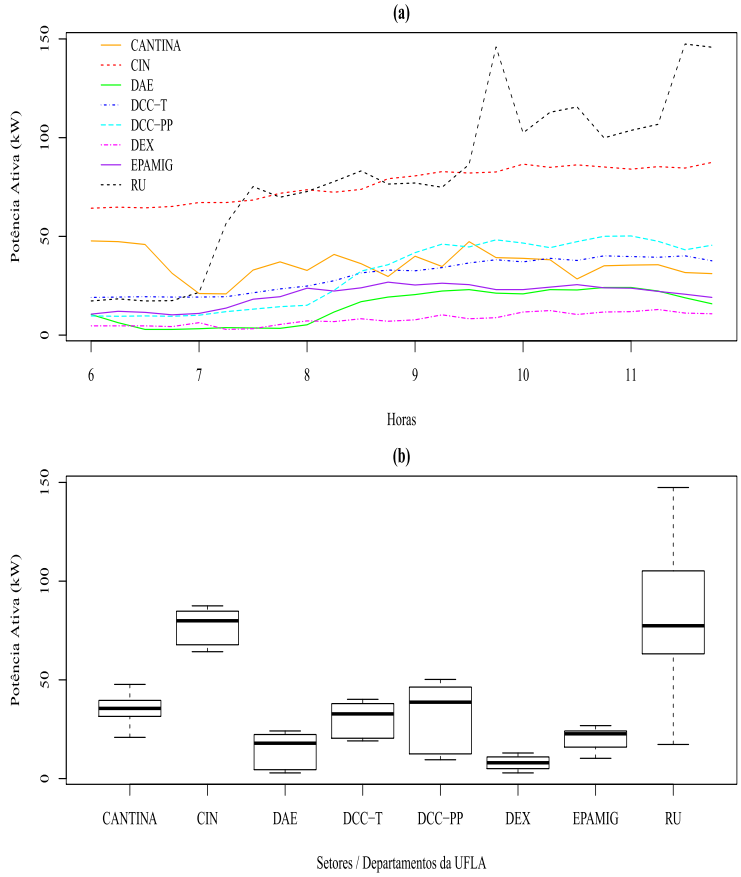

Figura 9 Séries temporais (a) e *boxplots* (b) da potência ativa em alguns setores/departamentos da UFLA, medida em kW, de 15 em 15 minutos no dia 12/08/2010, das 6 às 11h45min

Observa-se neste período que o RU apresenta uma tendência de crescimento na potência ativa medida. Esse comportamento é típico deste setor, pois as suas atividades começam logo no início da manhã, por volta das 7 horas. Na Figura 9 (a), pode-se observar que por volta das 9h30min existe um pico de potên-

cia, pois nesse momento estão sendo realizadas as frituras, em fritadeiras elétricas industriais, de alta potência. Já as 11h30min o pico de potência ocorre em virtude da higienização. Nesse horário, lava-se os pratos e talheres dos primeiros usuários do RU, que começa a servir almoço as 11 horas.

A CANTINA no período das 6 às 7 horas possui valores de potência em um patamar mais alto. Porém, a partir das 7 horas volta a valores mais baixos, comportamento esperado para este setor. No período das 6 às 7 horas, estão sendo feitos os alimentos para se servir aos alunos e funcionários do turno da manhã. A partir das 7 horas não há grandes alterações na potência, principalmente pelo fato da CANTINA utilizar fogão a gás na preparação de outros alimentos. O que pode ser observado de mais relevante são alguns pequenos picos na potência. Esses valores são explicados pela ocorrência de intervalos nas aulas, momentos em que os alunos estão indo até a CANTINA para lanchar. Portanto, liga-se alguns aparelhos elétricos, como sanduicheira, liquidificador, microondas, entre outros.

O DCC-PP e DCC-T mostram um crescimento na potência a partir das 8 horas, mais evidente no DCC-PP. Nesse horário, os alunos e professores estão chegando. Consequentemente, os computadores são ligados nas salas de aula, laboratórios, sala dos professores e setores administrativos.

O CIN tem seus valores de potência em um patamar mais alto durante todo o período indicado, isso ocorre pois esse setor possui equipamentos ligados durante todo o dia, dentre os quais pode-se citar: servidores, computadores, ar-condicionados, entre outros. Observa-se que este setor também tem um crescimento da potência a partir das 8, de baixa magnitude. Esse aumento ocorre pois, além dos aparelhos ligados todo o tempo, nesse horário liga-se também os computadores dos funcionários desse setor, que estão chegando para trabalhar.

Os outros departamentos (DAE, DEX, EPAMIG) têm durante o período

analisado leves mudanças, pois realizam principalmente atividades administrativas e/ou acadêmicas. Nesses setores/departamentos o crescimento da potência no início da manhã também é observado. Veja que para o DAE esse crescimento é maior, pois além de salas de aula e salas de professores, há também laboratórios que usam equipamentos de maior potência. O que não ocorre com a EPAMIG e o DEX. O primeiro tem características administrativas, e o segundo apenas salas de aulas, salas de professores e laboratórios de baixo consumo energético.

As maiores variabilidades de potência ativa estão no DCC-PP e RU, justamente os locais onde o crescimento é mais evidente.

Na Figura 10, tem-se o dendrograma para a variável potência ativa no período das 6 às 11 horas. O coeficiente de correlação cofenética deste dendrograma foi de 0,9375, mostrando uma boa qualidade do agrupamento.

Figura 10 Dendrograma para agrupamento de alguns setores/departamentos da UFLA, obtido a partir de dados da potência ativa da UFLA, medidas em kW, de 15 em 15 minutos no dia 12/08/2010, das 6 às 11h45min

Os cortes que poderiam ser feitos no dendrograma da Figura 10 são:

a) Corte 1: CIN, RU e CANTINA, DCC-T, DCC-PP, DAE, EPAMIG, DEX.

b) Corte 2: CIN; RU e CANTINA, DCC-T, DCC-PP, DAE, EPAMIG, DEX.

c) Corte 3: CIN; RU; CANTINA, DCC-T, DCC-PP e DAE, EPAMIG, DEX.

d) Corte 4: CIN; RU; CANTINA; DCC-T, DCC-PP e DAE, EPAMIG, DEX.

e) Corte 5: CIN; RU; CANTINA; DEX; DCC-T, DCC-PP e DAE, EPAMIG.

f) Corte 6: CIN; RU; CANTINA; DEX; DCC-T; DCC-PP e DAE, EPAMIG.

Por meio da Tabela 3, observa-se que, para o teste de permutação aleatória, o índice de Moran para os cortes 1 e 2 foram estatisticamente diferentes de zero, ao nível de significância de 5%, como eles estão no intervalo de 0 a 1, esta estatística indica que existe similaridade entre os vizinhos formados por este corte. Para o índice de Geary, a significância do teste foi obtida nos cinco primeiros cortes, com todos os valores do índice no intervalo entre 0 e 1, indicando também similaridade.

Dentre os cortes escolhidos, o que obteve maior valor positivo significativo pelo índice de Moran foi o corte 1. Nesse caso, os grupos formados são: CIN, RU e CANTINA, DCC-T, DCC-PP, DAE, EPAMIG, DEX. Para o índice de Geary, o corte que obteve estatística significativa mais próxima de zero foi o corte 5, os grupos neste caso são: CIN; RU; CANTINA; DEX; DCC-T, DCC-PP e DAE,

EPAMIG.

Tabela 3 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,8609	0,0176	0,1217	0,0272
Segundo	0,2645	0,0227	0,1599	0,0170
Terceiro	0,4117	0,0830	0,0311	0,0017
Quarto	0,5012	0,0655	0,0306	0,0061
Quinto	0,3282	0,1667	0,0114	0,0115
Sexto	0,5941	0,1246	0,0223	0,1604

Observa-se que os grupos formados foram bastante coerentes. Veja, por exemplo, que o grupo formado por CANTINA, DCC-T, DCC-PP, DAE, EPAMIG, DEX, inclui departamentos/setores com características muito parecidas, baixos valores de potência, crescimento a partir das 8 horas aproximadamente, portanto, em média esses locais se comportam de maneira similar. Já o CIN e o RU são casos particulares. Assim, devem realmente ficar em um grupo separado. Observa-se também que, esses, na grande maioria do tempo, possuem valores de potência acima dos demais, sendo o CIN com uma certa constância e o RU com tendência de crescimento.

A escolha entre o agrupamento formado pelos índices de Moran ou Geary, fica a cargo do pesquisador/administrador, ambas as situações são razoáveis. A diferença é que o índice de Geary separa mais os departamentos/setores, ou seja, produz um maior número de grupos.

Foi feito o diagrama de espalhamento de Moran referente ao primeiro corte no dendrograma da Figura 10. Observa-se que 100% dos departamentos/setores encontram-se nos quadrantes 1 e 2. Assim, há indícios fortes de similaridade entre os vizinhos obtidos pelo primeiro corte no dendrograma. No primeiro quadrante,

estão o CIN e o RU, áreas com valores altos para a variável em análise. Já no segundo quadrante, estão as áreas com valores baixos para a variável em análise cercadas por vizinhos que também apresentam baixos valores. Esses resultados corroboram as afirmações feitas com os índices de Moran e Geary. Tem-se que a inclinação positiva da reta também comprova a existência de autocorrelação positiva.

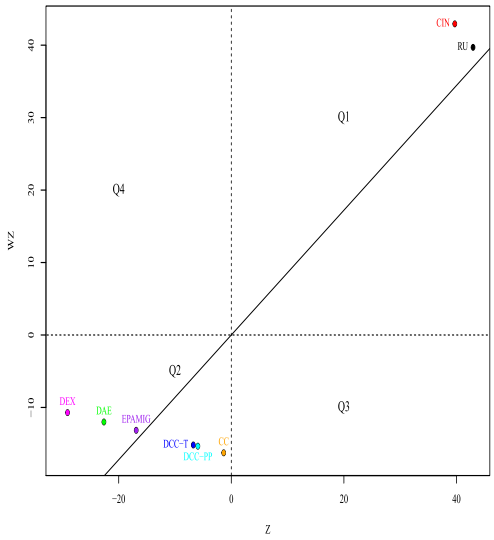

Figura 11 Diagrama de espalhamento de Moran referente ao primeiro corte no dendrograma de alguns setores/departamentos da UFLA (período das 6 às 11h45min)

4.1.1.2 Período das 12 às 17h45min

Na Figura 12, têm-se as séries temporais e os *boxplots* da potência ativa de alguns setores/departamentos da UFLA no período das 12 às 17h45min.

Observa-se neste período que o RU têm seus valores de potência ativa

altos até por volta das 13 horas, caindo seus valores a partir daí. Isso é reflexo do fim do período do almoço. Existe um pico de potência às 13, que é a hora em que os funcionários fazem a higienização do ambiente. Justamente devido a esse momento de transição, do início da tarde, que a maior variabilidade dos dados de potência ativa neste período está associada ao RU. Isso pode ser visualizado pelo seu *boxplot* na Figura 12 (b).

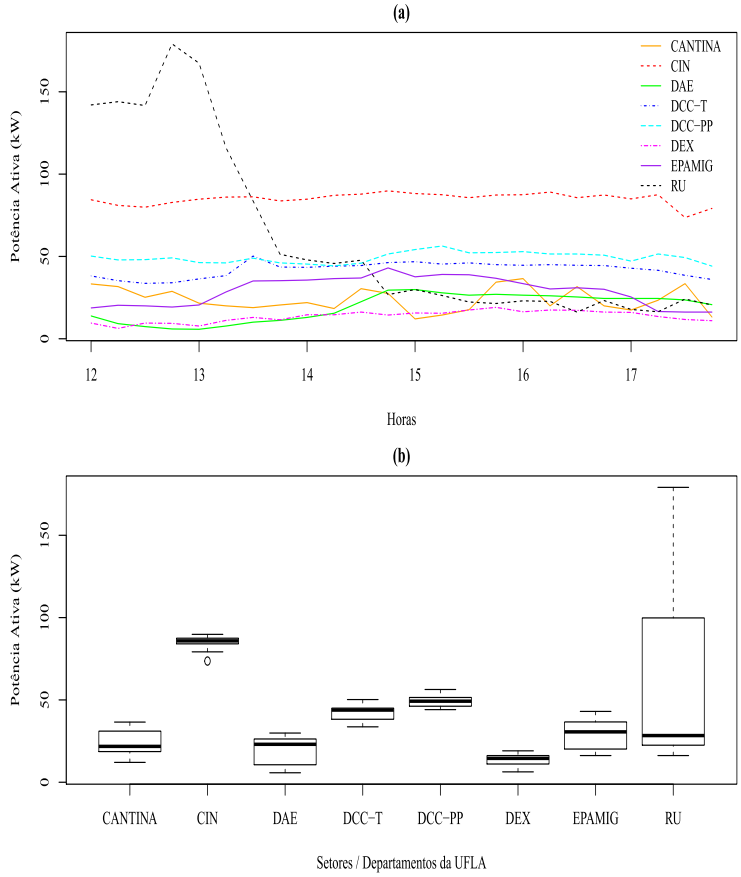

Figura 12 Séries temporais (a) e *boxplots* (b) da potência ativa em alguns setores/departamentos da UFLA, medida em kW, de 15 em 15 minutos no dia 12/08/2010, das 12 às 17:45 horas

A Figura 12 (a) mostra que o CIN tem seus valores praticamente estáveis, em torno de 90 kW. Esse comportamento pode ser observado em praticamente todo o período estudado, já que os aparelhos utilizados nesse setor têm uso contínuo. Essa estabilidade é observada também na Figura 12 (b). O *boxplot* para o CIN mostra a baixa variabilidade dos dados de potência para esse setor. Porém, nota-se também a presença de um *outlier*, que foi ocasionado por um valor baixo da variável analisada por volta das 17h30min. Este horário coincide com o fim das atividades dos funcionários deste local, portanto, esta pode ser a causa da queda da demanda de potência nesse local.

Os outros departamentos/setores têm seus comportamentos que se assemelham desde o início da manhã, ou seja, um padrão muito bem definido, com potências abaixo de 55 kW, sendo que neste grupo o DCC-T possui as maiores potências medidas e o DEX os menores valores. A CANTINA também continua com seu comportamento típico, com algumas pequenas flutuações ao longo do tempo, fruto de intervalos durante as aulas.

O dendrograma para a variável potência ativa no período das 12h10min às 17h45min encontra-se na Figura 13. O coeficiente de correlação cofenética deste dendrograma foi de 0,9442, indicando uma boa qualidade do agrupamento.

Têm-se que os cortes que poderiam ser feitos no dendrograma da Figura 13 são:

a) Corte 1: RU e CIN, CANTINA, DAE, DEX, EPAMIG, DCC-T, DCC-PP.

b) Corte 2: CIN; RU; CANTINA, DAE, DEX, EPAMIG, DCC-T, DCC-PP.

c) Corte 3: CIN; RU; DCC-T, DCC-PP e CANTINA, DAE, DEX, EPA-

MIG.

d) Corte 4: CIN; RU; EPAMIG; DCC-T, DCC-PP e CANTINA, DAE, DEX.

e) Corte 5: CIN; RU; EPAMIG; CANTINA; DCC-T, DCC-PP e DAE, DEX.

f) Corte 6: CIN; RU; EPAMIG, CANTINA; DCC-T; DCC-PP e DAE, DEX.

Figura 13 Dendrograma para agrupamento de alguns setores/departamentos da UFLA, obtido a partir de dados da potência ativa da UFLA, medidas em kW, de 15 em 15 minutos no dia 12/08/2010, das 12 às 17h45min

Na Tabela 4, observa-se que para o teste de permutação aleatória, o índice de Moran foi estatisticamente diferente de zero ao nível de significância de 5% somente no corte 2. Como o valor do índice é positivo, há o indicativo de que exista similaridade entre os vizinhos formados por esse corte. Para o índice de

Geary, a significância do teste foi obtida nos cortes 2, 3, 4 e 5, com todos os valores do índice entre 0 e 1, indicando também similaridade.

Tabela 4 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	-0,1516	0,5719	1,0307	0,5585
Segundo	0,1696	0,0345	0,3252	0,0171
Terceiro	0,4653	0,0695	0,0665	0,0011
Quarto	0,5484	0,0783	0,0458	0,0029
Quinto	0,5755	0,1320	0,0361	0,0166
Sexto	1,1214	0,0528	0,0266	0,0925

Para o índice de Moran, como somente um corte foi significativo, este foi o escolhido. Assim, tem-se que o corte 2 representa a melhor escolha, sendo os grupos dados por: CIN; RU; CANTINA, DAE, DEX, EPAMIG, DCC-T, DCC-PP. Para o índice de Geary, o corte que obteve estatística significativa mais próxima de zero foi o corte 5. Os grupos associados a esse corte são: CIN; RU; EPAMIG; CANTINA; DCC-T, DCC-PP e DAE, DEX.

Observa-se que os grupos formados pelo índice de Moran separam de forma satisfatória os departamentos/setores que foram considerados diferentes pela análise exploratória da Figura 12. O RU fica separado em um grupo, principalmente em virtude do período das 12 às 14 horas, momentos em que o setor esta finalizando as atividades do almoço. Após as 14 horas, o RU se aproxima do consumo de energia elétrica dos demais departamentos/setores, com exceção do CIN, que é diferente dos demais todo o tempo, sendo sua característica principal a constância no consumo de energia. Consequentemente o CIN também fica em um grupo separado dos demais. Todos os outros departamentos/setores ficam em um mesmo grupo, sendo que estes possuem valores de potência abaixo de 55 kW em

todo o período estudado.

Para o índice de Geary, houve uma separação maior dos departamentos/setores, observa-se que o DCC-T e o DCC-PP se juntaram em um grupo, o DAE e o DEX se uniram em outro, já todos os outros locais se mostraram distintos uns dos outros. Como pode ser observado na Figura 12, esse agrupamento também é pertinente.

Foi feito o diagrama de espalhamento de Moran referente ao segundo corte no dendrograma da Figura 14. Observa-se que 50% dos departamentos/setores encontram-se no segundo quadrante, ou seja, são locais que apresentam valores baixos para a variável potência. O CIN e o RU estão sobre o eixo das abscissas, visto que estes foram considerados sem nenhuma vizinhança, o DCC-T e o DCC-PP encontram-se no terceiro quadrante, ou seja, locais com altos valores de potência que se encontram agrupados com locais de baixos valores. No geral, esse diagrama mostra que realmente o segundo corte no dendrograma é razoável de ser realizado. A inclinação positiva da reta também comprova a existência de autocorrelação positiva.

Nos próximos tópicos, serão apresentadas as análises feitas para comparações de dias da semana segundo a variável potência ativa, comparações dos meses considerando a demanda de potência em horário de ponta e horário fora de ponta. Para os próximos tópicos, serão apresentados os resultados de forma mais resumida, considerando que as metodologias utilizadas são semelhantes às já apresentadas para o estudo dos departamentos. Os níveis de confiança considerados serão sempre 5%. A metodologia para selecionar o corte no dendrograma também será a mesma. Para o índice de Moran, o maior valor positivo e significativo será utilizado para se fazer essa escolha. Enquanto que para o índice de Geary, o grupo obtido será aquele que obtiver valor significativo mais próximo de zero.

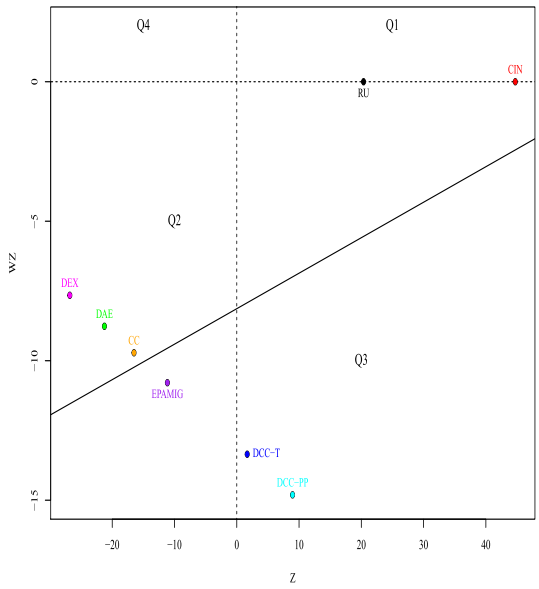

Figura 14 Diagrama de espalhamento de Moran referente ao segundo corte no dendrograma de alguns setores/departamentos da UFLA (período das 12 às 17h45min)

4.1.2 Dias da semana

Na Figura 15 (a), têm-se as séries temporais da potência ativa da UFLA, no período de 10/06 a 16/06/2013. O intuito com esta coleta de dados foi comparar os dias da semana com relação à variável potência ativa.

Tem-se que no período das 0 até as 7 horas, aproximadamente, a potência ativa se mantém em torno de 600 kW para todos os dias estudados. Isso é esperado, pois a universidade neste período tem suas atividades reduzidas, ficando ligados somente alguns equipamentos de uso contínuo, como por exemplo, os servidores no CIN e alguns aparelhos nos laboratórios.

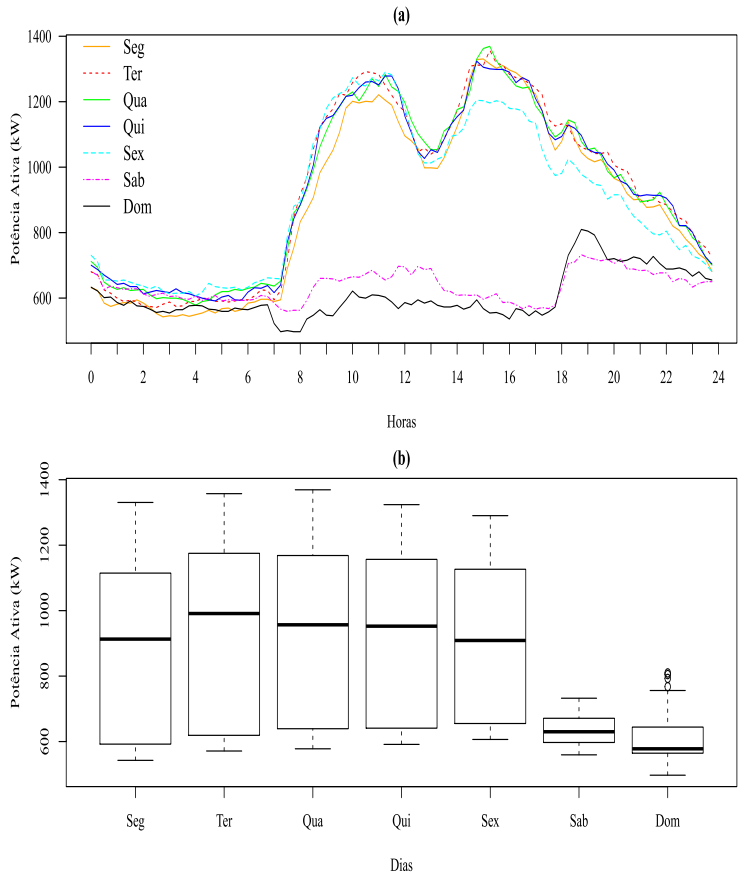

Figura 15 Séries temporais (a) e *boxplots* (b) da potência ativa da UFLA, medida em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013

Observa-se visualmente que não há uma diferença relevante de segunda a sexta-feira, mas no final de semana, como era de se esperar, há um decréscimo acentuado, pois existem poucas atividades no campus.

Na Figura 15 (a), vê-se também um padrão típico durante os dias da semana. Por volta das 12 horas há uma queda dos valores de potência, devido à ausência de aulas neste período e também pelo fato de ser horário de almoço dos funcionários e professores.

Observa-se também que no período da manhã, a segunda-feira possui valores ligeiramente inferiores aos demais dias. Provavelmente, isso é reflexo de que neste período as atividades acadêmicas não estão ainda todas sendo realizadas. Fato semelhante ocorre na sexta-feira, porém os valores de potência são inferiores no período da tarde, também pelo efeito da redução das atividades acadêmicas neste período, devido à ausência de alguns alunos que não se encontram mais no campus.

Uma menor variabilidade dos dados é encontrada no sábado e domingo (Figura 15b), comportamento esperado, visto que nesses dias a universidade tem suas atividades bastante reduzidas. Pode-se observar também que o sábado tem alguns horários em que a potência supera os valores do domingo. Isso decorre do fato de que no sábado há algumas atividades acadêmicas.

Outra situação relevante, que se pode observar na Figura 15(a), é que no domingo, a partir das 18 horas, há um acréscimo nos valores de potência, que chega a superar os valores de sábado no mesmo período. Isso pode ser explicado pela presença de alguns alunos que estão voltando para a universidade, especificamente para o alojamento estudantil.

Para confirmar o agrupamento obtido pela análise visual, procedeu-se com a AA obtendo-se o dendrograma (Figura 16). Verifica-se por meio deste que realmente dois grupos podem ser formados, um com os dias de segunda a sexta-feira e outro com os dias do fim de semana. Observa-se também que a amplitude nas distâncias de junção desses dois grupos formados é alta, mostrando assim que os grupos diferem bastante. O coeficiente de correlação cofenética aqui também foi alto (0,9959), indicando assim boa qualidade no agrupamento.

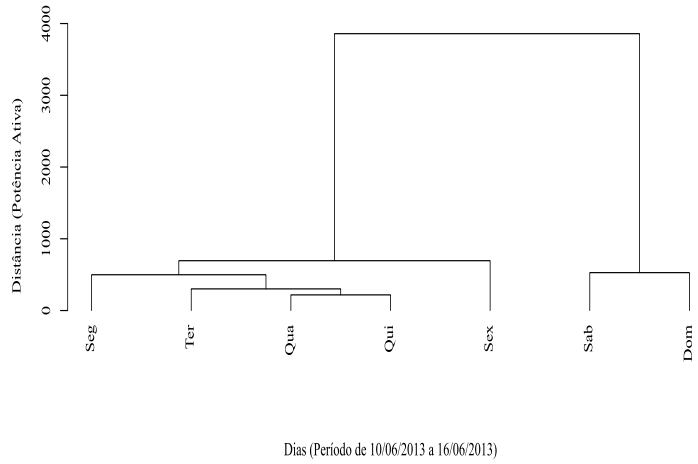

Figura 16 Dendrograma para o agrupamento dos dias da semana, obtido a partir de dados de potência ativa, medidas em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013

Os cortes que poderiam ser feitos no dendrograma da Figura 16 são:

a) Corte 1: Sab, Dom e Seg, Ter, Qua, Qui, Sex.

b) Corte 2: Sex; Sab, Dom e Seg, Ter, Qua, Qui.

c) Corte 3: Sex; Sab; Dom e Seg, Ter, Qua, Qui.

d) Corte 4: Sex; Sab; Dom; Seg e Ter, Qua, Qui.

e) Corte 5: Sex; Sab; Dom; Seg; Ter e Qua, Qui.

Na Tabela 5, observa-se que, para o índice de Moran, o corte escolhido seria o segundo. Os grupos formados por este são: Sex; Sab, Dom e Seg, Ter, Qua, Qui. Já para o índice de Geary o corte 4 foi considerado o melhor, sendo que os grupos formados são: Sex; Sab; Dom; Seg e Ter, Qua, Qui.

Tem-se que ambos os índices levaram a agrupamentos satisfatórios. Aqui também o índice de Geary levou ao maior número de grupos.

Observa-se que o índice de Moran encontrou três grupos. O grupo que contém o sábado e domingo juntos já era esperado pela análise visual dos dados. Para os demais dias da semana, esperava-se que ficassem em um outro grupo, porém juntos, mas o que ocorreu foi a separação da sexta dos demais. Esse fato também é compreensível, pois na sexta, a partir das 14 horas, aproximadamente, há uma queda no consumo de energia. Isso ocorre, como já mencionado, pela ausência de alguns alunos na universidade, que voltam para suas cidades para o fim de semana.

Os grupos formados pelo índice de Geary mostram algumas situações não obtidas pelo índice de Moran. Observa-se, por exemplo, que a separação entre sábado e o domingo pode ser explicada pela diferença de demanda de potência destes em grande parte do dia. O fato da segunda-feira ficar em um grupo sem outros elementos pode ser justificado pelo período da manhã, que diverge dos demais dias da semana.

Tabela 5 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,9797	0,0245	0,0174	0,0234
Segundo	1,1005	0,0153	0,0183	0,0121
Terceiro	0,4244	0,0439	0,0174	0,0404
Quarto	0,5281	0,0145	0,0000	0,0153
Quinto	0,5305	0,0735	0,0000	0,0675

Nos tópicos a seguir também realizaram-se partições. Estas foram feitas nos dias 10/06 a 16/06/2013, considerando os mesmos períodos enumerados anteriormente.

4.1.2.1 Período das 6 às 11h45min

Na Figura 17, têm-se as séries temporais e os *boxplots* da potência ativa dos dias da semana na UFLA, no período das 6 às 11h45min, considerando os dias 10/06 a 16/06/2013.

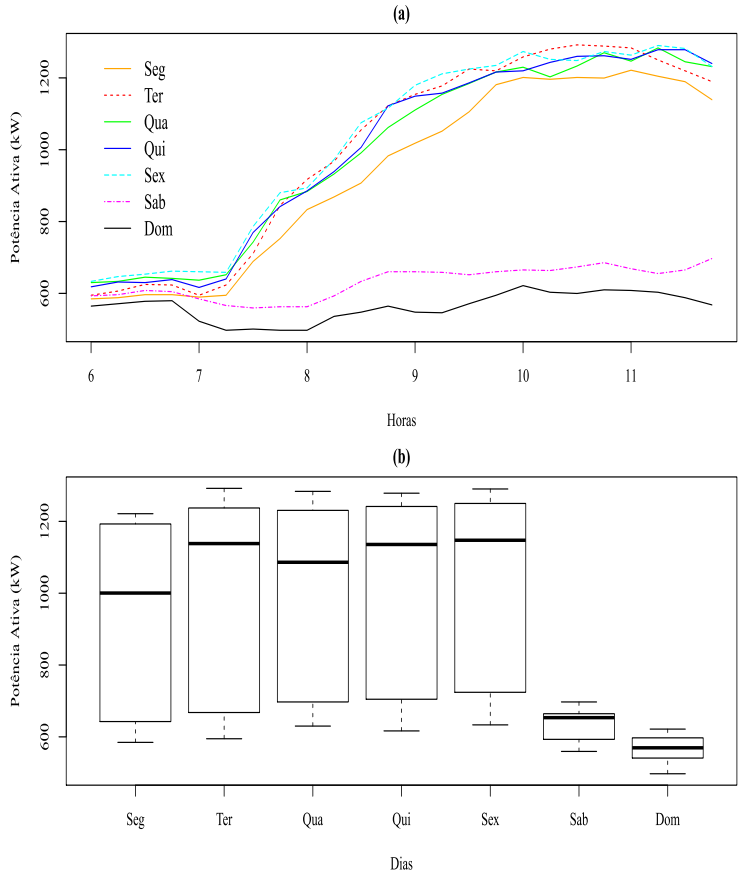

Figura 17 Séries temporais (a) e *boxplots* (b) da potência ativa da UFLA, medida em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013, das 6 às 11h45min

Observa-se que para todos os dias o comportamento é semelhante até por volta das 7 horas. A partir daí é nítido o crescimento dos valores de potência no

período de segunda à sexta, com uma tendência linear até as 10 horas, mantendo-se constante desse ponto em diante. A estabilização ocorre em torno de 1200 kW.

Visualiza-se pela série temporal da segunda-feira, que esse dia possui valores de potência inferiores aos demais, exceto sábado e domingo, que têm valores abaixo de todos os outros. Isto ocorre devido à ausência de alguns alunos na universidade na segunda no período da manhã, muitos destes moram em outras cidades e ainda não chegaram. Esse fato acarreta várias alterações no consumo de energia elétrica, pois menos refeições são feitas no RU, o alojamento dos estudantes não está com sua lotação real, etc.

O sábado e o domingo não têm variações relevantes de potência durante o período analisado. Isso pode ser analisado nos *boxplots* da Figura 17 (b). O que se observa nas séries temporais da Figura 17 (a) é que o sábado possui valores de potência ligeiramente maiores do que o domingo. Esse fato ocorre devido à ocorrência de algumas aulas durante o intervalo de tempo estudado.

De acordo com os *boxplots* e as séries temporais da Figura 17, dois grupos podem ser formados, um com os dias do fim de semana e outro com os demais dias. Observa-se pelos boxplots que os dias do fim de semana possuem medidas de posição e dispersão da potência visualmente menores do que dos demais dias.

Na Figura 18, tem-se o dendrograma para os dias da semana, considerando o período das 6 às 11h45min. O coeficiente de correlação cofenética aqui também foi alto (0,9891), indicando assim boa qualidade no agrupamento.

Os cortes que podem ser feitos no dendrograma da Figura 18 são:

a) Corte 1: Sab, Dom e Seg, Ter, Qua, Qui, Sex.

b) Corte 2: Seg; Sab, Dom e Ter, Qua, Qui, Sex.

c) Corte 3: Seg; Sab; Dom e Ter, Qua, Qui, Sex.

d) Corte 4: Seg; Sab; Dom; Ter e Qua, Qui, Sex.

e) Corte 5: Seg; Sab; Dom; Ter; Sex e Qua, Qui.

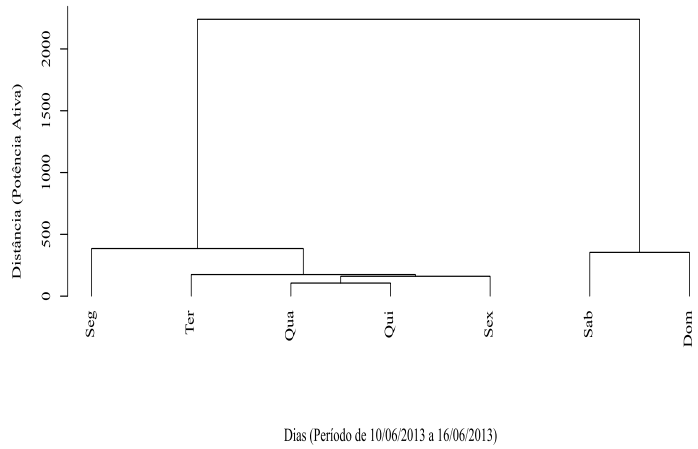

Figura 18 Dendrograma para o agrupamento dos dias da semana, obtido a partir de dados de potência ativa, medidas em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013, das 6 às 11h45min

A Tabela 6 mostra que, tanto para o índice de Moran quanto para o índice de Geary, o corte no dendrograma seria o segundo. Os grupos formados seriam: Seg; Sab, Dom e Ter, Qua, Qui, Sex.

Observa-se que os grupos formados retratam de forma satisfatória a separação dos dias, segundo a variável potência. Tem-se sábado e domingo juntos, o que é esperado para a variável analisada. A segunda-feira ficou isolada em um grupo, isso também é razoável, pois foi visto que no período observado, esse dia possui valores de potência abaixo dos demais dias úteis. Já os outros dias ficaram em um único grupo, isso também é pertinente considerando a Figura 17 (a).

Tabela 6 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,9566	0,0259	0,0372	0,0286
Segundo	1,1472	0,0063	0,0026	0,0158
Terceiro	0,4898	0,0137	0,0039	0,0161
Quarto	0,4950	0,0652	0,0057	0,1005
Quinto	0,4341	0,3210	0,0006	0,0657

4.1.2.2 Período das 12 às 17h45min

Na Figura 19, têm-se as séries temporais e os *boxplots* da potência ativa dos dias da semana na UFLA, no período das 12 às 17h45min, considerando os dias 10/06 a 16/06/2013.

Tem-se que neste período o sábado e o domingo também possuem valores de potência abaixo dos outros dias. Observa-se que o sábado permanece superior ao domingo, porém diminuindo a diferença à medida que se aproxima do fim da tarde.

Tem-se que às 13 horas há um leve decréscimo dos valores de potência, considerando os dias de segunda à sexta. Logo por volta das 14 horas, os valores aumentam, chegando a 1300 kW aproximadamente, e permanecem aproximadamente constantes até em torno das 17 horas, momento que coincide com o fim das atividades do período vespertino. Consequentemente, há uma queda nos valores de potência.

Pela Figura 19 (a), visualiza-se também que na sexta-feira há uma queda mais acentuada na potência, que faz com que esse dia se torne diferente dos demais dias úteis. Essa queda na potência é esperada, pois como já dito anteriormente a sexta é um dia atípico, já que tem suas atividades um pouco comprometidas no

período da tarde, pela ausência de alguns alunos no campus.

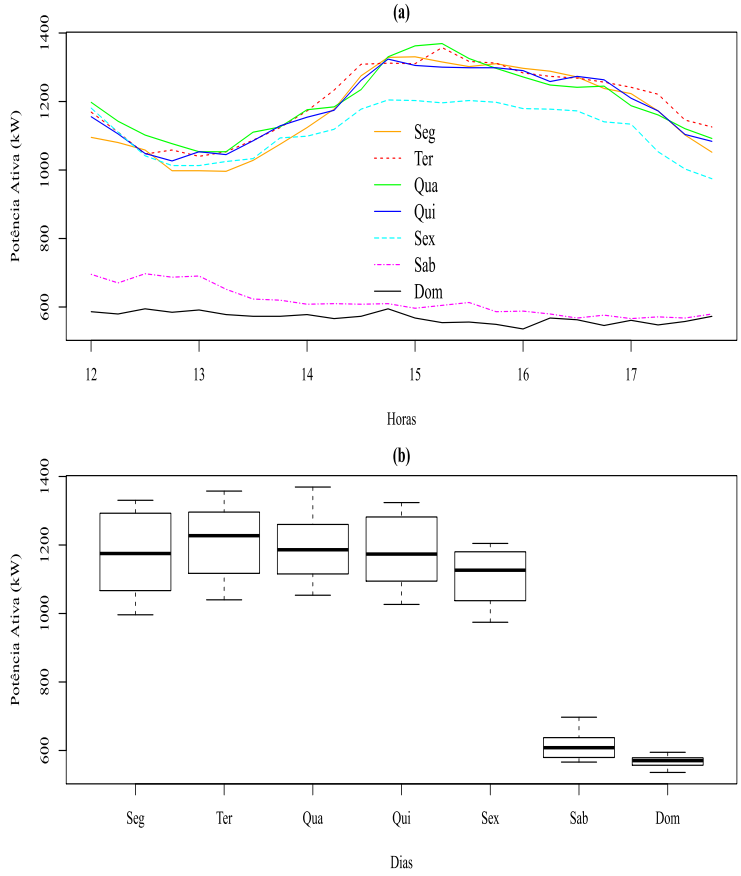

Figura 19 Séries temporais (a) e *boxplots* (b) da potência ativa da UFLA, medida em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013, das 12 às 17h45min

Conforme a Figura 19, dois grupos podem ser formados, um com os dias do fim de semana e outro com os demais dias. Pela Figura 19 (b), fica nítida a diferença das medidas de variabilidade e posição dos dados de potência associados aos dias de semana e fim de semana. Observa-se que os valores médios de potência para os dias durante a semana ficam em torno de 1200 kW, enquanto que o sábado

e domingo possuem em média uma demanda de 550 kW. Visualiza-se também que os dias durante a semana são mais dispersos quanto à variável analisada do que o sábado e o domingo. Isso é esperado, pois no fim de semana tem-se poucas atividades na universidade, ou seja, uma menor demanda de potência por quase todo o período do dia.

O dendrograma para os dias da semana, considerando o período das 12 às 17h45min, é apresentado na Figura 20. O coeficiente de correlação cofenética aqui também foi alto (0,9949), indicando assim boa qualidade no agrupamento.

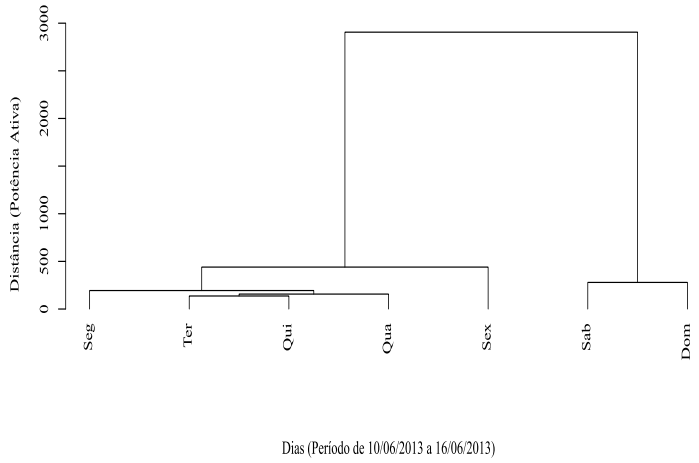

Figura 20 Dendrograma para o agrupamento dos dias da semana, obtido a partir de dados de potência ativa, medidas em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013, das 12 às 17h45min

No dendrograma da Figura 20, os cortes feitos são:

a) Corte 1: Sab, Dom e Seg, Ter, Qua, Qui, Sex.

b) Corte 2: Sex; Sab, Dom e Seg, Ter, Qua, Qui.

c) Corte 3: Sex; Sab; Dom e Seg, Ter, Qua, Qui.

d) Corte 4: Sex; Sab; Dom; Seg e Ter, Qua, Qui.

e) Corte 5: Sex; Sab; Dom; Seg; Qua e Ter, Qui.

A Tabela 7 mostra que para o índice de Moran o melhor corte é o segundo. Os grupos formados neste caso são: Sex; Sab, Dom e Seg, Ter, Qua, Qui. Para o índice de Geary o corte de melhor qualidade foi o corte 4, sendo que os grupos formados são: Sex; Sab; Dom; Seg e Ter, Qua, Qui. Os agrupamentos são bastante pertinentes. Observa-se que em ambos, terça, quarta e quinta-feira, não se separam. Este fato foi observado em todas as situações analisadas, mostrando que esses dias tendem a ser similares em todo o período observado quanto a variável analisada.

Tabela 7 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,9831	0,0238	0,0145	0,0254
Segundo	1,1327	0,0058	0,0058	0,0046
Terceiro	0,46815	0,0150	0,0020	0,0123
Quarto	0,49656	0,0132	0,0009	0,0168
Quinto	0,48963	0,1076	0,0018	0,1685

Para os dias da semana observou-se também que os agrupamentos “ótimos” obtidos, considerando tanto as séries originais quanto as partições, não são os mesmos. A justificativa para esse fato é a mesma apresentada para os departamentos/setores, o comportamento da variável em estudo se altera ao longo do dia.

4.1.3 Meses - demanda de potência registrada em horário de ponta (DPRHP)

Na Figura 21, têm-se as séries temporais e os *boxplots* da demanda de potência registrada em horário de ponta da UFLA, no período de janeiro de 1995 a dezembro de 2013.

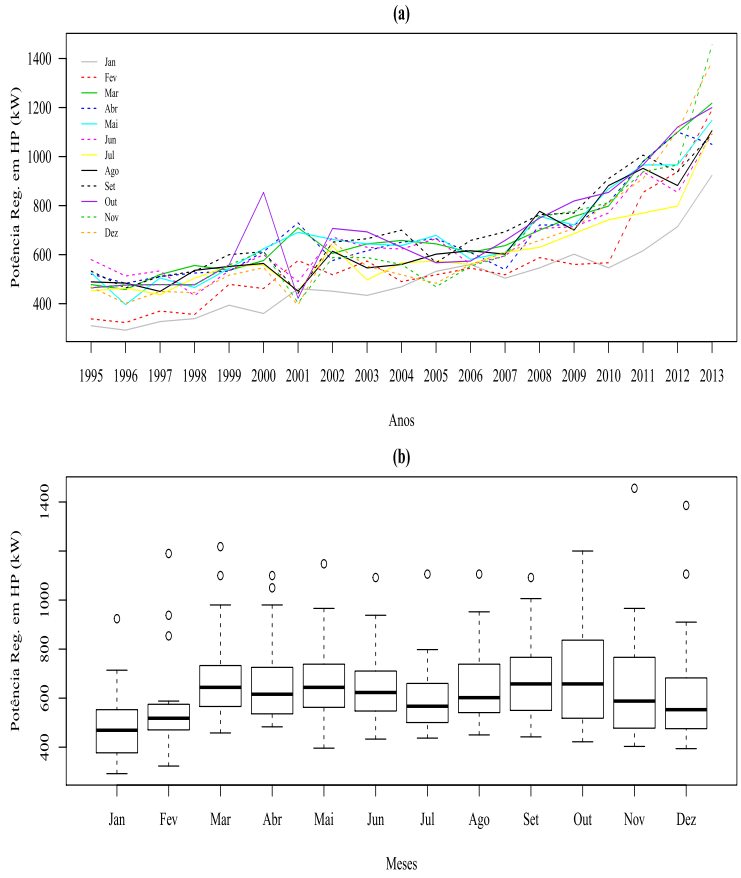

Figura 21 Séries mensais (a) e boxplots (b) da demanda de potência registrada em horário de ponta da UFLA, medida em kW, no período de janeiro de 1995 a dezembro de 2013

Observa-se pela Figura 21 que, nos meses de janeiro e fevereiro, há um menor registro de potência em horário de ponta. Isso pode ser explicado pelo fato

de que nesses meses a universidade normalmente está em período de férias.

Visualmente, percebe-se que há uma tendência de crescimento para todos os meses, o que é reflexo do avanço da UFLA durante estes anos analisados, principalmente a partir de 2010. Mesmo janeiro e fevereiro que são meses de menor demanda de potência, tiveram crescimento, principalmente devido ao uso cada vez mais constante de ares-condicionados. Esse uso é acarretado pelas altas temperaturas ocorridas nesses meses.

No horário de ponta da energia elétrica, que compreende o período noturno, houve um avanço considerável na universidade, pois novos cursos foram instalados, gerando assim reflexos no setor elétrico. Veja que no ano de 1995 a demanda de potência estava em torno de 400 kW e no ano de 2013 esse valor é de cerca de 1100 kW. A perspectiva é de mais crescimento nos próximos anos, já que novos prédios estão em construção e, portanto, mais consumo de energia elétrica.

Na Figura 21 (a), percebe-se que há um pico de potência no mês de outubro de 2000, este valor foi ocasionado por um evento ocorrido na universidade, chamado Rodeio Universitário. Esse evento fez com que a potência chegasse a um valor anormal para a época, porém não contribuiu para o aumento do consumo de energia elétrica, já que esta potência atingida foi momentânea, não perdurando por um tempo maior.

Outro fato importante ocorrido nesse período foi o racionamento de energia elétrica, promovido pelo governo federal, no ano de 2001. Observa-se que essa medida ocasionou uma queda na potência registrada em vários meses do referido ano. O período afetado compreende os meses de junho a dezembro de 2001. Houve também nesse mesmo ano uma greve nas universidades federais, inclusive na UFLA. A greve dos professores começou no fim do mês de agosto e terminou ao final do mês de novembro, com duração de quase 100 dias.

Por meio dos *boxplots* da Figura 21 (b), observa-se um comportamento sazonal, o que é esperado, pois em julho, dezembro, janeiro e fevereiro, os valores da variável em análise são realmente mais baixos, considerando que esses meses normalmente são as férias dos alunos. Visualiza-se também a presença de *outliers* em quase todos os meses analisados, somente outubro não teve a presença de valores como esses. Em todas as situações os *outliers* são valores de potência que estão na parte superior dos *boxplots*. Isso pode indicar medidas excessivas de demanda registrada de potência, podendo ser causadores de problemas para a universidade.

O dendrograma dos meses, para a variável analisada, considerando o período de janeiro de 1995 a dezembro de 2013, é apresentado na Figura 22. O coeficiente de correlação cofenética aqui também foi alto (0,8386), indicando assim boa qualidade no agrupamento.

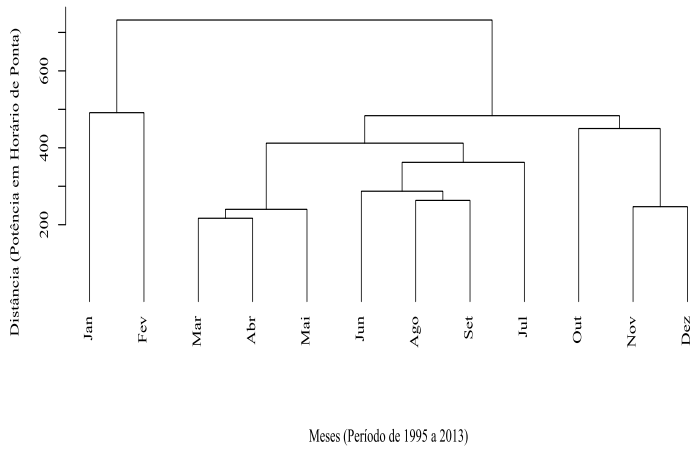

Figura 22 Dendrograma para agrupamento dos meses do ano obtido a partir de dados da demanda de potência registrada em horário de ponta da UFLA, medidas em kW, no período de janeiro de 1995 a dezembro de 2013

No dendrograma da Figura 22, os cortes feitos são:

a) Corte 1: jan, fev e mar, abr, mai, jun, jul, ago, set, out, nov, dez.

b) Corte 2: jan; fev e mar, abr, mai, jun, jul, ago, set, out, nov, dez.

c) Corte 3: jan; fev; out, nov, dez e mar, abr, mai, jun, jul, ago, set.

d) Corte 4: jan; fev; out; nov, dez e mar, abr, mai, jun, jul, ago, set.

e) Corte 5: jan; fev; out; nov, dez; mar, abr, mai e jun, jul, ago, set.

f) Corte 6: jan; fev; out; jul; nov, dez; mar, abr, mai e jun, ago, set.

g) Corte 7: jan; fev; out; jul; jun; nov, dez; ago, set e mar, abr, mai.

h) Corte 8: jan; fev; out; jul; jun; ago; set; nov, dez e mar, abr, mai.

i) Corte 9: jan; fev; out; jul; jun; ago; set; nov; dez e mar, abr, mai.

j) Corte 10: jan; fev; out; jul; jun; ago; set; nov; dez; mai e mar, abr.

A Tabela 8 mostra que, para o índice de Moran, o ponto no dendrograma a se realizar o corte é o primeiro, cujos grupos são: jan, fev e mar, abr, mai, jun, jul, ago, set, out, nov, dez. Já para o índice de Geary o corte selecionado é o nono, em que os grupos formados são: jan; fev; out; jul; jun; ago; set; nov; dez e mar, abr, mai.

As separações apresentadas pelos dois cortes, apesar de bastante diferentes, são realmente pertinentes. No caso do corte 1, tem-se a formação de dois grupos, um contendo os meses de janeiro e fevereiro, e o outro com todos os outros meses. De fato os meses de janeiro e fevereiro se comportam de maneira bastante peculiar. Estes geralmente apresentam valores de demanda de potência

inferiores aos demais meses, já que na universidade normalmente não ocorrem atividades acadêmicas. No corte 9, há uma separação maior dos meses, observa-se que somente os meses de março, abril e maio se juntaram em um único grupo. Estes verdadeiramente são bastante próximos quanto à demanda de potência.

Tabela 8 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,6780	0,0074	0,2952	0,0074
Segundo	0,1294	0,0086	0,2112	0,0078
Terceiro	0,1010	0,1096	0,2373	0,0154
Quarto	0,0886	0,1219	0,1966	0,0204
Quinto	0,1467	0,1606	0,1434	0,0101
Sexto	0,2339	0,1579	0,0596	0,0019
Sétimo	0,2489	0,1513	0,0771	0,0143
Oitavo	0,3112	0,1242	0,0256	0,0125
Nono	0,5283	0,0663	0,0112	0,0250
Décimo	0,5574	0,1461	0,0193	0,1765

Outros agrupamentos são razoáveis na separação dos meses segundo a DPRHP, como por exemplo, o obtido pelo corte 3, que gerou os seguintes grupos: jan; fev; out, nov, dez e mar, abr, mai, jun, jul, ago, set. A Tabela 8 mostra que este corte foi significativo pelo teste de Geary. Em comparação com o corte 1, ocorreram algumas mudanças. Houve a separação de janeiro e fevereiro, e a formação de um novo grupo, composto por outubro, novembro e dezembro. Essa separação é interessante, pois observa-se que nos meses de outubro, novembro e dezembro, que devido às temperaturas mais elevadas dos dias, há um gasto maior de energia elétrica, proporcionado por ares-condicionados, ventiladores, entre outros. O mesmo não ocorre em janeiro e fevereiro, que apesar das temperaturas do ar estarem elevadas, a demanda não se altera proporcionalmente, pois esses meses, como já mencionado, são de férias dos alunos.

A seguir fez-se uma análise restrita aos anos de 2010 a 2013. O objetivo foi obter informações de um período mais recente, visto que a UFLA passa por diversas mudanças nos últimos anos. Assim, dados mais atualizados retratam de forma mais fidedigna a realidade atual da universidade. Esse estudo foi feito tanto para a variável demanda de potência em horário de ponta, quanto para a mesma variável em horário fora de ponta.

4.1.3.1 Período de janeiro de 2010 a dezembro de 2013

Na Figura 23, encontram-se as séries temporais e os *boxplots* da demanda de potência registrada em horário de ponta da UFLA, no período de janeiro de 2010 a dezembro de 2013.

Observa-se que há uma tendência de crescimento da demanda neste período. Em 2012, percebe-se uma queda da variável analisada em alguns meses, devido a uma greve dos professores da universidade que durou quase 4 meses. Esse fato ocasionou um decréscimo no consumo de energia nos meses de maio, junho, julho, agosto e setembro.

Pela Figura 23 (a) observa-se que em 2013 uma situação interessante foi o salto de potência de novembro. Observa-se que esse mês chegou ao valor de aproximadamente 1400 kW, o que pode ser explicado pelas altas temperaturas ocorridas nesse período, acarretando assim o uso com maior frequência dos sistemas de refrigeração.

Tem-se que o salto de potência apresentado no mês de novembro contribuiu para o aumento da variabilidade dos dados de potência para esse mês. Isso pode ser visualizado no *boxplot* da Figura 23 (b). Outra característica importante indicada pelos *boxplots* é a sazonalidade presente nos meses do ano, com baixas medidas de potência nos meses de janeiro e julho, que são meses de férias dos

alunos da universidade.

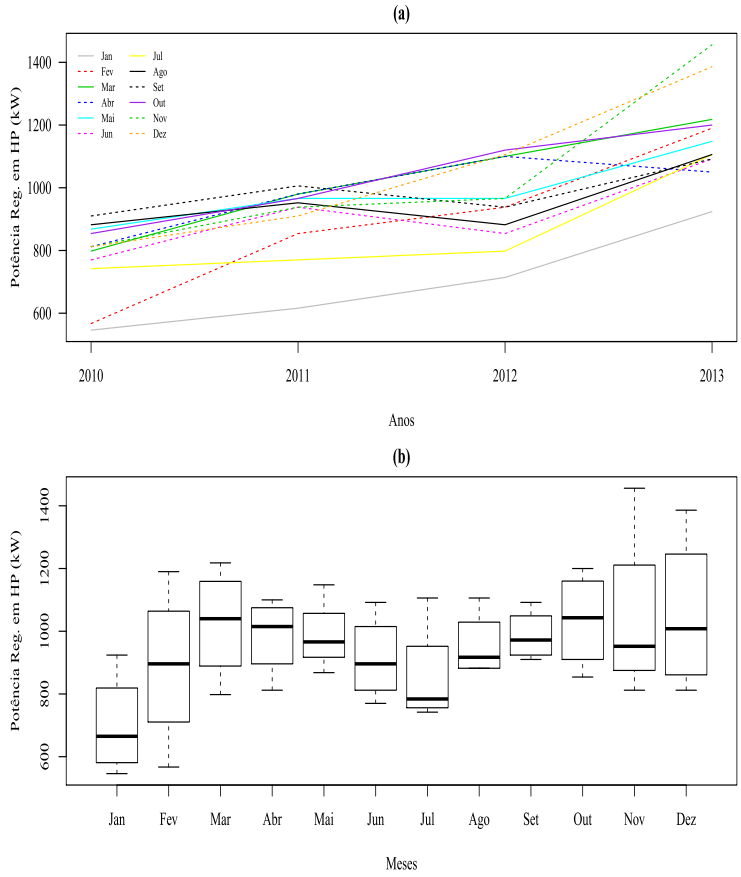

Figura 23 Séries mensais (a) e boxplots (b) da demanda de potência registrada em horário de ponta da UFLA, medida em kW, no período de janeiro de 2010 a dezembro de 2013

Por meio da Figura 23, tem-se a impressão de que o mês de janeiro se diferencia dos demais, formando assim, dois grupos.

O dendrograma dos meses, apresentado na Figura 24, também tem um coeficiente de correlação cofenética alto (0,8840), indicando boa qualidade no agrupamento.

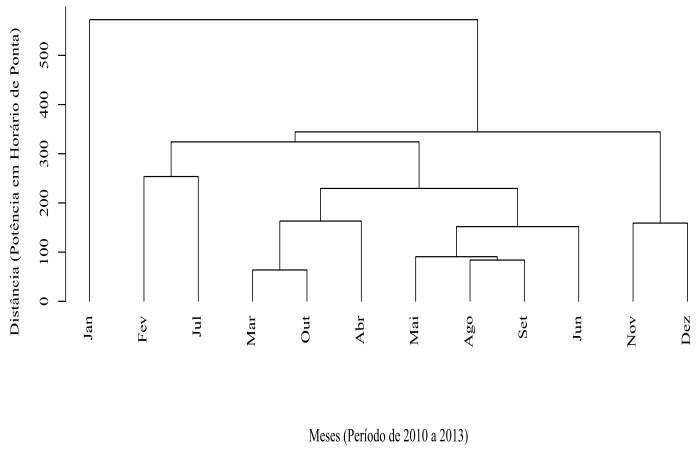

Figura 24 Dendrograma para agrupamento dos meses do ano obtido a partir de dados da demanda de potência registrada em horário de ponta da UFLA, medidas em kW, no período de janeiro de 2010 a dezembro de 2013

No dendrograma da Figura 24, os cortes feitos são:

a) Corte 1: jan e fev, mar, abr, mai, jun, jul, ago, set, out, nov, dez.

b) Corte 2: jan; nov, dez e fev, mar, abr, mai, jun, jul, ago, set, out.

c) Corte 3: jan; nov, dez; fev, jul e mar, abr, mai, jun, ago, set, out.

d) Corte 4: jan; fev; jul; nov, dez e mar, abr, mai, jun, ago, set, out.

e) Corte 5: jan; fev; jul; nov, dez; mar, abr, out e mai, jun, ago, set.

f) Corte 6: jan; fev; jul; abr; nov, dez; mar, out e mai, jun, ago, set.

g) Corte 7: jan; fev; jul; abr; nov; dez; mar, out e mai, jun, ago, set.

h) Corte 8: jan; fev; jul; abr; nov; dez; jun; mar, out e mai, ago, set.

i) Corte 9: jan; fev; jul; abr; nov; dez; jun; mai; mar, out e ago, set.

j) Corte 10: jan; fev; jul; abr; nov; dez; jun; mai; ago; set e mar, out.

Na Tabela 9, tem-se que, para o índice de Moran, o corte selecionado é o terceiro, e para o índice de Geary o corte escolhido é o oitavo. O Corte 3 é dado por: jan; nov, dez; fev, jul e mar, abr, mai, jun, ago, set, out, e o corte 8 é composto por: jan; fev; jul; abr; nov; dez; jun; mar, out e mai, ago, set. Este último separa mais os meses, contudo é um agrupamento que tem coerência com a realidade. O corte 3 também é pertinente, observa-se por exemplo, a formação de dois grupos, fevereiro e julho, novembro e dezembro, que são coerentes pois, no primeiro estão os meses que possuem menor demanda em decorrência de os alunos estarem de férias, e no segundo contém meses considerados de maiores medições de demanda de potência, visto que maiores temperaturas são observadas nesse período.

Tabela 9 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,0136	0,0519	0,4252	0,0355
Segundo	0,1451	0,0565	0,3046	0,0052
Terceiro	0,3531	0,0495	0,1140	0,0001
Quarto	0,2815	0,0443	0,1274	0,0028
Quinto	0,3385	0,0689	0,0752	0,0009
Sexto	0,3912	0,0789	0,0614	0,0016
Sétimo	0,1951	0,1787	0,0801	0,0130
Oitavo	0,2841	0,1595	0,0214	0,0034
Nono	0,3231	0,2125	0,0263	0,0253
Décima	0,6337	0,1444	0,0059	0,0825

Para o agrupamento formado pelo índice de Moran, os máximos de potência obtidos pelos grupos foram: jan = 924 kW, nov, dez = 1456 kW, sendo

novembro o maior valor, fev, jul = 1190 kW, sendo fevereiro o maior valor e mar, abr, mai, jun, ago, set, out = 1218 kW, sendo que março atingiu o maior valor de potência.

Considerando o agrupamento formado pelo índice de Geary, os máximos encontrados nos grupos foram: jan = 924 kW, fev = 1190 kW, jul = 1106 kW, abr = 1100 kW, nov = 1456 kW, dez = 1386 kW, jun = 1092 kW, mar, out = 1218 kW, sendo março o maior valor e mai, ago, set = 1148 kW, sendo este valor alcançado no mês de maio.

Observou-se com as análises no período de ponta que a demanda de potência muda o comportamento ao longo dos anos, acarretando, por exemplo, a formação de agrupamentos "ótimos" diferentes quando se utiliza as séries originais e a partição que compreende os anos de 2010 a 2013.

4.1.4 Meses - demanda de potência registrada em horário fora de ponta - DPRHFP

Na Figura 25, têm-se as séries temporais e os *boxplots* da demanda de potência registrada em horário fora de ponta da UFLA, no período de janeiro de 1995 a dezembro de 2013.

O período fora de ponta na UFLA têm valores de demanda de potência maiores do que o período de ponta, devido ao fato de que a universidade tem a maioria das atividades realizadas no período diurno, apesar de que nos últimos anos esta disparidade tem diminuído.

Observa-se, nas séries temporais da Figura 25, uma tendência crescente nos dados de potência em horário fora de ponta da UFLA. Observa-se que os valores no horário fora de ponta são superiores aos registrados no período de ponta, o que é explicado pelos motivos já mencionados.

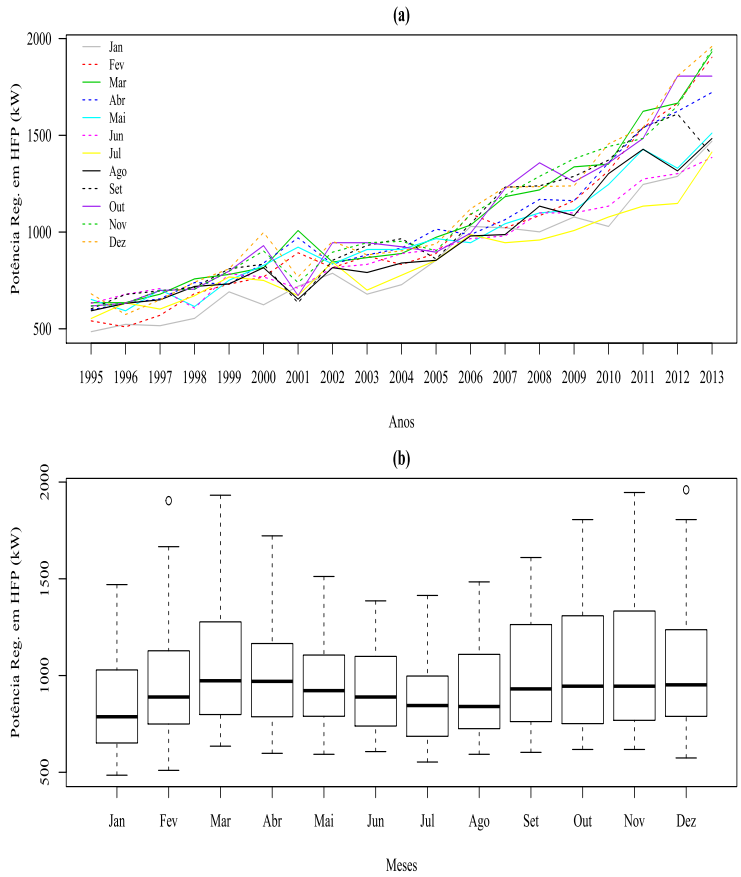

Figura 25 Séries mensais (a) e boxplots (b) da demanda de potência registrada em horário fora de ponta da UFLA, medida em kW, no período de janeiro de 1995 a dezembro de 2013

Nos *boxplots* da Figura 25 (b), observa-se um comportamento sazonal nos dados, cenário também esperado quando se trata da variável analisada. Visualiza-se também a presença de dois *outliers*, um no mês de fevereiro e outro em dezembro. Isso ocorreu principalmente devido às altas temperaturas que ocorreram nos últimos anos da série analisada, que fizeram com que o uso de ares-condicionados ocorressem com mais frequência.

A partir da Figura 25 (a), observa-se também que de junho a dezembro de 2001 há uma queda nos valores da DPRHFP, o que também ocorreu no período de ponta. Como já explicado, nesse período houve racionamento de energia elétrica para todos os consumidores, inclusive para a UFLA e por isso esses valores ocorreram nas séries de demanda da universidade.

Assim como no período de ponta, aqui também os valores de potência foram afetados pela greve nas universidades federais, que ocorreu no ano de 2001. A greve foi do fim do mês de agosto até o final do mês de novembro, totalizando quase 100 dias.

O dendrograma dos meses, para a variável analisada, considerando o período de janeiro de 1995 a dezembro de 2013, é apresentado na Figura 26. O coeficiente de correlação cofenética aqui também foi alto (0,8129), indicando assim boa qualidade no agrupamento.

No dendrograma da Figura 26, os cortes feitos são:

a) Corte 1: jan, mai, jun, jul, ago e fev, mar, abr, set, out, nov, dez.

b) Corte 2: set; jan, mai, jun, jul, ago e fev, mar, abr, out, nov, dez.

c) Corte 3: set; jan, jul; mai, jun, ago e fev, mar, abr, out, nov, dez.

d) Corte 4: set; jan, jul; mai, jun, ago; fev, mar, abr e out, nov, dez.

e) Corte 5: set; fev; jan, jul; mar, abr; mai, jun, ago e out, nov, dez.

f) Corte 6: set; fev; mai; jan, jul; mar, abr; jun, ago e out, nov, dez.

g) Corte 7: set; fev; mai; jan; jul; mar, abr; jun, ago e out, nov, dez.

h) Corte 8: set; fev; mai; jan; jul; mar; abr; jun, ago e out, nov, dez.

i) Corte 9: set; fev; mai; jan; jul; mar; abr; jun; ago e out, nov, dez.

j) Corte 10: set; fev; mai; jan; jul; mar; abr; jun; ago; out e nov, dez.

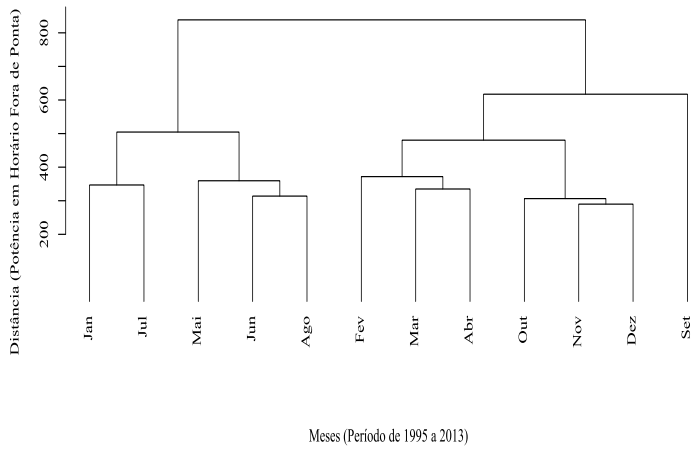

Figura 26 Dendrograma para agrupamento dos meses do ano, obtido a partir de dados da demanda de potência registrada em horário fora de ponta da UFLA, medidas em kW, no período de janeiro de 1995 a dezembro de 2013

A Tabela 10 mostra que o corte selecionado pelo índice de Moran é o sexto, cujos grupos são: set; fev; mai; jan, jul; mar, abr; jun, ago e out, nov, dez. Para o índice de Geary, os grupos obtidos são: set; fev; mai; jan; jul; mar; abr; jun, ago e out, nov, dez, referentes ao corte 8.

Observa-se que os dois agrupamentos obtidos são interessantes, o segundo um pouco mais segmentado, com 9 grupos, o primeiro com menos grupos, 7. A escolha fica a cargo do pesquisador/administrador, que pode, a partir de um acordo com a concessionária de energia, firmar um contrato com mais ou menos segmentações.

Tabela 10 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,7256	0,0007	0,2515	0,0007
Segundo	0,7918	0,0003	0,2621	0,0006
Terceiro	0,9486	0,0001	0,1184	0,0001
Quarto	0,9810	0,0001	0,0887	0,0002
Quinto	1,1157	0,0000	0,0633	0,0002
Sexto	1,2594	0,0000	0,0449	0,0005
Sétimo	0,8002	0,0327	0,0556	0,0016
Oitavo	0,9202	0,0317	0,0184	0,0015
Nono	1,1383	0,0213	0,0219	0,0111
Décimo	1,3089	0,0537	0,0118	0,0829

4.1.4.1 Período de janeiro de 2010 a dezembro de 2013

A Figura 27 mostra as séries temporais e os *boxplots* da demanda de potência registrada em horário fora de ponta da UFLA, no período de janeiro de 2010 a dezembro de 2013.

Por meio dos *boxplots* apresentados na Figura 27 (b), observa-se um efeito claramente sazonal da variável analisada, visualiza-se também que fevereiro, março, abril, outubro, novembro e dezembro são os meses que possuem maior demanda de potência.

O constante crescimento na demanda é observado de forma nítida no período aqui analisado. Pela Figura 27 (a), visualiza-se uma queda dos valores de potência em alguns meses do ano de 2012, mais especificamente, nos meses de maio, junho, julho, agosto e setembro. A explicação para esse fato se deve a uma greve dos professores que ocorreu nesse período.

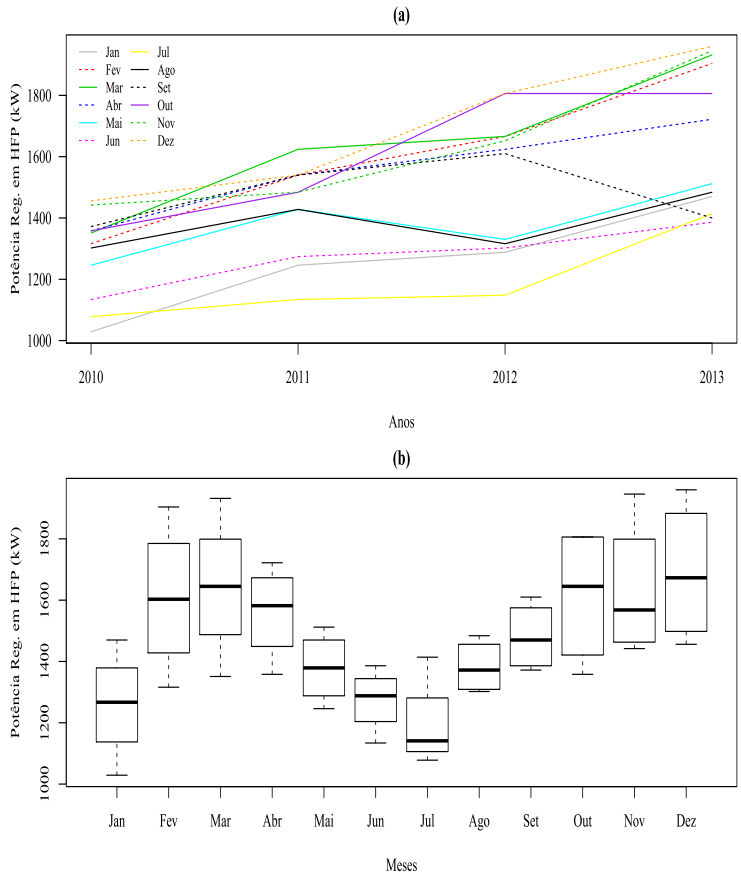

Figura 27 Séries mensais (a) e boxplots (b) da demanda de potência registrada em horário fora de ponta da UFLA, medida em kW, no período de janeiro de 2010 a dezembro de 2013

No período aqui analisado, uma queda visível na demanda ocorre no mês de setembro de 2013, essa medida é reflexo da greve de 2013, pois, no calendário acadêmico, esse mês foi destinado para férias dos estudantes, sendo que geralmente as férias do meio do ano são em julho. Outra consequência da greve foi que o mês de janeiro de 2013 superou em termos de demanda o mês de fevereiro, o que normalmente não ocorre.

Um diagnóstico relevante que se faz a partir desses últimos anos é que fevereiro está assumindo maiores valores de demanda em horário fora de ponta, talvez devido à ocorrência de temperaturas mais altas. O mês de janeiro não tem a mesma tendência, apesar de ter também dias com altas temperaturas, pois neste mês geralmente as atividades na universidade são menos intensas.

Outra constatação que se faz a partir das séries de DPRHFP, no período aqui analisado, é que há uma tendência em se formar dois grupos: o primeiro contendo os meses de janeiro, maio, junho, julho e agosto e o segundo grupo com fevereiro, março, abril, setembro, outubro, novembro e dezembro.

O dendrograma dos meses, para a variável analisada, considerando o período de janeiro de 2010 a dezembro de 2013, é apresentado na Figura 28. O coeficiente de correlação cofenética é consideravelmente alto (0,8633), indicando assim boa qualidade no agrupamento.

No dendrograma da Figura 28, os cortes feitos são:

a) Corte 1: jan, mai, jun, jul, ago, set e fev, mar, abr, out, nov, dez.

b) Corte 2: set; jan, mai, jun, jul, ago e fev, mar, abr, out, nov, dez.

c) Corte 3: set; mai, ago; jan, jun, jul e fev, mar, abr, out, nov, dez.

d) Corte 4: set; mai, ago; jan, jun, jul; abr, out e fev, mar, nov, dez.

e) Corte 5: set; abr; out; mai, ago; jan, jun, jul e fev, mar, nov, dez.

f) Corte 6: set; abr; jul; out; mai, ago; jan, jun e fev, mar, nov, dez.

g) Corte 7: set; abr; jul; out; dez; mai, ago; jan, jun e fev, mar, nov.

h) Corte 8: set; abr; jul; out; dez; nov; mai, ago; jan, jun e fev, mar.

i) Corte 9: set; abr; jul; out; dez; nov; jan; jun; mai, ago e fev, mar.

j) Corte 10: set; abr; jul; out; dez; nov; jan; jun; fev; mar e mai, ago.

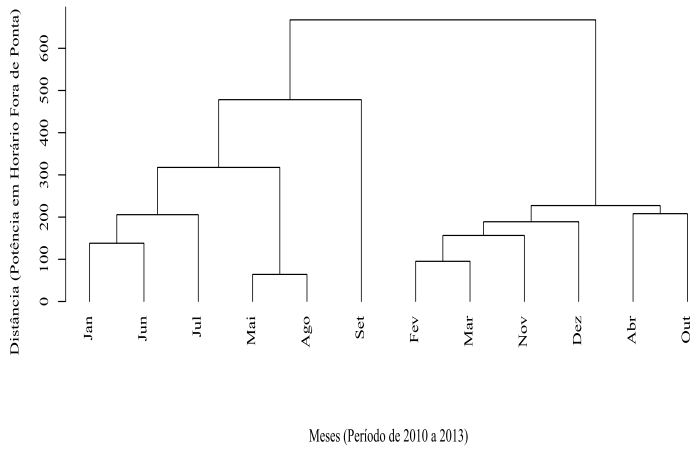

Figura 28 Dendrograma para agrupamento dos meses do ano obtido a partir de dados da demanda de potência registrada em horário fora de ponta da UFLA, medidas em kW, no período de janeiro de 2010 a dezembro de 2013

A Tabela 11 mostra que, para o índice de Moran, o melhor corte é o quinto, e, para o índice de Geary, é o corte 7. Tem-se que o corte 5 é dado por: set; abr; out; mai, ago, jan, jun, jul e fev, mar, nov, dez e o corte 7 é dado por: set; abr; jul; out; dez; mai, ago; jan, jun e fev, mar, nov.

Levando em consideração o agrupamento formado pelo índice de Moran, os máximos de potência ativa obtidos pelos grupos foram: set = 1414 kW, abr = 1722 kW, out = 1806 kW, mai, ago = 1512 kW, valor obtido em maio, jan, jun, jul = 1470 kW, valor observado em janeiro e fev, mar, nov, dez = 1960 kW, valor alcançado no mês de dezembro.

Tabela 11 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,7657	0,0008	0,2148	0,0011
Segundo	0,9424	0,0000	0,1361	0,0001
Terceiro	1,0357	0,0000	0,0505	0,0000
Quarto	1,0468	0,0000	0,0403	0,0000
Quinto	1,1845	0,0001	0,0390	0,0000
Sexto	0,9904	0,0036	0,0220	0,0000
Sétimo	0,9118	0,0180	0,0063	0,0000
Oitavo	0,9154	0,0416	0,0090	0,0006
Nono	0,5660	0,1860	0,0114	0,0082
Décimo	0,3333	0,3699	0,0002	0,0074

Para o agrupamento formado pelo índice de Geary, os máximos encontrados nos grupos foram: set = 1610 kW, abr = 1722 kW, jul = 1414 kW, out = 1806 kW, dez = 1960 kW, mai, ago = 1512 kW, observado em maio, jan, jun = 1470 kW, valor observado em janeiro e fev, mar, nov = 1946 kW, encontrado no mês de novembro.

Observou-se com as análises no período fora de ponta que a demanda de potência se altera ao longo dos anos, formando assim agrupamentos “ótimos” distintos quando se utilizou as séries originais e a partição que compreende os anos de 2010 a 2013. Este resultado é semelhante ao encontrado para os departamentos/setores, dias da semana e meses no período de ponta.

Nas análises de departamentos/setores, dias da semana e meses, uma observação relevante a se considerar é que os índices de Moran e Geary podem levar à escolha de agrupamentos distintos. Portanto, na prática, as análises devem ser feitas considerando somente um dos índices para que não haja confronto entre os resultados, levando a interpretações incorretas por parte dos administradores. Geralmente o índice de Moran é o preferido, pois, como já mencionado, Moran é

mais poderoso que Geary.

4.2 Componentes independentes (CI's), análise de agrupamento e medidas de similaridades

A seguir, são apresentadas as análises referentes aos componentes independentes obtidos das séries de potência ativa de alguns departamentos/setores e dos dias da semana da UFLA.

4.2.1 Departamentos/setores

Na Figura 29, têm-se CI's obtidos considerando as séries de potência ativa de alguns setores/departamentos da UFLA no dia 12/08/2010. No cálculo dos componentes, foi utilizado o algoritmo FastICA, considerando a negentropia como medida de não gaussianidade e a técnica de branqueamento para reduzir a dimensão dos dados. Com apenas três componentes, pode-se explicar 97,58% da variabilidade dos dados.

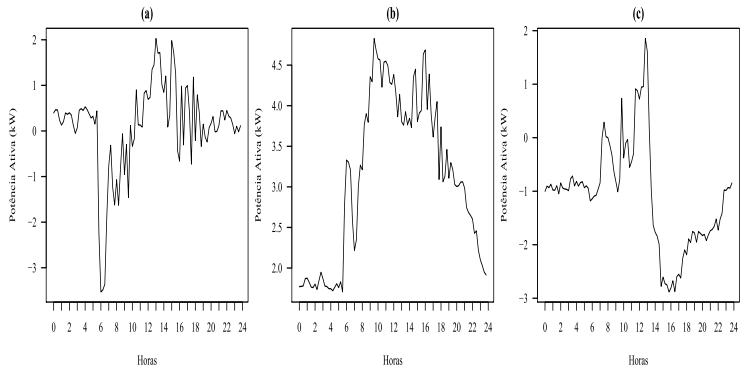

Figura 29 Primeiro (a), segundo (b) e terceiro (c) componente independentes obtidos com o algoritmo FastICA para as séries de potência ativa de alguns setores/departamentos da UFLA, medidas em kW, de 15 em 15 minutos no dia 12/08/2010

Na Figura 30, tem-se o periodograma para cada um dos componentes independentes obtidos anteriormente. No eixo das abcissas do gráfico, foram colocadas as frequências. Sem perda de interpretações, poderia-se utilizar os períodos ao invés das frequências. O estudo do periodograma é importante para que se analise possíveis comportamentos sazonais ou cíclicos da variável em análise.

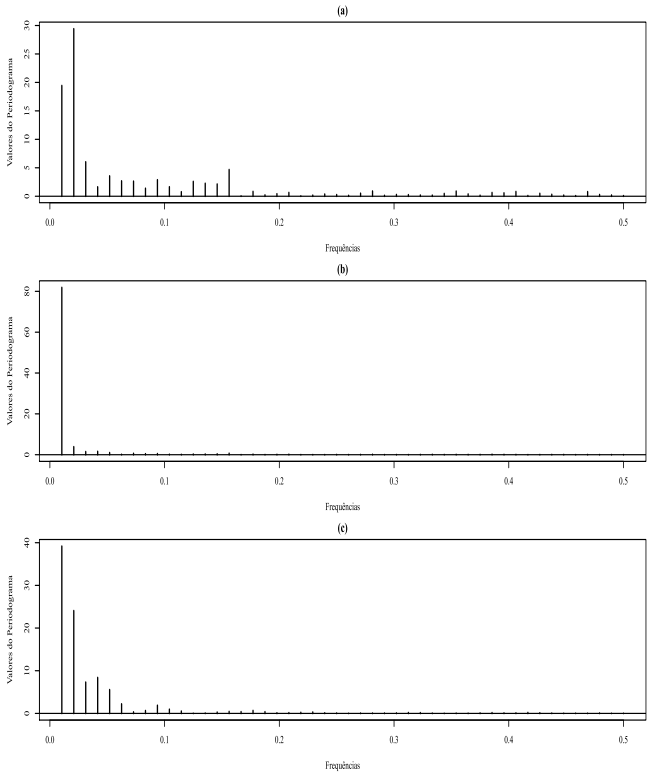

Figura 30 Periodogramas dos componentes independentes, a, b e c, referentes às séries de potência ativa de alguns setores/departamentos da UFLA, medidas em kW, de 15 em 15 minutos no dia 12/08/2010

Por meio do algoritmo FastICA, obteve-se também a matriz de mistura que neste caso foi denominada de $\hat{\mathbf{A}}_1$ (equação 4.1). Considerando uma coluna qualquer desta matriz, tem-se em cada célula um peso referente a cada departamento/setor. Assim, a ideia é comparar os departamentos considerando essas

células para cada componente independente, ou seja, para cada coluna. Um fato relevante a se considerar na matriz de mistura é que quando se tem um coeficiente negativo para um departamento/setor, este tem comportamento inverso ao que esta sendo indicado pelo componente na proporção do valor do referido coeficiente.

$$\hat{\mathbf{A}}_1 = \begin{array}{c} \begin{matrix} \overset{(a)}{\uparrow} & \overset{(b)}{\uparrow} & \overset{(c)}{\uparrow} \end{matrix} \\ \begin{pmatrix} -6,5241 & 8,3766 & 2,1653 \\ 3,2710 & 8,0861 & -0,8013 \\ 1,6279 & 4,1047 & -3,2470 \\ 3,7624 & 8,2553 & -2,5610 \\ 6,6143 & 14,9950 & -3,6501 \\ 1,3535 & 2,9071 & -1,9894 \\ 2,4342 & 6,4084 & -2,1586 \\ 8,6639 & 26,0495 & 32,6856 \end{pmatrix} \end{array} \begin{array}{c} \\ \rightarrow CANTINA \\ \rightarrow CIN \\ \rightarrow DAE \\ \rightarrow DCC-T \\ \rightarrow DCC-PP \\ \rightarrow DEX \\ \rightarrow EPAMIG \\ \rightarrow RU \end{array} \tag{4.1}$$

A partir da matriz dada em 4.1, fez-se a comparação dos departamentos/setores, considerando cada componente independente separadamente. Realizou-se AA, construção dos dendrogramas, e posteriormente a obtenção dos melhores agrupamentos, considerando os índices de Moran e de Geary, e os respectivos testes a 5% de significância. Os coeficientes de correlação cofenética para os dendrogramas associados aos componentes (a), (b) (c), foram 0,9204, 0,9517 e 0,9972, respectivamente, indicando assim boa qualidade nos agrupamentos. Os resultados referentes aos dendrogramas, estimativas dos coeficientes de autocorrelação e respectivos valores-p podem ser encontrados no (Anexo B). Os agrupamentos para os 8 departamentos/setores, de acordo com os três componentes, estão na Tabela 12.

A seguir, têm-se as observações com relação aos componentes e os respectivos agrupamentos dos departamentos/setores.

No primeiro componente (Figura 29a), tem-se um período atípico para os valores da potência ativa, ou seja, observa-se uma queda nos valores entre as 6 e 7 horas aproximadamente. Após as 7 horas, tem-se uma tendência de crescimento nos valores, que tem seu ponto de maior magnitude por volta das 13 horas. Em seguida há um período de baixos valores, mas em torno das 15 horas existe um pico na quantidade de potência ativa. Depois das 15 a variável analisada segue com algumas oscilações de menores proporções.

Tabela 12 Melhores agrupamentos (departamentos/setores) por componente, considerando os índices de Moran e de Geary, e os respectivos testes a 5% de significância

Componente	Moran
(a)	(CANTINA), (DCC-PP, RU) e (DAE, DEX, CIN, DCC-T, EPAMIG)
(b)	(RU), (DCC-PP), (DAE, DEX) e (CANTINA, CIN, DCC-T, EPAMIG)
(c)	(RU), (CANTINA), (CIN), (DCC-T), (DAE, DCC-PP) e (DEX, EPAMIG)
Componente	Geary
(a)	(CANTINA), (DCC-PP), (RU), (CIN), (DCC-T), (EPAMIG) e (DAE, DEX)
(b)	(RU), (CIN), (DAE), (DCC-PP), (DEX), (EPAMIG) e (CANTINA, DCC-T)
(c)	(RU, CANTINA, CIN, DAE, DCC-T, DCC-PP) e (DEX, EPAMIG)

Na Figura 30 (a), que apresenta o periodograma para o componente (a), visualiza-se um pico no período 96 (freqüência 0,0104). Porém, este não tem sentido neste estudo, pois existem somente 96 observações na série analisada. Neste trabalho, em todas as situações analisadas, este período não foi levado em consideração. Outro pico observado indica periodicidade de ordem 48 (frequência 0,0208). Com o intuito de testar a existência do efeito sazonal de período 48, aplicou-se o teste de Fisher. Este teste foi descrito por Priestley (1989). Para o teste, tem-se que g é a estatística teste, e z é o valor crítico, que está relacionado ao nível de significância. Por meio das análises, chegou-se às estatísticas $g = 0,3065$ e $z_{0,05} = 0,1359$, como $g > z$, a série apresenta sazonalidade de ordem 48, a 5% de significância. Portanto, para o componente (a) existe uma regularidade da

variável em análise de 12 em 12 horas.

Para o componente (a), o coeficiente associado à CANTINA possui valor negativo (ver em 4.1), indicando comportamento contrário aos mencionados. Já para os outros departamentos/setores, os valores dos coeficientes são positivos. Os maiores pesos considerando esse componente são atribuídos ao DCC-PP, a CANTINA e ao RU.

Para o índice de Moran, os grupos de melhor qualidade formados, considerando o componente (a) foram: CANTINA; DCC-PP, RU e DAE, DEX, CIN, DCC-T, EPAMIG. Observa-se que nesse agrupamento a CANTINA é o único setor que possui peso negativo para o componente (a), indicando que este possui comportamento inverso ao apresentado pela Figura 30 (a), ou seja, observa-se, por exemplo, que a CANTINA tem alta na potência ativa entre as 6 e 7 horas, período em que são preparados os alimentos. Após as 7 horas, a CANTINA possui uma tendência de decrescimento nos valores e uma sequência de oscilações. Estas podem ser explicadas pelos intervalos que os alunos têm entre as aulas. Veja que estas características se enquadram de forma satisfatória ao perfil da CANTINA.

Outro grupo obtido é o DCC-PP e o RU, que têm comportamentos similares, fato que pode ser observado nos coeficientes de ambos para este componente.

O RU possui o maior coeficiente positivo dentre todos os departamentos/setores. Isso indica que os comportamentos descritos para o mesmo são mais intensos, com ênfase ao crescimento ocorrido pela manhã, comportamento realmente esperado para este setor, visto que durante a manhã é realizada a preparação dos alimentos para o almoço.

Tem-se também que a maior magnitude na potência, por volta das 13 horas (Figura 29a), pode ser explicada para esse setor, o valor ocorrido é devido ao período pós almoço, em que há a limpeza do ambiente, e utilização de máquinas

pelos funcionários deste setor.

Os demais departamentos/setores se juntaram em um único grupo, sendo os coeficientes/pesos mais baixos no componente (a).

Para o índice de Geary, o melhor corte foi o sexto, sendo o agrupamento selecionado formado pelos grupos CANTINA; DCC-PP; RU; CIN; DCC-T; EPA-MIG e DAE, DEX. Observa-se que nesse corte sete grupos foram encontrados, houve a junção somente entre DAE e DEX, que possuem os menores pesos para o componente (a). Todos os outros departamentos/setores ficaram separados. Esse corte também é considerado pertinente, porém com uma segmentação maior dos departamentos/setores.

Para o segundo componente independente (Figura 29b), inicialmente observa-se o mesmo comportamento descrito no componente anterior, ocorrido entre as 6 e 7 horas, porém em sentido contrário, ou seja, há uma alta nos valores neste período. A partir daí os valores começam a crescer, chegando a uma quantidade máxima por volta das 10 horas. Entre as 10 e 15, o comportamento da potência forma uma "parábola" com concavidade para cima, ou seja, há um decrescimento e logo em seguida um crescimento da variável analisada. Depois das 15 horas a variável potência segue uma tendência de decrescimento, que continua até o fim da noite.

A periodicidade do componente (b) pode ser analisada a partir da Figura 30 (b). Por meio da figura, observa-se um pico menor no período 48 (frequência 0,0208). Aplicou-se então o teste de Fisher para confirmar tal periodicidade, as estatísticas obtidas foram $g = 0,2754$ e $z_{0,05} = 0,1383$, como $g > z$, a série apresenta sazonalidade de ordem 48, a 5% de significância. Portanto, a variável analisada apresenta uma regularidade de 12 em 12 horas.

Considerando os pesos para esse componente, que estão na segunda co-

luna da matriz $\hat{\mathbf{A}}_1$ dada em 4.1, observa-se que para todos os departamentos/setores, os valores são positivos, indicando que estes departamentos/setores possuem comportamento como o apresentado na Figura 29 (b).

Por meio dos coeficientes apresentados na segunda coluna da matriz $\hat{\mathbf{A}}_1$ e pela AA, encontrou-se o grupo RU; DCC-PP; DAE, DEX e CANTINA, CIN, DCC-T, EPAMIG, para o índice de Moran, e RU; CIN; DAE; DCC-PP; DEX; EPAMIG e CANTINA, DCC-T, segundo o índice de Geary. Estes foram os grupos considerados de melhor qualidade.

O RU possui o maior coeficiente positivo para este componente, indicando assim, que os comportamentos descritos para o mesmo são mais intensos para este setor. Assim, há um crescimento da potência no período da manhã, fato explicado pela preparação dos alimentos, após as 10 horas há uma queda, momento em que é servido o almoço, e logo após o almoço, devido à higienização do ambiente, a potência tende a crescer.

O DCC-PP possui o segundo maior coeficiente, enquanto que os demais departamentos/setores possuem pesos inferiores para esse componente, sendo o DAE e o DEX os que possuem menores valores.

Para o DCC-PP, a série temporal apresentada pelo componente (b) também tem interpretação prática, visto que nesse departamento há um crescimento na potência durante a manhã, uma queda em torno das 12 horas, e uma alta no período da tarde. Esse comportamento é típico para departamentos com as características do DCC-PP, que possuem salas de aula, salas de professores e laboratórios.

Assim como no componente (a), pode-se visualizar também no componente (b) o comportamento apresentado pela CANTINA no período das 6 às 7 horas. Como o coeficiente deste setor para este componente foi positivo, o comportamento é como o apresentado pela Figura 29 (b), ou seja, forma-se um pico de

crescimento para a potência no período mencionado.

Para o componente (c), Figura 29 (c), verifica-se uma sazonalidade no comportamento da potência no período das 5 às 15 horas, formando três intervalos. Esses intervalos tem picos às 7, 10 e 13 horas aproximadamente, sendo este último o de maior proporção.

A partir das 13 horas, há um decaimento abrupto na potência ativa, atingindo valores mínimos às 15 horas. Nos momentos subsequentes, a variável em análise volta a crescer de forma mais lenta.

Na Figura 30 (c), pode-se observar uma periodicidade de ordem 48 (frequência 0,0208) no componente (c), indicando regularidade da variável em análise de 12 em 12 horas. Para testar a existência do efeito sazonal no período indicado, aplicou-se o teste de Fisher. Chegou-se as estatísticas $g = 0,4237$ e $z_{0,05} = 0,1383$, como $g > z$, a série apresenta sazonalidade de ordem 48, a 5% de significância.

Na terceira coluna da matriz $\hat{\mathbf{A}}_1$ (4.1), observa-se que somente os coeficientes para a CANTINA e o RU foram positivos, o que indica que para esses setores o comportamento é como apresentado na Figura 29 (c).

Os grupos de departamentos/setores formados para o componente (c) são: RU; CANTINA; CIN; DCC-T; DAE, DCC-PP e DEX, EPAMIG segundo o índice de Moran, e RU; CANTINA; CIN; DAE; DCC-T; DCC-PP e DEX, EPAMIG para o índice de Geary.

Observa-se que para o RU o comportamento é o esperado. Têm-se oscilações sazonais no período da manhã e os picos são explicados pelo uso de utensílios elétricos na cozinha, devido à preparação dos alimentos para o almoço. O pico presente às 13 horas é em virtude da limpeza do ambiente após o almoço. Depois das 13 horas, a tendência é decrescer, pois neste setor não há realização de atividades

de maiores proporções.

O componente (c) mostra um fato bastante interessante. Observa-se que existem dois padrões de comportamento, um durante a manhã e outro à tarde. Para o RU essa separação é compreensível, pois durante a manhã as potências são mais altas, já que estão sendo preparados os alimentos do almoço, enquanto que a tarde o consumo de energia tende a valores mais baixos, já que no período analisado não se servia o jantar. Essa variabilidade fica evidente por meio desse componente. Por meio da matriz $\hat{\mathbf{A}}_1$, pode-se confirmar esse resultado, visto que o RU possui o maior peso para o componente (c).

Neste tópico, fez-se uma associação entre as metodologias de ICA, AA e índices de Moran e Geary com aplicação nas séries de potência ativa dos departamentos/setores em estudo. O objetivo foi mostrar que a partir dos componentes independentes encontrados, pode-se obter relevantes informações para cada um dos departamentos/setores, e também encontrar os grupos de melhor qualidade segundo os índices de Moran e Geary, em que haja similaridade entre os elementos.

4.2.2 Dias da semana

Na Figura 31, têm-se os componentes independentes obtidos a partir das séries diárias de potência ativa da UFLA. Considerando a técnica de branqueamento reduziu-se o número de componentes para três. Por meio desses componentes, pode-se explicar 99,67% da variabilidade dos dados.

Na Figura 32, tem-se o periodograma para cada um dos componentes independentes. Assim como na análise dos componentes referentes aos departamentos/setores, o periodograma aqui apresentado considera, no domínio, as frequências.

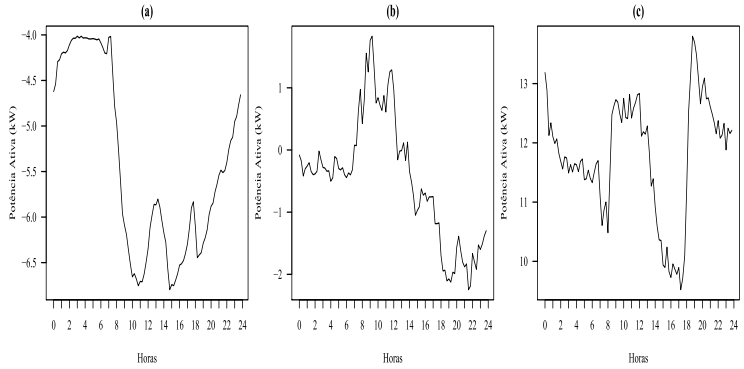

Figura 31 Primeiro (a), segundo (b) e terceiro (c) componente independentes obtidos com o algoritmo FastICA para as séries diárias de potência ativa da UFLA, medida em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013

A matriz de mistura para este caso é apresentado em 4.2. A ideia aqui é comparar os dias da semana considerando cada coluna separadamente, ou seja, cada componente independente, semelhante ao que foi realizado no estudo dos departamentos/setores.

$$\hat{\mathbf{A}}_2 = \begin{array}{c} \begin{matrix} (a) & (b) & (c) \\ \uparrow & \uparrow & \uparrow \end{matrix} \\ \begin{pmatrix} -257,9953 & 4,2025 & -46,4942 \\ -263,2760 & 22,7527 & -37,7031 \\ -250,1663 & 22,5697 & -35,0736 \\ -248,1179 & 24,0545 & -32,1635 \\ -219,3128 & 67,4281 & -19,6191 \\ -17,5804 & -10,2280 & 39,1202 \\ -16,9462 & -50,7016 & 45,9384 \end{pmatrix} \end{array} \begin{array}{l} \\ \rightarrow Seg \\ \rightarrow Ter \\ \rightarrow Qua \\ \rightarrow Qui \\ \rightarrow Sex \\ \rightarrow Sab \\ \rightarrow Dom \end{array} \tag{4.2}$$

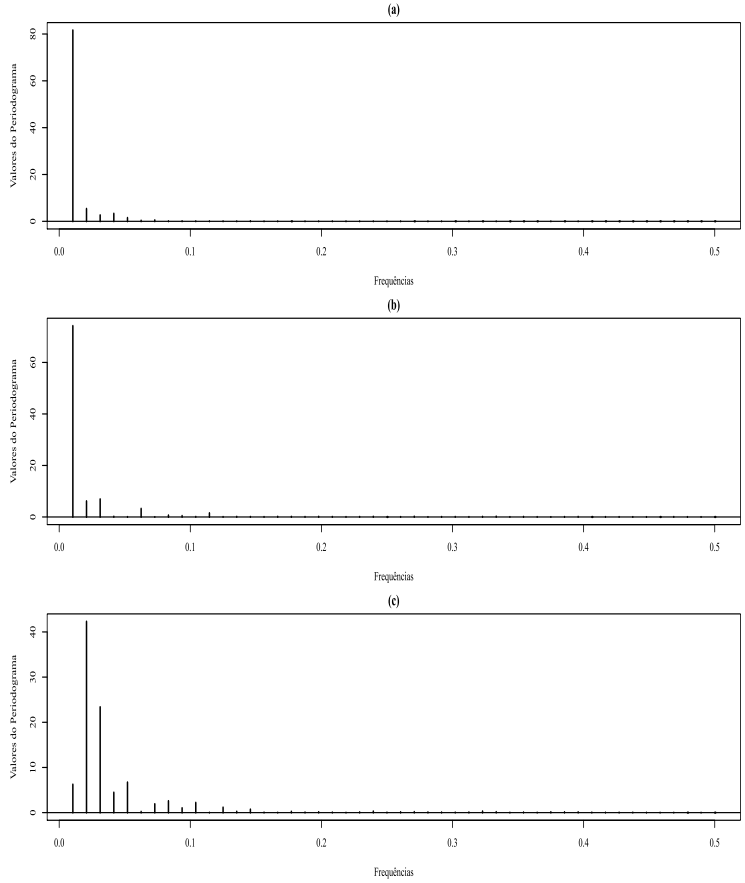

Figura 32 Periodogramas dos componentes independentes, a, b e c, referentes às séries de potência ativa dos dias da semana na UFLA, medidas em kW, de 15 em 15 minutos no dia 12/08/2010

Os coeficientes de correlação cofenética para os dendrogramas associados aos componentes (a), (b) (c), foram 0,9948, 0,8638 e 0,9785 respectivamente, indicando assim boa qualidade nos agrupamentos. Maiores detalhes dos resultados podem ser encontrados no (Anexo B). Os agrupamentos para os sete dias, segundo os três componentes, estão na Tabela 13.

Tabela 13 Melhores agrupamentos (dias) por componente, considerando os índices de Moran e de Geary, e os respectivos testes a 5% de significância

Componente	Moran (p.a.)
(a)	(Seg, Ter, Qua, Qui, Sex) e (Sab, Dom)
(b)	NENHUM
(c)	(Seg), (Sex), (Sab, Dom) e (Ter, Qua, Qui)
Componente	Geary (p.a.)
(a)	(Seg), (Ter), (Qua), (Qui), (Sex) e (Sab, Dom)
(b)	(Seg), (Qui), (Sex), (Sab), (Dom) e (Ter, Qua)
(c)	(Seg), (Qui), (Sex), (Sab), (Dom) e (Ter, Qua)

A seguir têm-se observações com relação aos componentes e seus respectivos agrupamentos dos dias.

No componente (a), observa-se uma regularidade sazonal no comportamento da potência no período das 9 às 20 horas, aproximadamente. Formam-se três intervalos, como pode ser visto na Figura 31 (a). Esses intervalos têm picos de decaimento às 11, 15 e 18 horas, aproximadamente, sendo este último o de menor proporção. A partir das 18 horas, a variável em análise passa a seguir uma tendência de crescimento que vai até o fim da noite.

Na Figura 32 (a), pode-se observar uma periodicidade de ordem 48 (frequência 0,0208) para o componente (a), indicando regularidade da variável em análise de 12 em 12 horas. Aplicou-se o teste de Fisher para confirmar a periodicidade, chegou-se às estatísticas $g = 0,3735$ e $z_{0,05} = 0,1383$, como $g > z$, a série apresenta sazonalidade de ordem 48, a 5% de significância.

Na primeira coluna da matriz $\hat{\mathbf{A}}_2$, têm-se os pesos para o componente (a). Observa-se que para todos os dias, os coeficientes são negativos, o que indica que para todos os dias estudados, o comportamento é inverso ao apresentado na Figura 31 (a). O sinal negativo para os coeficientes são compreensíveis quando se observa as características do componente (a), principalmente devido ao seu com-

portamento periódico. Observa-se, por exemplo, que no período das 12 às 15 horas há um crescimento dos valores, e na realidade o que ocorre é o inverso, ou seja, há um decrescimento da potência ativa nesse período.

Segundo o índices de Moran o agrupamento de melhor qualidade foi: Seg, Ter, Qua, Qui, Sex e Sab, Dom. Já para o índice de Geary, o agrupamento escolhido foi: Seg; Ter; Qua; Qui; Sex e Sab, Dom.

Observa-se que para os dias do fim de semana, os pesos para o componente (a) são os menores e se contrastam com os demais pesos, que possuem valores altos, em módulo. Isso é esperado, pois sábado e domingo são dias em que não há atividades acadêmicas na universidade e, portanto, o consumo de energia tende a ser mínimo em comparação com os dias úteis.

A periodicidade apresentada pelo componente (a) mostra com clareza o perfil da universidade durante os dias úteis, no período da manhã há um crescimento da potência, com pico por volta das 11 horas. Durante o intervalo do almoço, que vai das 12 às 14 horas, forma-se um "vale", com decrescimento por volta das 12 horas e crescimento da potência em torno das 14 horas, o qual é reflexo da interrupção das atividades acadêmicas para o horário do almoço. Um pico de potência é formado às 15 horas, horário que geralmente há um maior consumo de energia na universidade. A partir daí há um decrescimento nos valores da variável analisada, observando-se uma queda até por volta das 18 horas. Logo depois, forma-se um novo intervalo de crescimento, porém com um pico menor, o que pode ser explicado pelo fato de que à noite o número de cursos de graduação é menor, portanto a potência medida tende a ser menor.

Na Figura 31 (b), tem-se o segundo componente independente. Observa-se que, a partir das 6 horas, a potência ativa começa a crescer, atingindo um valor máximo por volta das 9 horas. Na sequência, o comportamento é de decresci-

mento, porém de maneira sazonal, em períodos de aproximadamente três horas. O menor valor assumido pela variável neste período de decréscimo é atingido por volta das 21 horas.

Pode-se observar na Figura 32 (b) uma periodicidade de ordem 48 (frequência 0,0208) para o componente (b), indicando regularidade da variável em análise de 24 em 24 horas. Aplicou-se então o teste de Fisher para confirmar a periodicidade indicada. As estatísticas obtidas foram $g = 0,4134$ e $z_{0,05} = 0,1407$, como $g > z$, a série apresenta sazonalidade de ordem 48, a 5% de significância. Portanto a variável analisada apresenta uma regularidade de 12 em 12 horas.

Observa-se na matriz dada em 4.2, que os valores da segunda coluna são negativos somente para o sábado e domingo. Portanto, tem-se que nesses dias o comportamento é inverso ao apresentado pelo componente (b).

Outra observação a se levar em conta, é que numericamente o valor associado a sexta-feira, para esse componente, é relativamente alto quando comparado aos demais dias. Isso pode ser um indício de formação de grupos com características semelhantes, em que, por exemplo, a sexta-feira seria um dia dissimilar a qualquer outro, assim como o domingo que possui o maior valor negativo.

O peso para o componente (b) é menor para a segunda-feira e maior para a sexta-feira. Baseado no componente (b), esses valores são compreensíveis, pois na segunda as atividades estão no início, sendo que no período da manhã algumas destas ainda não são realizadas, já na sexta ocorre o inverso, o período da tarde tem um menor ritmo.

Observações semelhantes podem ser feitas também para o sábado e domingo. Tem-se que o sábado tende a ter potências maiores durante a manhã e menores à tarde, pois durante a manhã algumas atividades ainda são realizadas. Já o domingo tem valores inversos, ou seja, baixos valores durante a manhã, com

potências em crescimento à tarde, visto que neste período os alunos do alojamento estão retornando para a universidade. Portanto, os coeficientes negativos para esses dias são coerentes com as observações realizadas.

Com as observações feitas anteriormente, pode-se concluir por meio da ICA que a segunda e sexta não são dias típicos, ou seja, o comportamento da variável potência nesses dias segue algumas particularidades que não são encontradas nos demais dias da semana. Isso vale também para dias do fim de semana. Jesus (2011), por meio de métodos estatísticos descritivos, já havia observado que segunda e sexta-feira não seriam dias representativos. Segundo esse autor, nesses dias há alunos retornando e saindo da universidade.

Nenhum agrupamento significativo foi identificado ao utilizar o índice de Moran. Nesse caso, fica a cargo do pesquisador fazer a escolha do agrupamento que melhor representa os dados, ou utilizar outro critério de escolha, como por exemplo, o índice de Geary. Para este componente, o índice de Geary indicou como melhor escolha o agrupamento: Seg; Qui; Sex; Sab; Dom e Ter, Qua.

Observa-se que terça e quarta-feira ficaram juntos em um grupo, como era de se esperar, considerando os valores próximos dos seus coeficientes para o componente (b). Enquanto que os outros dias se separaram, formando grupos distintos.

Veja que o domingo tem comportamento inverso ao apresentado, ou seja, valores de potência baixos pela manhã, crescendo a partir do fim da tarde e início da noite, momentos em que os alunos estão voltando para o alojamento estudantil. Portanto, o consumo de energia tende a aumentar. Já a sexta-feira tem um comportamento bem similar ao apresentado pelo componente (b), com altos valores pela manhã, decaindo de forma significativa após as 13 horas.

O componente (c) pode ser visualizado na Figura 31 (c). Observa-se que a

potência ativa no período das 0 às 6 horas não tem comportamento constante, como foi apresentado nos componentes anteriores. Neste período há um decaimento dos valores, chegando a uma estabilidade que vai das 2 às 6 horas.

Pode-se observar na Figura 32 (c) uma periodicidade de ordem 48 (freqüência 0,0208) para o componente (c), indicando regularidade da variável em análise de 12 em 12 horas. Para testar a existência do efeito sazonal no período indicado, aplicou-se o teste de Fisher. Chegou-se às estatísticas $g = 0,4406$ e $z_{0,05} = 0,1359$, como $g > z$, a série apresenta sazonalidade de ordem 48, a 5% de significância.

Na terceira coluna da matriz dada em 4.2, têm-se os pesos dos dias para o componente (c). Observa-se que neste caso somente os pesos associados ao sábado e ao domingo são positivos. Assim, para estes o comportamento do componente (c) é como o apresentado na Figura 31 (c).

Tem-se também que o peso para a segunda-feira é numericamente maior, o da sexta-feira é menor, e os demais terça, quarta e quinta ficam em posição intermediária. Estes valores já indicam uma provável formação de grupos, neste caso um total de três.

Considerando os pesos do componente (c) para cada dia, pode-se observar que, na segunda-feira o maior peso negativo indica uma característica bastante peculiar desse dia, que é de possuir valores de alta potência à tarde e baixos valores durante a manhã, fato explicado pela ausência de alunos na universidade neste período. Já à tarde, os alunos estão chegando, e consequentemente há um crescimento nos valores da variável em estudo.

Situação semelhante ocorre na sexta-feira, em que os alunos durante a manhã estão realizando atividades na universidade, enquanto que à tarde alguns desses já não se encontram mais no campus. Essas observações foram feitas de

forma similar no componente (b).

Portanto, o componente (c) retrata de forma clara algumas variabilidades presentes em certos dias da semana.

De acordo com o índice de Moran, formaram-se quatro grupos para este componente, um com a segunda, outro com a sexta, outro com o sábado e domingo, e por fim um contendo terça, quarta e quinta.

Segundo o índice de Geary, os grupos de melhor qualidade foram: Seg; Qui; Sex; Sab; Dom e Ter, Qua.

Observa-se ser bastante coerente com a realidade os agrupamentos propostos pelos índices de Moran e Geary.

É importante salientar que a análise apresentada aqui, envolvendo ICA, AA e índices de Moran e de Geary, proporcionaram, além das características encontradas pelos componentes, a obtenção de informações relevantes a partir da análise de agrupamento, uma vez que pode-se formar grupos com características comuns, considerando cada componente.

Assim, como nas análises anteriores, uma observação a se considerar tanto para os departamentos/setores quanto para os dias da semana, é que os índices de Moran e Geary podem levar à escolha de agrupamentos distintos. Normalmente, escolhe-se o índice de Moran, pois é mais poderoso que o índice de Geary.

4.2.3 Fitotecnia, Química e Sementes

Nesta seção, têm-se a análise de três departamentos/setores, cujas medições foram feitas em 28/10/2014. Esta foi uma maneira de se mostrar a metodologia de ICA em dados atualizados. A variável potência foi observada no departamento de Química e nos setores de Fitotecnia e Sementes.

Na Figura 33, têm-se as séries de potência dos três departamentos/setores,

juntamente com os *boxplots* de cada um. Observa-se que a Química possui comportamento diferente dos apresentados pela Fitotecnia e Sementes. Estes últimos se mostram com certa semelhança ao longo do dia, sendo que o Setor de Sementes tem os valores da variável analisada superiores aos da Fitotecnia.

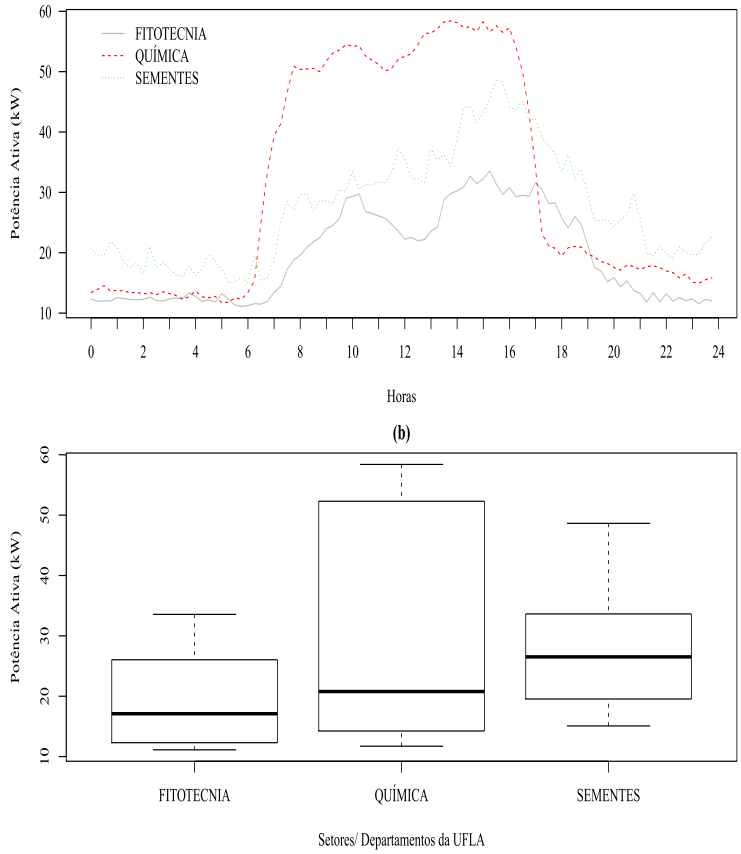

Figura 33 Séries temporais (a) e *boxplots* (b) da potência ativa da Fitotecnia, Química e Sementes, medida em kW, de 15 em 15 minutos no dia 28/10/2014

A Química possui valores de potência acima dos outros dois locais por ter em seu prédio laboratórios com aparelhos que consomem uma maior quantidade

de energia elétrica.

Observa-se nas séries da Figura 33 (a) que na Fitotecnia e na Química, o comportamento de decaimento da potência no período do almoço é mais evidente do que na Sementes, sendo menos perceptível na Química. Neste último, verifica-se um crescimento na potência que começa por volta 7 horas e atinge um máximo em torno de 15 horas.

O comportamento sazonal é observado de forma nítida nas séries de potência da Química e Fitotecnia, sendo que os momentos de maior consumo de energia estão de 10 às 15 horas, o que é esperado para a maioria dos departamentos na UFLA.

Para o setor de Sementes, observa-se que durante o período de 0 às 7 horas, a potência fica acima dos outros dois locais amostrados. Isso ocorre devido ao uso de equipamentos de análise de sementes nesse setor no horário mencionado.

No fim da tarde, após as 17 horas, o departamento de Química praticamente encerra as suas atividades e, por isso, observa-se na série de potência uma queda relevante dos valores nesse horário. Isso leva esse local a possuir uma maior variabilidade nos dados (Figura 33b). Os valores de potência saem de quase 60 kW para 10 kW, aproximadamente, como pode ser observado na Figura 33 (a). Já o setor de Sementes tem claramente atividades até por volta das 22 horas e a partir desse horário há uma estabilidade em torno de 20 kW na potência ativa. O setor de Sementes tem laboratório de prestação de serviços, sala de computação, usina de beneficiamento de sementes, sala de secagem e armazéns climatizados. Toda a estrutura citada faz com que o setor de Sementes tenha um maior consumo de energia e por um período maior do dia. O setor de Fitotecnia tem seus valores de potência reduzidos um pouco mais cedo, em torno das 20 horas.

Na Figura 34, têm-se os componentes independentes selecionados a partir das séries de potência ativa da Fitotecnia, Química e Sementes. Considerando a técnica de branqueamento, reduziu-se a dimensão das dados. Por meio de dois componentes pode-se explicar 99,1457% da variabilidade dos dados.

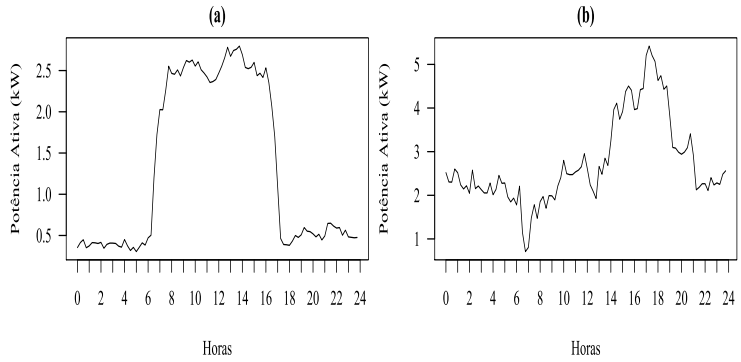

Figura 34 Componentes independentes obtidos com o algoritmo FastICA para as séries de potência ativa da Fitotecnia, Química e Sementes, medidas em kW, de 15 em 15 minutos no dia 28/10/2014

Na Figura 35, tem-se o periodograma para cada um dos componentes independentes, obtidos na Figura 34.

Em 4.3, tem-se a matriz de mistura para este caso. A ideia aqui é comparar os departamentos considerando cada coluna separadamente, ou seja, cada componente independente, semelhante ao que foi realizado nos estudos anteriores.

$$\hat{\mathbf{A}}_3 = \begin{array}{c} \begin{matrix} (a) \uparrow & (b) \uparrow \end{matrix} \\ \begin{pmatrix} 5,3514 & 4,9195 \\ 18,3190 & 2,7043 \\ 6,0600 & 6,8626 \end{pmatrix} \end{array} \begin{array}{l} \\ \rightarrow FITOTECNIA \\ \rightarrow QUIMICA \\ \rightarrow SEMENTES \end{array} \tag{4.3}$$

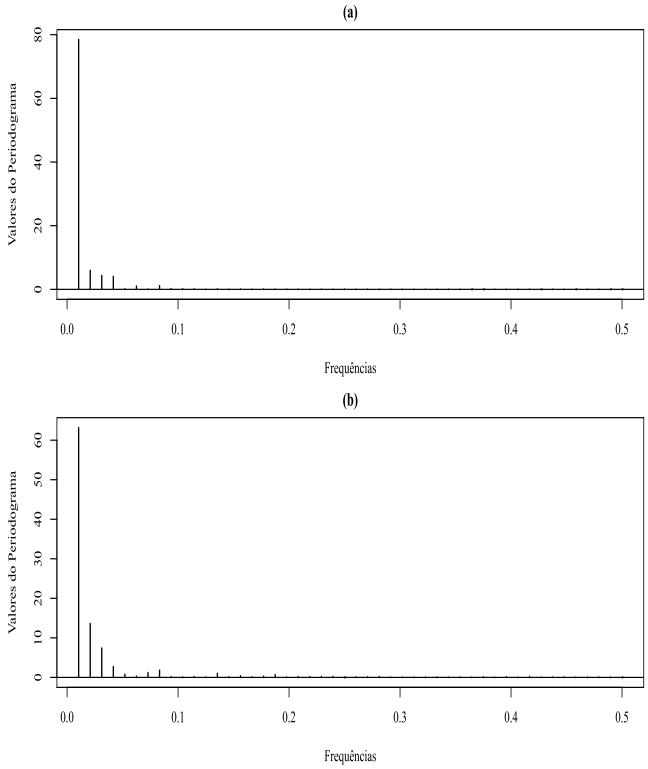

Figura 35 Periodogramas dos componentes independentes, referentes às séries de potência ativa da Fitotecnia, Química e Sementes, medidas em kW, de 15 em 15 minutos no dia 28/10/2014

Na Figura 34 (a), tem-se o componente independente (a) e observa-se que este representa de forma bastante clara o comportamento do departamento de Química. Tanto que o coeficiente para este local foi o maior valor dentre os três em estudo para o componente (a), como pode ser visto na primeira coluna da matriz de mistura 4.3.

Na Figura 35 (a), pode-se observar uma periodicidade de ordem 48 (frequência 0,0208) para o componente (a), indicando regularidade da variável em análise de 12 em 12 horas. Para confirmar essa periodicidade, fez-se o teste de Fisher. As estatísticas obtidas foram $g = 0,3380$ e $z_{0,05} = 0,1383$, como $g > z$, a série

apresenta sazonalidade de ordem 48, a 5% de significância. Portanto a variável analisada apresenta uma regularidade de 12 em 12 horas.

Pode-se ver também pela matriz que todos os coeficientes da primeira coluna são positivos, ou seja, seguem o comportamento de potência como é apresentado pela série do componente (a).

O componente (b) pode ser visto na Figura 34 (b). Veja que o comportamento está mais associado às características do setor de Sementes, e depois ao setor de Fitotecnia. Observa-se pelos coeficientes que se encontram na segunda coluna da matriz de misturas (4.3), que realmente isso ocorre.

Observou-se pela Figura 34 (b) que o componente (b) também tem uma periodicidade de ordem 48 (frequência 0,0208), indicando regularidade da variável em análise de 12 em 12 horas. Para testar a existência do efeito sazonal no período indicado, aplicou-se também o teste de Fisher. Chegou-se às estatísticas $g = 0,4143$ e $z_{0,05} = 0,1383$, como $g > z$, a série apresenta sazonalidade de ordem 48, a 5% de significância.

Observa-se que esses componentes não trouxeram grandes contribuições. Porém, mostra-se que a ICA pode separar de forma eficiente os comportamentos adjacentes de um conjunto de séries de dados.

5 CONCLUSÃO

De um modo geral, os resultados apresentados nesta tese foram satisfatórios. Por meio do conjunto de métodos estatísticos propostos, pôde-se responder a questões até então não respondidas acerca de variáveis energéticas na UFLA.

A nova maneira de construir a matriz de proximidades, em que o elemento para ser vizinho não necessariamente precisa estar próximo fisicamente, possibilitou que os índices de Moran e Geary, que até então eram usados somente na estatística espacial, fossem utilizados em um novo contexto.

A proposta de combinar os índices de Moran e Geary com a AA, por intermédio da matriz sugerida, se mostrou eficiente, pois possibilita além da determinação do número "ótimo" de grupos, como é feito normalmente pelos métodos presentes na literatura, a obtenção de outros agrupamentos que poderiam ser utilizados, pois possuem índices de dependência significativos. Assim, o pesquisador tem liberdade na escolha dos agrupamentos, podendo considerar a alternativa "ótima" ou optar por qualquer outro agrupamento pertinente segundo os índices de Moran e Geary.

Além disso, por meio de exemplos, os índices de Moran e Geary mostraram que grupos diferentes dos indicados pelo dendrograma levam a resultados esperados, que é de ausência de dependência ou dissimilaridades entre os elementos dos grupos.

Dentre os departamentos/setores analisados com os dados de 2010, o RU e o CIN são os que necessitam maior demanda de potência, sendo que o CIN possui comportamento constante durante todo o período do dia, com um leve crescimento durante a manhã e tarde. Enquanto o RU tem um pico de consumo por volta das 12 horas, os outros departamentos/setores têm baixo consumo, com apenas um fato relevante, que é a CANTINA, a qual possui uma maior demanda de potência entre

as 6 e 7 horas.

Na análise dos departamentos/setores, Fitotecnia, Química e Sementes, pode-se observar que, no período das 7 às 17 horas, o DQI possui o maior consumo de energia elétrica dentre os três. Já no período das 0 às 7 e das 17 às 0 horas há uma maior demanda de potência por parte do Setor de Sementes.

Para os dias da semana, avaliados em 2013, os agrupamentos encontrados levaram a conclusões esperadas, com o sábado e domingo ficando em grupos separados dos demais dias. Considerando que os períodos da manhã na segunda-feira e a tarde na sexta-feira possuem menores demandas de potência, outro agrupamento pode ser formado, com esses dias em grupos distintos. Isso ocorre pois, nas segundas e sextas-feiras, há alunos retornando e saindo da universidade.

De forma geral observou-se que a periodicidade da potência ocorre de 12 em 12 horas, ou seja, o comportamento do consumo de energia tende a se repetir no período mencionado. Isso se mostra coerente com a realidade da universidade.

No período de ponta, a demanda de potência registrada mensalmente, se mostrou crescente desde 1995, sendo que o aumento maior está ocorrendo nos últimos quatro anos, ou seja, a partir de 2010.

No período de 2010 a 2013, os meses segundo a variável DPRHP foram agrupados de forma diferente pelos índices de Moran e Geary. Para o índice de Moran o agrupamento de melhor qualidade foi: jan; nov, dez; fev, jul e mar, abr, mai, jun, ago, set, out. Já para o índice de Geary os grupos encontrados foram: jan; fev; jul; abr; nov; dez; jun; mar, out e mai, ago, set.

Para o agrupamento formado pelo índice de Moran, os máximos obtidos pelos grupos foram: jan = 924 kW, nov, dez = 1456 kW, sendo novembro o maior valor, fev, jul = 1190 kW, sendo fevereiro o maior valor e mar, abr, mai, jun, ago, set, out = 1218 kW, sendo que março atingiu o maior valor de potência.

Considerando o agrupamento formado pelo índice de Geary, os máximos encontrados nos grupos foram: jan = 924 kW, fev = 1190 kW, jul = 1106 kW, abr = 1100 kW, nov = 1456 kW, dez = 1386 kW, jun = 1092 kW, mar, out = 1218 kW, sendo março o maior valor e mai, ago, set = 1148 kW, sendo este valor alcançado no mês de maio.

No horário fora de ponta, em que a demanda de energia da universidade é maior do que no horário de ponta, os valores de potência estão também com tendência de crescimento no período analisado, de 1995 a 2013.

No horário fora de ponta, o índice de Moran indicou como melhor agrupamento dos meses set; abr; out; mai, ago; jan, jun, jul e fev, mar, nov, dez. Para o índice de Geary, o agrupamento escolhido foi set; abr; jul; out; dez; mai, ago; jan, jun e fev, mar, nov.

Levando em consideração o agrupamento formado pelo índice de Moran, os máximos obtidos pelos grupos foram: set = 1414 kW, abr = 1722 kW, out = 1806 kW, mai, ago) = 1512 kW, valor obtido em maio, jan, jun, jul = 1470 kW, valor observado em janeiro e fev, mar, nov, dez = 1960 kW, valor alcançado no mês de dezembro.

Para o agrupamento formado pelo índice de Geary, os máximos encontrados nos grupos foram: set = 1610 kW, abr = 1722 kW, jul = 1414 kW, out = 1806 kW, dez = 1960 kW, mai, ago = 1512 kW, observado em maio, jan, jun = 1470 kW, valor observado em janeiro e fev, mar, nov = 1946 kW, encontrado no mês de novembro.

A importância de se utilizar os valores máximos dos grupos encontrados para os meses, é que esses valores poderão contribuir para futuras contratações de demanda de potência. A administração da universidade poderá fazer contratos para a demanda de ponta e fora de ponta, considerando os agrupamentos e cada

máximo obtido por grupo.

As metodologias aqui propostas poderão contribuir com a UFLA, por exemplo, reduzindo as penalizações nas contas de energia elétrica por ultrapassagem de demanda contratada. Os métodos de forma integrada poderão servir para se fazer um planejamento para os próximos anos.

Para finalizar, pode-se concluir que os métodos ICA, AA e índices de Moran e Geary, podem em conjunto ser uma maneira eficiente de monitorar a energia elétrica em uma universidade ou numa empresa qualquer.

5.1 Trabalhos futuros

Como continuidade deste trabalho pretende-se:

a) ampliar as análises, considerando um maior número de departamentos/setores da Universidade Federal de Lavras;

b) realizar as mesmas análises com outras variáveis energéticas da UFLA;

c) comparar o método proposto nesta tese com outros critérios para escolha do número de grupos em um dendrograma;

d) aplicar em outros conjuntos de dados a combinação de métodos aqui propostos.

REFERÊNCIAS

ANSELIN, L. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: FISCHER, M.; SCHOLTEN, H. J.; UNWIN, D. (Ed.). **Spatial analytical perspectives on Gis**. London: Taylor e Francis, 1996. p.111-125.

ARANHA, F. Autocorrelação espacial na área de loja de supermercados nos municípios paulistas: mensuração por meio do Índice de Geary. **Revista de Administração de Empresas**, São Paulo, v.39, n.4, p.38-45, out./dez. 1999.

ASSUNÇÃO, R. M. **Estatística espacial com aplicações em epidemiologia, economia e sociologia**. São Carlos: Universidade Federal de São Carlos, 2001.

BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**. Essex: Longman Scientific, 1995.

BARROSO, L. P.; ARTES, R. Análise multivariada. In: SIMPÓSIO DE ESTATÍSTICA APLICADA A EXPERIMENTAÇÃO AGRONÔMICA, 10.; REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 48., 2003, Lavras. **Minicursos...** Lavras: Editora da UFLA, 2003.

BELL, A. J.; SEJNOWSKI T.J. An information-maximization approach to blind separation and blind deconvolution. **Neural Computation**, Cambridge, v. 7, n. 6, p. 1129-1159, Apr. 1995.

BUSSAB, W. O.; MIAZAKI, É. S.; ANDRADE, D. F. **Introdução à análise de agrupamentos**. São Paulo: ABE, 1990.

CALINSKI, T.; HARABASZ, J. A Dendrite Method for Cluster Analysis. **Communications in Statistics**, New York, v. 3, n. 1, p. 1-27, Jun. 1974.

CARDOSO, J. F.; SOULOUMIAC, A. Blind beamforming for non Gaussian signals. **IEE Proceedings - Part F**, London, v. 140, n. 6, p. 362-370, Dec. 1993.

CASELLA, G.; BERGER, R. L. **Inferência estatística**. São Paulo: Cengage Learning, 2010.

CLIFF A. D.; ORD, K. **Spatial processes**: models and applications. London: Pion, 1981.

COMON, P. Independent component analysis, a new concept? **Signal Processing**, New York, v. 36, n. 3, p. 287-314, Apr. 1994.

CRESSIE, N. A. C. **Statistics for spatial data**. New York: John Wiley & Sons, 1993.

DRUCK, S. et al. **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2004.

FARIA, P. N. **Avaliação de métodos para determinação do número ótimo de clusters em estudo de divergência genética entre acessos de pimenta**. 2009. 54 p. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, 2009.

FELIX, F. N. **Aplicando bootstrap para determinação de intervalos de confiança para o número de grupos no procedimento hierárquico aglomerativo de Ward**. 2004. 108 p. Dissertação (Mestrado em Estatística) - Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

FERREIRA, D. D. **Análise de distúrbios elétricos em sistemas de potência**. 2010. 210 p. Tese (Doutorado em Engenharia Elétrica) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

FERREIRA, D. F. **Estatística multivariada**. Lavras: Editora da UFLA. 2008.

FERREIRA, L. M. **Mapeamento dos casos de dengue na cidade de Lavras-MG, no período de 2007 - 2010**. 2012. 82 p. Dissertação (Mestrado em Modelagem de Sistemas Biológicos) - Universidade Federal de Lavras, Lavras, 2012.

FERREIRA, R. V. **Previsão de demanda: um estudo de caso para o sistema interligado nacional**. 2006. 116 p. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal de Minas Gerais, Belo Horizonte, 2006.

GEARY, R. C. The Contiguity Ratio and Statistical Mapping. **The Incorporated Statistician**, Oxford, v. 5, n. 3, p. 115-145, Nov. 1954.

GRIFFITH, D.A. **Spatial Autocorrelation:** A primer. Association of American Geographers, 1987.

GUILHON, D.; BARROS, A. K.; MEDEIROS, E. ECG data compression by independent component analysis. In: IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, 2005, Mystic. **Procedings...** Mystic: [s.n.], 2005. p. 189-193.

HYVÄRINEN, A.; OJA, E. A fast fixed-point algorithm for independent component analysis. **Neural Computation**, Cambridge, v. 9, n. 7, p. 1483-1492, 1997.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. **Neural Networks**, New York, v. 13, n. 4-5, p. 411-430, May/June 2000.

HYVÄRINEN, A.; KARHUNEN, J.; OJA, E. **Independent component analysis**. New York: John Wiley & Sons, 2001.

JESUS, C. D. de. **Monitoramento de parâmetros relacionados à energia elétrica em ambiente universitário atípico:** um estudo de caso. Dissertação (Mestrado em Engenharia Agrícola) - Universidade Federal de Lavras, Lavras, 2011.

LEITE, I.C.C.; SÁFADI, T.; CARVALHO, M.L.M. Evaluation of seed radiographic images by independent component analysis and discriminant analysis. **Seed Science and Technology**, Zurich, v. 41, n. 2, p. 235-244, Aug. 2013.

MANUEL, L. **Modelos de regressão linear com efeitos espaciais na análise da mortalidade infantil**. 2011. 82 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2011.

MARTINS, M. D. **Monitoramento da energia elétrica e gerenciamento pelo lado do consumidor**. 2008. 151 p. Dissertação (Mestrado em Ciências de Engenharia Elétrica) - Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, 2008.

MARTINS, M. R. F. O.; PEDRO, S.; ROSA, S. **Escolha do número de grupos e validação da solução em análise classificatória: da teoria à prática**. Lisboa: Instituto Superior de Estatística e Gestão de Informação, 2004. Disponível em: <http://hdl.handle.net/10362/7686>. Acesso em: 19 set. 2014.

MATLAB for windows users guide. [S.l.]: The Math Works, 2011.

MANLY, B. F. J. **Métodos estatísticos multivariados:** uma introdução. 3. ed. Porto Alegre: Bookman, 2008.

MILLINGAN, G. W. An algorithm for generating artificial test clusters. **Psychometrika**, Williamsburg, v. 50, n. 1, p. 123-127, Mar. 1985.

MILLINGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, Williamsburg, v. 50, n. 2, p. 159-179, June 1985.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Editora da UFMG, 2005.

MOJENA, R. Hierarchical grouping methods and stopping rules: an evaluation. **Computer Journal**, London, v. 20, n. 4, p. 359-363, 1977.

MORAN, P. A. F. Notes on continuons stochastic phenomena. **Biometrika**, London, v.37, n.1, p. 17-23, June 1950.

MORETO, F. A. de L. **Análise de componentes independentes aplicada à separação de sinais de áudio**. 2008. 83 p. Dissertação (Mestrado em Engenharia Elétrica) - Universidade de São Paulo, São Paulo, 2008.

NASCIMENTO, M.; SÁFADI, T.; SILVA, F. F. Aplicação da análise de agrupamento de dados de expressão gênica temporal a dados em painel. **Pesquisa Agropecuária Brasileira**, Brasília, v.46, n.11, p.1489-1495, Nov. 2011.

OHTSUKA, Y.; OGA, T; KAKAMU, K. Forecasting electricity demand in Japan: A bayesian spatial autoregressive ARMA approach. **Computational Statistics & Data Analysis**, Amsterdam, v. 54, n. 11, p. 2721-2735, Nov. 2010.

PECK, R.; FISCHER, L.; NESS, J. V. Aproximate confidence intervals for the number of cluster. **Journal of the American Statistical Association**, New York, v. 84, n. 405, p. 184-191 Mar. 1989.

PRADO, J.R. do. **Modelos para demanda e consumo de energia elétrica utilizando séries temporais na Universidade Federal de Lavras**. 2011. 114 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2011.

PRIESTLEY, M. B. **Spectral analysis and time series**. London: Academic Press, 1989.

R CORE TEAM **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2014.

RENCHER, A. C. **Methods of multivariate analysis**. 2. ed. New York: John Wiley & Sons, 2002.

SÁFADI, T. Using independent component for clustering of time series data. **Applied Mathematics and Computation**, New York, v. 243, p. 522-527, 2014.

SHARMA, S. **Applied multivariate techniquies**. New York: John Wiley & Sons, 1996.

SOKAL, R.R.; ODEN, N.L. Spatial autocorrelation in biology. **Biological Journal of the Linnean Society**, London, v. 10, n. 2, p. 199-228, June 1978.

VILLAMAGNA, M. R. **Seleção de modelos de séries temporais e redes neurais artificiais na previsão de consumo e demanda de energia elétrica**. 2013. 117 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2013.

WALLER, L. A.; GOTWAY, C. A. **Applied spatial statistics for public health data**. New York: John Wiley & Sons, 2004.

ANEXOS

ANEXO A - Agrupamentos e tabelas

a) Agrupamento 1: CIN, DCC-T, DAE e CANTINA, DCC-PP, EPAMIG, DEX, RU.

b) Agrupamento 2: CANTINA, DCC-T, DEX, RU e CIN, DAE, DCC-PP, EPAMIG.

Tabela 14 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,3381	0,2970	1,1709	0,2007
Agrupamento 2	-0,3291	0,1664	1,1627	0,1677

a) Agrupamento 1: CANTINA, DAE, EPAMIG, DCC-PP e DEX, DCC-T, CIN, RU.

b) Agrupamento 2: CANTINA, DCC-T, DEX, EPAMIG, RU e CIN, DAE, DCC-PP.

Tabela 15 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,0716	0,7263	0,9377	0,2662
Agrupamento 2	-0,2117	0,5073	1,0602	0,5100

a) Agrupamento 1: Seg, Qua, Sab e Ter, Qui, Sex, Dom.

b) Agrupamento 2: Seg, Qui, Dom e Ter, Qua, Sex, Sab.

Tabela 16 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,3909	0,4015	1,1922	0,3846
Agrupamento 2	-0,3875	0,4499	1,1893	0,4285

a) Agrupamento 1: Seg, Qua, Sab e Ter, Qui, Sex, Dom.

b) Agrupamento 2: Seg, Ter, Dom e Qua, Qui, Sex, Sab.

Tabela 17 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,1465	0,6215	1,0375	0,4918
Agrupamento 2	-0,3537	0,5043	1,1603	0,5002

a) Agrupamento 1: Seg, Qua, Sab e Ter, Qui, Sex, Dom.

b) Agrupamento 2: Seg, Ter, Dom e Qua, Qui, Sex, Sab.

Tabela 18 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,1565	0,5998	1,0467	0,4687
Agrupamento 2	-0,4100	0,1366	1,2086	0,1353

a) Agrupamento 1: jan, abr, ago, out, nov e fev, mar, mai, jun, jul, set, dez.

b) Agrupamento 2: jan, mar, jun, dez e fev, abr, mai, jul, ago, set, out, nov.

Tabela 19 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,2095	0,0827	1,1087	0,0823
Agrupamento 2	-0,1680	0,3041	1,0706	0,3063

a) Agrupamento 1: jan, fev, mar, mai, jun, nov e abr, jul, ago, set, out, dez.

b) Agrupamento 2: jan, abr, jun, set, dez e fev, mar, mai, jul, ago, out, nov.

Tabela 20 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,1120	0,5755	1,0194	0,5719
Agrupamento 2	-0,1665	0,2752	1,0678	0,2949

a) Agrupamento 1: jan, fev, mai, set, out e mar, abr, jun, jul, ago, nov, dez.

b) Agrupamento 2: jan, mar, jun, dez e fev, abr, mai, jul, ago, set, out, nov.

Tabela 21 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,1814	0,3035	1,0829	0,2969
Agrupamento 2	-0,2403	0,0051	1,1369	0,0053

a) Agrupamento 1: jan, mai, set, out, nov e fev, mar, abr, jun, jul, ago, dez.

b) Agrupamento 2: jan, mar, ago, out, dez e fev, abr, mai, jun, jul, set, nov.

Tabela 22 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para dois cortes que não são obtidos no dendrograma

Corte	Moran	valor-p	Geary	valor-p
Agrupamento 1	-0,1896	0,0858	1,0922	0,0572
Agrupamento 2	-0,1499	0,4837	1,0540	0,4853

ANEXO B - Dendrogramas e tabelas

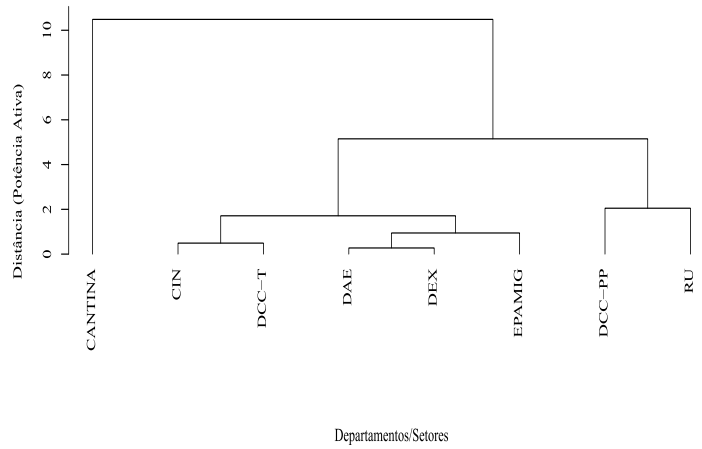

Figura 36 Dendrograma para agrupamento de alguns setores/departamentos da UFLA, segundo os pesos do componente independente (a), e considerando a variável potência ativa da UFLA, medida em kW, de 15 em 15 minutos no dia 12/08/2010

Tabela 23 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma do componente (a)

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,0378	0,1144	0,3676	0,1164
Segundo	0,3803	0,0214	0,0679	0,0023
Terceiro	-0,0107	0,3110	0,0532	0,0091
Quarto	0,0365	0,3244	0,0118	0,0009
Quinto	0,0574	0,3652	0,0039	0,0026
Sexto	0,0755	0,3013	0,0019	0,0177

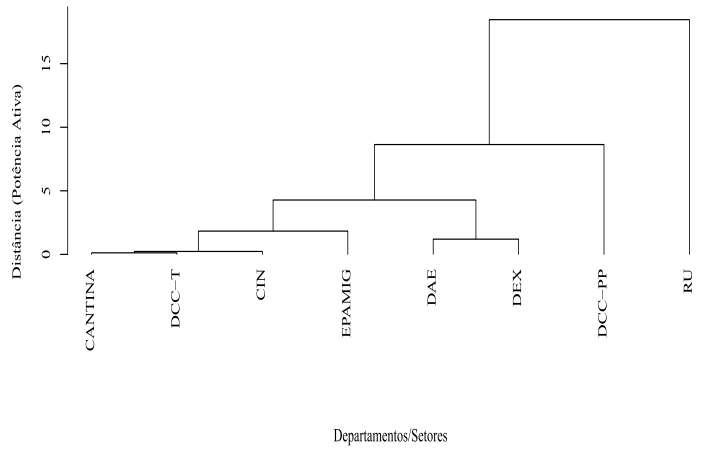

Figura 37 Dendrograma para agrupamento de alguns setores/departamentos da UFLA, segundo os pesos do componente independente (b), e considerando a variável potência ativa da UFLA, medida em kW, de 15 em 15 minutos no dia 12/08/2010

Tabela 24 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma do componente (b)

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,0645	0,0833	0,2742	0,0671
Segundo	0,2385	0,0202	0,0993	0,0239
Terceiro	0,3355	0,0372	0,0145	0,0016
Quarto	0,3663	0,0601	0,0054	0,0008
Quinto	0,0563	0,3983	0,0004	0,0088
Sexto	0,0513	0,5532	0,0001	0,0165

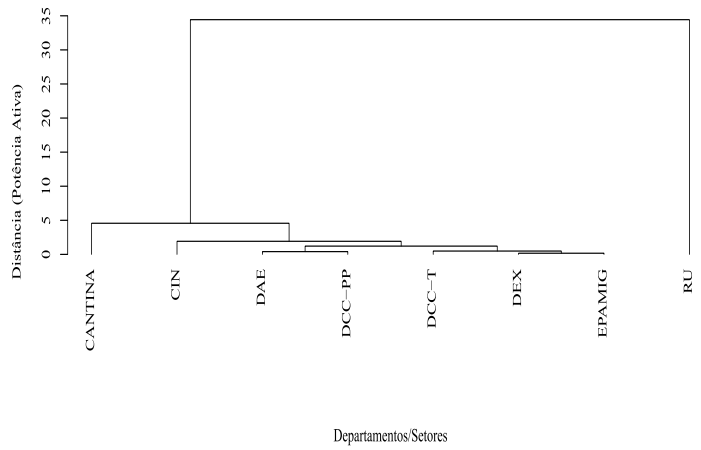

Figura 38 Dendrograma para agrupamento de alguns setores/departamentos da UFLA, segundo os pesos do componente independente (c), e considerando a variável potência ativa da UFLA, medida em kW, de 15 em 15 minutos no dia 12/08/2010

Tabela 25 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma do componente (c)

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,0645	0,0815	0,2742	0,0620
Segundo	0,1841	0,0029	0,0067	0,0175
Terceiro	0,2093	0,0082	0,0033	0,0138
Quarto	0,2124	0,0010	0,0006	0,0010
Quinto	0,2166	0,0310	0,0003	0,0022
Sexto	0,1616	0,3392	0,0001	0,0172

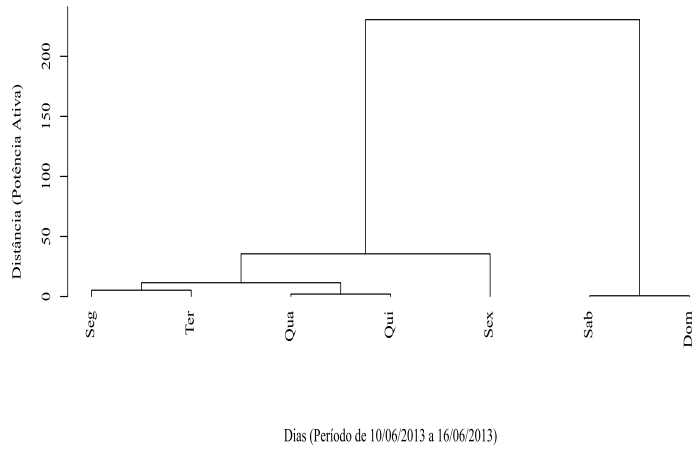

Figura 39 Dendrograma para o agrupamento dos dias da semana, segundo os pesos do componente independente (a), e considerando a variável potência ativa da UFLA, medida em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013

Tabela 26 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma do componente (a)

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,9812	0,0230	0,0161	0,0216
Segundo	1,1425	0,0045	0,0026	0,0053
Terceiro	1,1450	0,0050	0,0004	0,0052
Quarto	1,4364	0,0466	0,0004	0,0036
Quinto	2,4623	0,0233	0,0000	0,0230

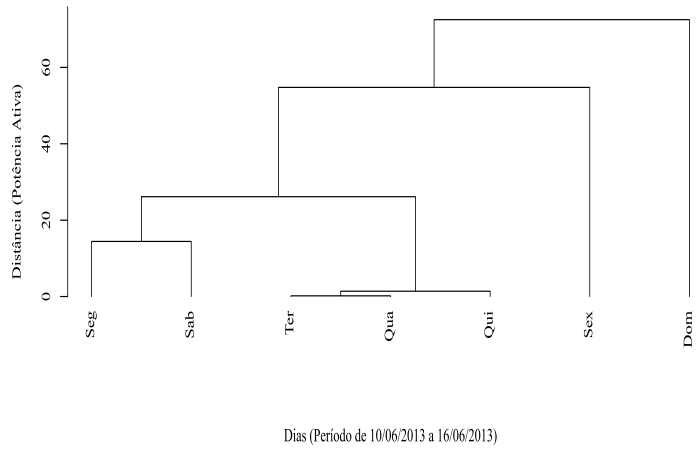

Figura 40 Dendrograma para o agrupamento dos dias da semana, segundo os pesos do componente independente (b), e considerando a variável potência ativa da UFLA, medida em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013

Tabela 27 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma do componente (b)

Corte	Moran	valor-p	Geary	valor-p
Primeiro	-0,0061	0,0067	0,5182	0,0707
Segundo	-0,0395	0,2665	0,1750	0,0236
Terceiro	0,1276	0,2490	0,0318	0,0030
Quarto	0,1204	0,1598	0,0005	0,0133
Quinto	0,1112	0,4008	0,0000	0,0235

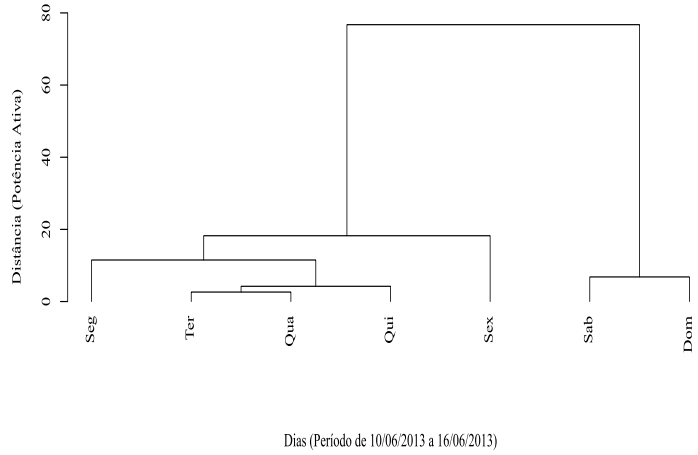

Figura 41 Dendrograma para o agrupamento dos dias da semana, segundo os pesos do componente independente (c), e considerando a variável potência ativa da UFLA, medida em kW, de 15 em 15 minutos no período de 10/06 a 16/06/2013

Tabela 28 Índices de Moran e Geary e seus respectivos testes de permutação aleatória, para cada um dos cortes feitos no dendrograma do componente (c)

Corte	Moran	valor-p	Geary	valor-p
Primeiro	0,9812	0,0232	0,0161	0,0233
Segundo	1,1331	0,0047	0,0226	0,0049
Terceiro	1,1946	0,0172	0,0095	0,0022
Quarto	0,4069	0,1278	0,0052	0,0135
Quinto	0,4599	0,2237	0,0024	0,0218