UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL

Aline Schaefer da Rosa

Análise de agrupamentos aplicada à obtenção de modelos de referência para estudos de desempenho térmico de edificações

Florianópolis
2019

Aline Schaefer da Rosa

Análise de agrupamentos aplicada à obtenção de modelos de referência para estudos de desempenho térmico de edificações

Tese submetida ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal de Santa Catarina para a obtenção do título de doutor em Engenharia Civil.
Orientador: Prof. Enedir Ghisi, PhD.

Florianópolis

2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Rosa, Aline Schaefer
Análise de agrupamentos aplicada à obtenção de modelos de referência para estudos de desempenho térmico de edificações / Aline Schaefer Rosa ; orientador, Enedir Ghisi, 2019.
175 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Civil, Florianópolis, 2019.

Inclui referências.

1. Engenharia Civil. 2. Engenharia civil. 3. Desempenho térmico em edificações. 4. Análise de agrupamentos. 5. Habitação de interesse social. I. Ghisi, Enedir . II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Engenharia Civil. III. Título.

Aline Schaefer da Rosa

Análise de agrupamentos aplicada à obtenção de modelos de referência para estudos de desempenho térmico de edificações

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Aldomar Pedrini, PhD.
Universidade Federal do Rio Grande do Norte

Prof. Eduardo Grala da Cunha, Dr.
Universidade Federal de Pelotas

Prof. Roberto Lamberts, PhD.
Universidade Federal de Santa Catarina

Profª Michele Fossati, Drª
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Engenharia Civil.

Documento assinado digitalmente
Poliana Dias de Moraes
Data: 13/01/2020 16:54:35-0300
CPF: 613.571.209-82

Profª Poliana Dias de Moraes, Drª
Coordenadora do Programa

Enedir Ghisi:81375018949

Assinado de forma digital por Enedir Ghisi:81375018949
DN: cn=Enedir Ghisi:81375018949, ou=UFSC - Universidade Federal de Santa Catarina, o=ICPEdu
Dados: 2020.01.13 09:31:08 -03'00'

Prof. Enedir Ghisi, PhD.
Orientador

Florianópolis, 2019.

Dedico este trabalho a toda a turminha do /*lunchtrain*.

AGRADECIMENTOS

O caminho percorrido para a concretização desse estudo não foi fácil e há muito a agradecer as pessoas que me ajudaram na caminhada.

Em primeiro lugar, agradeço a Deus pelo dom da vida e por sua presença constante e amorosa nos momentos de dificuldade e fraqueza. Não há força mais poderosa que a fé e a oração para nos manter perseverantes mesmo nos momentos mais difíceis.

Agradeço também ao professor Enedir, pela oportunidade e pela forma serena como conduziu as orientações. Agradeço por toda a dedicação, correções, coautorias, pelas reuniões e conversas ao longo desses anos. Por ser nossa melhor referência. Ah! E pela enooorme paciência!

Agradeço à banca, composta pelos professores Aldomar Pedrini, Eduardo Grala da Cunha, Roberto Lamberts e Michele Fossati, pela avaliação deste volume e contribuições.

Aos meus irmãos, pelo enorme carinho e apoio. Aos meus pais, pela presença constante, por cada abraço e apoio ao longo dos anos (apesar dos insistentes "isso não acaba nunca, não?"). Aos meus sogros, pela compreensão quanto à distância e ausência. Ao meu marido Flavio, por aceitar comigo o desafio dessa caminhada. Ao meu amado filho Davi, por todo carinho e amor, por cada sorriso e cada lembrança que me ajudou a seguir em frente.

Aos meus amigos-irmãos TDFs, por estarem sempre perto, mesmo que distantes. "Não importa onde vocês estiverem, vocês sempre estarão lá! "

A todos os colegas do LabEEE, os que passaram, os que chegaram e os que ainda estão. Vocês têm sido uma verdadeira família. Em especial, à turminha do RU, pelos almoços divertidos, pela terapia em grupo, troca de conhecimento e amizade. Jamais teria chegado até aqui se não fosse por vocês. Não poderia me esquecer das "Lab-mães" e "Lab-babies", por compartilhar os momentos de descontração e também de desespero. Aos queridos amigos Arthur e Laiane, que o destino quis levar para longe, mas que vão morar sempre no meu coração.

Agradeço à Pri, por todo apoio, ajuda, conversa, dicas e etc.

Por fim, agradeço à CAPES, por financiar meus estudos.

“Nobody knows where you are,
how near or how far.
Shine on, you crazy diamond”
(WRIGHT, WATERS and GILMOUR, 1975)

RESUMO

Estudos têm utilizado modelos de referência para obter indicadores de quais medidas de conservação de energia seriam mais eficientes quando aplicadas a um estoque de edificações. A análise de agrupamentos tem sido frequentemente utilizada para obtenção desses modelos. Essa análise é composta por diversos parâmetros, cuja combinação leva a resultados distintos. Entretanto, não se encontra na literatura estudos que analisem quais combinações levam a melhores resultados no processo de agrupamento. Nesse contexto, este trabalho tem como objetivo o desenvolvimento de um método de aplicação da análise de agrupamentos visando a obtenção de modelos de referência, para uso em estudos de desempenho térmico de edificações, a partir de diferentes configurações de clusterização. Para isso, propôs-se um método dividido em seis etapas e implementado em um estudo de caso de habitações de interesse social de Florianópolis. Inicialmente, uma matriz de dados foi formada com dados referentes à geometria das habitações. Na sequência, submeteu-se a matriz de dados a três diferentes tratamentos estatísticos, formando cinco matrizes distintas. Na terceira etapa, a combinação das cinco matrizes com cinco medidas de similaridade e cinco algoritmos de partição deram origem a 105 métodos de agrupamentos. Modelos de referência foram designados a cada agrupamento a partir da habitação mais próxima ao centroide na quarta etapa. Os agrupamentos formados foram submetidos a dois processos de validação: interna e relativa. Foi possível, assim, definir o método cuja partição resultou na melhor formação de agrupamento e modelo de referência. Na última etapa, os agrupamentos e modelos formados foram descritos a partir de suas características geométricas e desempenho térmico. O método que combinou o algoritmo K-means e dados tratados a partir da ponderação dos fatores e detecção de objetos atípicos foi selecionado como formação mais adequada. Obteve-se dois agrupamentos, para os quais determinou-se dois modelos de referência. O primeiro representa uma habitação com 64m^2, sala independente e com orientação solar norte e leste, e três dormitórios, voltados majoritariamente a oeste. O segundo modelo caracteriza-se por uma habitação de 37m^2, sala e cozinha conjugadas com orientação oeste, e dois dormitórios com orientação leste. Testes de hipótese mostraram que os grupos diferem para a maioria dos indicadores de desempenho, especialmente quanto ao resfriamento dos dormitórios. Algumas configurações apresentaram resultados menos satisfatórios em relação às demais. O algoritmo Ligação Simples não obteve boa formação dos agrupamentos independente da configuração. A correlação de Pearson também não apresentou bons resultados. Os algoritmos Ligação Completa, Ward e K-means apresentaram, de forma geral, boas formações assim como as medidas City-block, distância euclidiana e distância euclidiana quadrada. A ponderação dos fatores foi o tratamento de dados que mais contribuiu para a obtenção de boas soluções de agrupamento. Concluiu-se, ao final do estudo, que o método proposto é capaz de identificar a combinação de parâmetros que resultam no melhor método de agrupamento e que é uma técnica aplicável à determinação de modelos de referência de edificações.

Palavras-chave: Desempenho térmico de edificações. Análise de agrupamento. Habitações unifamiliares. Estatística aplicada.

ABSTRACT

Studies have used reference models to obtain indicators of which energy conservation measures would be most effective when applied to a building stock. Cluster analysis has often been used to obtain these models. This analysis consists of several parameters, the combination of which leads to different results. However, no studies in the literature considers the different configurations in cluster analyses. In this context, this thesis aims to develop a method of applying cluster analysis to obtain reference models for use in thermal performance studies of buildings from different clustering configurations. For this, a method divided into six steps was proposed and implemented in a case study of social housing in Florianópolis. Initially, a data matrix was formed with data regarding the geometry of the dwellings. Subsequently, the data matrix was submitted to three different statistical treatments, forming five distinct matrices. In the third step, the combination of the five matrices with five similarity measures and five partition algorithms gave rise to 105 clustering methods. Reference models were assigned to each cluster from the nearest house to the centroid in the fourth step. The clusters were submitted to two validation processes: internal and relative. It was thus possible to define the method whose partition resulted in the best cluster formation and reference model. In the last step, the clusters and models provided were described from their geometric characteristics and thermal performance. The method that combined the K-means algorithm and data treated by the weighting factor and detection of atypical objects was selected as the most appropriate formation. Two clusters were obtained, for which two reference models were determined. The first represents a 64m^2 dwelling, north-east solar orientation living room, and three bedrooms, mostly west facing. The second model is characterized by a 37m^2 dwelling, combined living room and kitchen with west orientation, and two bedrooms, mostly east facing. Hypothesis tests have shown that clusters differ for most performance indicators, especially for bedrooms cooling. Some cluster configurations have shown less satisfactory results than the others. The Single Linkage algorithm has not performed good cluster formation regardless of the configuration. The Pearson correlation has not shown good results either. The algorithms Complete Linkage, Ward and K-means resulted in good formations as well as City-block, Euclidean Distance and Squared Euclidean Distance similarity measures. The weighting factor was the data treatment that most contributed to obtain good clustering solutions. It was concluded, at the end of the study, that the proposed method is able to identify the combination of parameters that results in the best clustering method and that this is an applicable technique to obtain reference building models.

Keywords: Thermal performance of buildings. Cluster analysis. Single-family dwellings. Applied statistics.

LISTA DE FIGURAS

Figura 1 - Esquema de um método adotado para realização do balanço térmico de edificações a partir do uso de modelos de referência. 24
Figura 2 - Métodos para obtenção de modelos de referência. 26
Figura 3 - Quantidade de publicações por ano e fonte com o termo *reference building*..... 27
Figura 4 - Estágios 1-3 do diagrama de decisão segundo Hair et al. (2009). 35
Figura 5 - Estágios 4-6 do diagrama de decisão segundo Hair et al. (2009). 36
Figura 6 – Medidas de similaridade: (a) Distância City-block. (b) Distância Euclidiana. (c) Distância Chebyshev. 38
Figura 7 - Construção da matriz de similaridade com a distância Euclidiana Quadrada. 39
Figura 8 - Representação do processo de agrupamento com o método hierárquico. 40
Figura 9 – Representação de alguns algoritmos de partição: (a) Ligação Média. (b) Ligação Simples. (c) Ligação Completa. 41
Figura 10 - Representação do processo de agrupamento com o método não hierárquico..... 42
Figura 11 - Fluxograma do método proposto. 52
Figura 12 - Composição dos materiais do envelope. 56
Figura 13 - Padrão de operação das aberturas. 59
Figura 14 - Composição dos materiais do envelope. 60
Figura 15 - Formação das Matrizes de dados. 64
Figura 16 – Comparação entre medidas de proximidade e de correlação. 71
Figura 17 - Representação do algoritmo de partição: Ligação Simples. 73
Figura 18 - Procedimentos de partição: Ligação Simples. 73
Figura 19 - Representação do algoritmo de partição: Ligação Completa. 74
Figura 20 - Representação do algoritmo de partição: Centroide 74
Figura 21 - Representação do algoritmo de partição: Ward. 75
Figura 22 - Procedimentos de partição: K-means. 76
Figura 23 - Exemplo de dendograma obtido a partir da técnica hierárquica. 79
Figura 24 – Fluxograma do processo de determinação do modelo de referência. 83
Figura 25 – Representação das medidas *intra-cluster* e *inter-cluster*. 84
Figura 26 – Procedimentos para validação interna. 88
Figura 27 – Localização dos pontos de levantamento: (a) Vargem Grande. (b) Foz do Rio. (c) Maciço Central. (d) Jardim El Dourado (e) Jardim Aquárius. 95

Figura 28 – Resumo das características geométricas da amostra: áreas e ambientes.....97
Figura 29 – Resumo das características geométricas da amostra: relação entre paredes e aberturas.....98
Figura 30 – Representação gráfica em planta baixa de algumas habitações com *layouts* variados. 99
Figura 31 – Resumo dos indicadores de desempenho das habitações obtidos por meio de simulações computacionais. 101
Figura 32 – Valores padronizados a partir da medida *z-scores*. 103
Figura 33 - Valores de F obtidos para cada variável. 107
Figura 34 - Dendograma obtido a partir do método M6_A. 109
Figura 35 - Dendograma obtido a partir do método M19_D. 109
Figura 36 - Dendograma obtido a partir do método M10_B. 109
Figura 37 - Dendograma obtido a partir do método M10_E. 110
Figura 38 - Dendograma obtido a partir do método M19_C. 110
Figura 39 - Dendograma obtido a partir do método M3_A. 113
Figura 40 - Dendograma obtido a partir do método M4_B. 113
Figura 41 - Dendograma obtido a partir do método M20_E. 114
Figura 42 - Resultados da análise de variância. 120
Figura 43 – Mapa em árvore da frequência de ocorrência dos modelos. 126
Figura 44 - Representação gráfica em planta baixa dos modelos de maior ocorrência. 127
Figura 45 – Perspectivas da maquete eletrônica utilizadas nas simulações: (a) Vista leste e norte do modelo de referência 1. (b) Vista leste e norte do modelo de referência 2. (c) Vista oeste e sul do modelo de referência 1. (d) Vista oeste e sul do modelo de referência 2. 133
Figura 46 – Modelo de referência do agrupamento 1. 134
Figura 47 – Modelo de referência do agrupamento 2. 134
Figura 48 – Indicadores de desempenho do Modelo 1 para aquecimento e resfriamento (ºCh). 135
Figura 49 – Indicadores de desempenho do Modelo 2 para aquecimento e resfriamento (ºCh). 136
Figura 50 – Diagrama de caixas referente ao indicador de desempenho de todas as habitações da amostra, para cada agrupamento. 138

LISTA DE QUADROS

Quadro 1 - Variáveis que compõem a matriz de dados inicial. 54
Quadro 2 - Variáveis submetidas à medida de validação relativa. 63
Quadro 3 – Conjunto preliminar de métodos de clusterização. 77
Quadro 4 - Variáveis submetidas à medida de validação relativa. 89
Quadro 5 – Teste de hipóteses utilizados nas análises. 94
Quadro 6 - Métodos eliminados do estudo devido à configuração do dendograma. 114
Quadro 7 - Decisão quanto à quantidade de grupos para cada método da Matriz A. 121
Quadro 8 - Decisão quanto à quantidade de grupos para cada método da Matriz B. 121
Quadro 9 - Decisão quanto à quantidade de grupos para cada método da Matriz C. 122
Quadro 10 - Decisão quanto à quantidade de grupos para cada método da Matriz D. 122
Quadro 11 - Decisão quanto à quantidade de grupos para cada método da Matriz E. 123
Quadro 12 - Decisão quanto à quantidade de grupos para cada método. 123
Quadro 13 – Modelos de referência obtidos para todos os métodos resultantes da análise de agrupamentos. 124

LISTA DE TABELAS

Tabela 1 – Propriedades térmicas dos sistemas construtivos. 56
Tabela 2 – Padrão de uso e densidade de carga interna de equipamentos. 57
Tabela 3 – Taxas metabólicas adotadas por ambiente. 57
Tabela 4 – Padrão de ocupação. 58
Tabela 5 – Padrão de uso e densidade de carga interna de iluminação. 58
Tabela 6 – Propriedades térmicas dos novos sistemas construtivos. 61
Tabela 7 - Identificação de potenciais objetos atípicos com a medida D^2 de Mahalanobis. .. 104
Tabela 8 - Características dos objetos identificados como atípicos. 105
Tabela 9 - Programa de aglomeração do método M19_C. 111
Tabela 10 - Regra de parada do método M19_C a partir do coeficiente de aglomeração. 112
Tabela 11 - Histórico de interação a partir dos centroides dos grupos para a Matriz A. 116
Tabela 12 - Histórico de interação a partir dos centroides dos grupos para a Matriz B. 117
Tabela 13 - Histórico de interação a partir dos centroides dos grupos para a Matriz C. 117
Tabela 14 - Histórico de interação a partir dos centroides dos grupos para a Matriz D. 118
Tabela 15 - Histórico de interação a partir dos centroides dos grupos para a Matriz E. 118
Tabela 16 – Inércia global, inércia *inter-cluster* e inércia *intra-cluster* por método. 129
Tabela 17 – Índices estatísticos ponderados para quantificação de erros por método. 130
Tabela 18 – Perfil dos agrupamentos a partir das variáveis qualitativas. 132
Tabela 19 – Perfil dos agrupamentos a partir das variáveis quantitativas. 133
Tabela 20 – Valores de média, mediana e erro padrão dos grupos para cada indicador de desempenho. 136
Tabela 21 - Hipótese de igualdade quanto à distribuição de valores do indicador de desempenho ao longo das categorias de agrupamento através do teste U de Mann-Whitney (nível de significância $p<0,05$). 139
Tabela 22 -Teste de Medianas para amostras independentes ao longo das categorias de agrupamento (nível de significância $p<0,05$). 140

SUMÁRIO

1. Introdução..... 17
1.1. Objetivos..... 21
1.1.1. Objetivo geral..... 21
1.1.2. Objetivos específicos..... 21
1.2. Estrutura da tese..... 21
2. Revisão de literatura..... 23
2.1. Uso de modelos de referência em estudos de desempenho térmico em edificações..... 23
2.1.1. Conceitos gerais..... 23
2.1.2. Estudos de desempenho térmico em edificações a partir de modelos de referência..... 26
2.2. Análise de agrupamentos..... 33
2.2.1. Conceito de análise de agrupamento..... 33
2.2.2. Procedimentos..... 34
2.2.3. Estudos de desempenho térmico e energético de edificações envolvendo a análise de agrupamentos..... 43
2.3. Síntese do capítulo..... 48
3. Método..... 50
3.1. Composição do banco de dados inicial..... 53
3.1.1. Dados referentes à geometria das edificações..... 53
3.1.2. Dados referentes ao desempenho térmico das habitações..... 55
3.1.2.1. Configurações gerais dos arquivos de simulação..... 55
3.1.2.2. Modelagem da geometria..... 59
3.1.2.3. Variações na configuração dos arquivos de simulação..... 60
3.1.2.4. Composição da matriz de desempenho..... 61
3.2. Formação das matrizes de dados..... 63
3.2.1. Padronização dos dados..... 64
3.2.2. Detecção de objetos atípicos..... 65
3.2.3. Ponderação dos fatores..... 66
3.3. Aplicação da análise de *cluster*..... 68

3.3.1. Medidas de similaridade..... 68
3.3.2. Algoritmos de partição..... 71
3.3.2.1. Técnica hierárquica de agrupamento..... 72
3.3.2.2. Técnica não hierárquica de agrupamento..... 75
3.3.3. Conjunto preliminar de soluções..... 77
3.3.4. Conjunto de soluções final..... 78
3.3.4.1. Técnica hierárquica de agrupamento..... 78
3.3.4.2. Técnica não hierárquica de agrupamento..... 80
3.4. Determinação dos modelos de referência..... 82
3.5. Validação dos métodos..... 83
3.5.1. Validação: medida interna..... 84
3.5.2. Validação: medida relativa..... 88
3.6. Interpretação e caracterização dos resultados..... 93
3.6.1. Perfis a partir das características geométricas..... 93
3.6.2. Perfis a partir do desempenho térmico..... 93
4. Resultados..... 95
4.1. Composição do banco de dados inicia..... 95
4.1.1. Dados referentes à geometria das edificações..... 95
4.1.2. Dados referentes ao desempenho térmico das edificações..... 100
4.2. Formação das matrizes de dados..... 102
4.2.1. Padronização dos dados..... 102
4.2.2. Detecção de objetos atípicos..... 104
4.2.3. Ponderação dos fatores..... 106
4.3. Aplicação da análise de *cluster*..... 108
4.3.1. Formação e quantidade de agrupamentos a partir da técnica hierárquica... 108
4.3.2. Formação e quantidade de agrupamentos a partir da técnica não hierárquica..... 115
4.3.3. Conjunto final de soluções..... 121
4.4. Determinação dos modelos de referência..... 123
4.5. Validação dos métodos..... 128
4.5.1. Validação: medida interna..... 128
4.5.2. Validação: medida relativa..... 130
4.6. Caracterização do objeto final..... 131

4.6.1. Perfis a partir da geometria..... 131
4.6.2. Perfis a partir do desempenho térmico..... 135
5. Conclusão..... 142
5.1. Limitações do trabalho..... 146
5.2. Sugestões para trabalhos futuros..... 147
REFERÊNCIAS..... 149
APÊNDICE A – Matriz A..... 155
APÊNDICE B – Características dos objetos identificados como atípicos..... 157
APÊNDICE C – Resultados da análise de variância para os métodos formados a partir de técnicas não hierárquicas de agrupamento 159
APÊNDICE D – Inércia global, inércia inter-cluster e inércia intra-cluster..... 162
APÊNDICE E – Índices estatísticos ponderados para quantificação de erros de cada método de agrupamento..... 163
APÊNDICE F – Características descritivas dos modelos e agrupamentos..... 164
ANEXO A – Representação gráfica das habitações..... 170

1. Introdução

A indústria da construção civil, em especial, o setor residencial, exerce grande impacto ambiental. O consumo energético atribuído às edificações representa importante nicho de pesquisa, visto que a energia é normalmente o recurso mais gasto quando considerada toda sua vida útil (UNEP, 2011). Globalmente, 40% do consumo energético é atribuído às edificações, com o setor residencial representando aproximadamente 30% desse valor (IEA, 2018). Na Europa, o consumo de energia em relação às edificações residenciais representa aproximadamente 26% do consumo total (EC, 2016), enquanto nos EUA representa 22,5% (DOE, 2011). No Brasil, segundo dados do Balanço Nacional de Energia, as edificações representam aproximadamente metade do consumo de energia nacional, das quais o setor residencial conta com mais de um quarto (EPE, 2019). Estima-se que o crescimento da construção civil apresente aumento de até 28% do consumo em 2030, fato que tem levado entidades governamentais e também do setor privado a investirem em desenvolvimento de pesquisas visando a redução do consumo e racionalização do uso de recursos ambientais.

Paralelamente a esse cenário, o grande déficit habitacional no Brasil motivou o desenvolvimento de políticas públicas, como o Programa Minha Casa Minha Vida e a Lei Federal 11.888/2008 (Lei de Assistência Técnica Pública e Gratuita). Tais programas estimulam a reforma, ampliação e criação de novas edificações, voltadas especialmente para as habitações de interesse social. As habitações de interesse social (considerando famílias com renda até três salários mínimos) representam uma área de interesse especial, visto que englobam cerca de 80% do déficit habitacional brasileiro, estimado em 5,8 milhões de unidades (IPEA, 2013). Com a redução do déficit habitacional, é esperado também aumento no consumo de energia, proporcionado pela construção e ocupação das novas unidades. Além disso, o baixo investimento financeiro nesse tipo de edificação tem por resultado a criação de edificações que não consideram importantes condicionantes ambientais, acabando por adotar materiais que resultam em baixo desempenho e replicação de projetos sem considerar variáveis bioclimáticas no seu desenvolvimento e implantação. Esse contexto aponta a importância desse setor como área de pesquisa sobre eficiência energética e desempenho térmico de edificações.

Diversos estudos são realizados para compreender os fenômenos que envolvem o desempenho térmico das edificações (BODACH; HAMHABER, 2010; BROWN et al., 2014, CICELSKY; MEIR, 2014; YOUG et al., 2017). O foco desses estudos volta-se para a identificação das variáveis que ocasionam maior impacto no desempenho das edificações e

determinação de quais são as estratégias que garantem mais conforto e eficiência (ALAIDROOS; KRARTI, 2015, EL-DARWISH; GOMAA, 2017; LOUKAIDOU, et al., 2017, CHARISI, 2017).

Por se tratar de um processo tão complexo, muitos pesquisadores adotam simulação computacional para auxiliar na realização dos estudos. Com essa ferramenta, é possível obter vários indicadores de eficiência e medidas de desempenho a partir da inserção de dados como as características da edificação, localização geográfica e clima, composição da envoltória, condicionamento do ar, ocupação e muitos outros. Entretanto, quando se deseja propor medidas governamentais e em larga escala, é necessário acessar o desempenho e as estratégias para todo um estoque de edificações, o que seria inviável a partir de simulações computacionais de cada edificação existente. Encontrar a solução para essa problemática representa, portanto, um importante passo para o desenvolvimento de políticas públicas com foco no desenvolvimento de edificações com melhor desempenho.

O uso de modelos de referência é apontado como solução para esse problema. Modelos de referência, como apontam Corgnati et al. (2013), são ferramentas que representam de forma aproximada as edificações do mesmo tipo, sob mesmas condições de uso e região climática. Esse conceito também pode ser estendido à criação de padrões, por exemplo, como quanto ao comportamento do usuário (YU et al., 2011). Esses modelos são desenvolvidos a partir da coleta e análise de dados de uma amostra de estoque edificado, permitindo que diferentes cenários sejam analisados com uma quantidade muito menor de simulações.

Atualmente, não há um método normatizado para a criação desses modelos. Entretanto, como apontado por Schaefer e Ghisi (2016), a maioria dos estudos segue os mesmos passos. Inicialmente define-se qual é o objeto de estudo, seu uso, região climática, abrangência (regional ou nacional, por exemplo), etc., obtendo-se então dados a respeito dessas edificações através de levantamentos em campo ou em bases pré-existentes. As variáveis coletadas também variam de estudo para estudo. Enquanto alguns coletam dados como área de piso e consumo de energia, outros focam o interesse no ano de construção, tipo de instalações existentes, etc. Essas variáveis são tratadas de forma a se determinar as características mais representativas daquela amostra, podendo ser adotado um tratamento baseado em métodos estatísticos ou empíricos, uni ou multivariados, ou até mesmo apenas uma classificação do que foi levantado conforme um padrão existente. Por fim, o modelo é criado, baseando-se nas características mais representativas encontradas na amostra.

Organizações governamentais têm realizado esforços para obter modelos de referência que possam ser utilizados em estudos de desempenho térmico e energético. O Departamento de Energia dos EUA possui vários modelos de referência de edifícios em sua base de dados, sendo dois para representar edificações residenciais (um para habitações unifamiliares e outro para multifamiliares). Os modelos são como um ponto de partida para medir o progresso dos alvos de eficiência energética em edificações, a partir da aplicação de novas normas de conservação de energia (DOE, 2019). Na Europa, também foram feitos esforços para determinar modelos de referência do estoque de edifícios de cada nação. Em particular, o projeto TABULA (LOGA et al., 2008, LOGA et al., 2016) é citado como referência para a determinação de modelos de construção residencial e tem a colaboração de vinte países. No Brasil, estudos que buscam obter uma tipologia representativa também têm sido desenvolvidos por alguns pesquisadores, como Brandão (2003), Teixeira et al. (2015), Triana et al. (2015), Schaefer e Ghisi (2016) e Giacomin et al. (2019).

Não há um procedimento padronizado para a obtenção desses modelos. Uma forma de obtê-los é por meio da análise de agrupamentos. A análise de agrupamentos é uma análise multivariada exploratória e não inferencial que tem por objetivo identificar dentro de uma amostra de dados grupos com características semelhantes (BUSSAB et al., 1990). Essa divisão é feita de modo a formar agrupamentos com alta homogeneidade interna e alta heterogeneidade externa, ou seja, objetos semelhantes são mantidos em um mesmo agrupamento, enquanto objetos distintos são mantidos em agrupamentos distintos (HAIR et al., 2009). Dessa forma, é possível selecionar um elemento de cada agrupamento como o modelo de referência. Esse modelo seria utilizado em estudos de desempenho térmico, por exemplo, e os resultados obtidos, estendidos a todo o grupo. Esse método foi adotado por alguns pesquisadores, como Yu et al. (2011), Giglio et al. (2014), Schaefer e Ghisi (2016) e Li et al. (2018). Na área de eficiência energética em edificações, a análise de agrupamentos tem sido utilizada em diferentes aplicações, como a investigação do potencial de economia de energia (YU et al., 2011) e a classificação de edificações quanto a sua eficiência energética (GAITANI et al., 2010).

A análise de agrupamento é um procedimento complexo, que envolve várias etapas e processos subsequentes de tomada de decisão. De forma geral, inicialmente organiza-se um banco de dados com as variáveis relevantes para o estudo e aplica-se algum tratamento aos dados, como para a identificação de elementos atípicos ou padronização dos dados. Na sequência, uma medida de similaridade (medida matemática que determina a semelhança entre os objetos da amostra) é aplicada, transformando a matriz de dados em uma matriz de

similaridade. Por fim, são aplicados algoritmos de partição. Esses algoritmos especificam as regras que determinarão a designação dos objetos a um agrupamento ou outro. Eles podem ser determinados a partir de técnicas hierárquicas ou não-hierárquicas. Alguns exemplos de medidas de similaridade e algoritmos de partição podem ser encontrados em Bussab et al. (1990), Kaufman e Rousseeuw (2005) e Mingoti (2007).

É grande a quantidade de medidas de similaridade e algoritmos de partição existentes, e sabe-se que as diferentes combinações entre eles podem levar a diferentes resultados, pois cada método consegue interpretar melhor a estrutura de cada conjunto de dados. Entretanto, a maioria dos estudos não parece adotar algum critério para a seleção de um ou outro, passando a adotá-los de forma indiscriminada, a partir daquele que foi mais recorrente em outros estudos realizados. Esse quadro evidencia a necessidade de estudar a influência das diferentes combinações entre os procedimentos da análise de agrupamentos e seus efeitos sobre o resultado final.

Outro ponto não identificado nos estudos que utilizam a análise de agrupamentos na área de edificações é a aplicação de um método de validação dos resultados. A análise de agrupamentos é uma técnica exploratória, ou seja, não há um p-valor que possa qualificá-la como adequada. Nesse sentido, adicionar uma etapa de validação capaz de avaliar a qualidade das partições é imprescindível para assegurar bons resultados.

Como visto, as edificações contribuem de forma expressiva para o consumo de energia e impactos ambientais, motivo pelo qual diversos estudos têm se voltado a oferecer soluções para reduzir o consumo. Para resolver esse problema, é necessário adotar medidas que envolvam todo o estoque de edificações. Devido à inviabilidade de se acessar o desempenho de cada uma das edificações de todo um estoque, o uso de modelos de referência tem crescido e se apresentado como estratégia eficiente e acessível, contribuindo para a obtenção de resultados de forma mais prática e objetiva. Uma das formas de obtenção desses modelos é através da aplicação da análise de agrupamentos. Entretanto, essa análise parte da combinação de vários procedimentos envolvendo diferentes tratamentos de dados, medidas de similaridade e algoritmos de partição, sendo necessário conhecer os efeitos das diferentes combinações desses procedimentos sobre o resultado final. Tais evidências não puderam ser encontradas em estudos na literatura referente ao desempenho térmico em edificações, indicando a necessidade de desenvolvimento de estudos que preencham essa lacuna.

1.1. Objetivos

1.1.1. Objetivo geral

O objetivo geral deste trabalho foi desenvolver um método de aplicação da análise de agrupamento a partir do qual seja possível obter modelos de referência de edificações para uso em estudos de desempenho térmico, a partir de diferentes configurações de clusterização.

1.1.2. Objetivos específicos

Alguns objetivos específicos foram traçados:

a) Caracterizar um estoque inicial de habitações de interesse social como estudo de caso para o desenvolvimento do método de aplicação da análise de agrupamento;
b) Identificar subgrupos na amostra de habitações com características similares, a partir da combinação de diferentes tratamentos de dados, medidas de similaridade e algoritmos de partição, usando a análise de agrupamentos;
c) Determinar modelos de referência para cada subgrupo, encontrados a partir de diferentes métodos;
d) Encontrar o método que teve como resultado a melhor formação dos agrupamentos e modelos, a partir da aplicação de medidas de validação internas e relativas;
e) Comparar os modelos e agrupamentos selecionados quanto às suas características geométricas e desempenho térmico;
f) Avaliar a qualidade da combinação dos diferentes parâmetros de clusterização na formação dos agrupamentos.

1.2. Estrutura da tese

Este trabalho foi dividido em cinco capítulos. Inicialmente, são apresentados no primeiro capítulo os conceitos que norteiam essa proposta de estudo. Faz-se uma contextualização geral da forma como o estoque de edificações tem sido estudado a partir de modelos de referência para simplificar os estudos de desempenho de edificações. Verifica-se a inexistência de um método padronizado para obtenção desses modelos, ponto identificado

como principal problema de pesquisa e que direcionou os objetivos desse estudo. Também são definidos nesse capítulo os objetivos gerais e específicos, e apresentada a estrutura geral desse trabalho.

No segundo capítulo, apresenta-se uma revisão de literatura dividida em dois tópicos: o uso de modelos de referência em estudos de desempenho térmico de edificações e a análise de agrupamentos. O primeiro tópico aborda a problemática de estudar o estoque edificado e apresenta-se o uso de modelos de referência como solução. São apresentados os conceitos que norteiam essa ferramenta, além de diferentes estudos que a utilizaram. No segundo tópico, a análise de agrupamentos é apresentada, por tratar-se de uma etapa fundamental desse trabalho. São apresentados seus conceitos, aplicações e procedimentos para a obtenção dos agrupamentos a partir da adoção de diferentes medidas e algoritmos.

Na sequência, no terceiro capítulo, é apresentado o método que será aplicado nesse estudo. Esse método foi dividido em seis etapas sucessivas. Inicia-se com a criação de matrizes de dados, criadas a partir de diferentes tratamentos estatísticos. Na sequência, aplica-se a análise de agrupamentos a partir da combinação de cinco medidas de similaridade e cinco algoritmos de partição e determinam-se os modelos de referência. Medidas de validação interna e relativa são aplicadas a todos os métodos provindos das diferentes combinações e seleciona-se aquele que gerou a melhor formação. Por fim, os grupos formados e seus modelos são descritos a partir da geometria e do desempenho térmico dos objetos que os compõem.

No capítulo quatro são apresentados os resultados obtidos ao longo do estudo, como resposta ao objetivo geral e a cada objetivo específico.

No quinto e último capítulo, apresentam-se as principais conclusões obtidas com esse estudo e recomendações para aplicação da análise de agrupamentos, as limitações desse trabalho e sugestões para trabalhos futuros.

2. Revisão de literatura

2.1 Uso de modelos de referência em estudos de desempenho térmico em edificações

2.1.1 Conceitos gerais

Para que estratégias de redução do consumo de energia e melhora do desempenho térmico tenham sucesso, é importante que sejam pensadas a nível global, ou seja, possam ser aplicadas a todo um estoque de edificações, e que cada estratégia seja elaborada para edificações com características semelhantes (DASCALAKI et al., 2010).

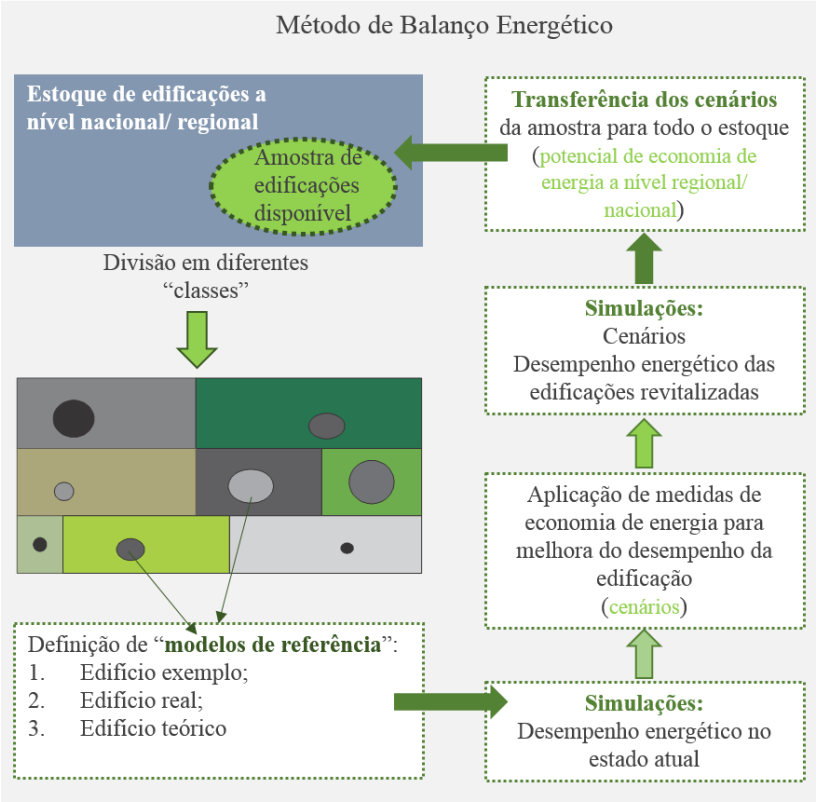
A simulação computacional é uma ferramenta interessante para obtenção de indicadores de eficiência, auxiliando na proposição de medidas sustentáveis que favoreçam o bom desempenho de edificações (VERSAGE, 2015). Entretanto, como fazer para simular todo um estoque edificado? Essa tarefa parece inviável, dada a grande quantidade de edificações existentes. Como solução plausível a este problema, diversos autores defendem o uso de modelos de referência (McKENNA et al., 2013, É et al., 2014, SOKOL et al., 2017). Os modelos de referência são estratégias capazes de representar todo um estoque de edificações em estudos de diversos âmbitos, incluindo, por exemplo, estudos de desempenho térmico de edificações.

Os modelos de referência têm se mostrado fundamentais para estudos na área de pesquisa de desempenho de edificações, o que pode ser comprovado pelo aumento de publicações envolvendo o tema nos últimos anos.

Embora seja um tema atual e sua aplicação cada vez mais frequente, ainda não há uma definição padronizada do que seria um modelo de referência. De forma geral, atribuem-se ao termo modelos que seriam capazes de representar um grupo de objetos (que, neste estudo, seriam as edificações e seu uso) cujas características e/ou comportamento se assemelham. Segundo Sanches e David (2007), trata-se de um modelo que representa a realidade do que será analisado. Para Dascalaki et al. (2010) o termo se aplica a uma classificação de edifícios a partir de algumas características que se relacionam ao desempenho da edificação. Na Diretiva de Desempenho Energético de Edificações (*Energy Performance Building Directive* - EPBD *recast* (UE, 2010), os modelos de referência são definidos como edifícios característicos e representativos da funcionalidade, localização geográfica e condições ambientais internas e externas do estoque de edificações (CORGNATI et al., 2013).

São várias as aplicações dos modelos de referência. Eles têm sido aplicados, por exemplo, em pesquisas de custos e aplicação de novas tecnologias, desenvolvimento de normas e diretrizes construtivas, para acessar os efeitos de medidas de conservação de energia e fazer projeções futuras para diferentes situações. Uma das principais aplicações dos modelos de referência, como apontam Dascalaki et al. (2011), é que podem ser utilizados para acessar o balanço térmico de edificações, com o uso de simulações computacionais. Assim, é possível inferir sobre o impacto causado no desempenho quando medidas de economia de energia são aplicadas. A Figura 1 mostra um esquema de um método aplicado pelo governo italiano. Primeiramente, dados sobre as edificações são obtidos em uma amostra do estoque edificado. As edificações levantadas são divididas em subgrupos conforme suas características. Para cada subgrupo, é definido um modelo de referência. Esse modelo é então submetido a simulações térmicas conforme suas características no estado atual e também a partir de novos cenários, quando medidas de economia de energia são adotadas. A partir dos resultados, é possível avaliar quais medidas apresentam resultados eficientes, sendo então transferidas para o estoque edificado (LOGA et al., 2008).

Figura 1 - Esquema de um método adotado para realização do balanço térmico de edificações a partir do uso de modelos de referência.



Fonte: Adaptado de Loga et al. (2008)

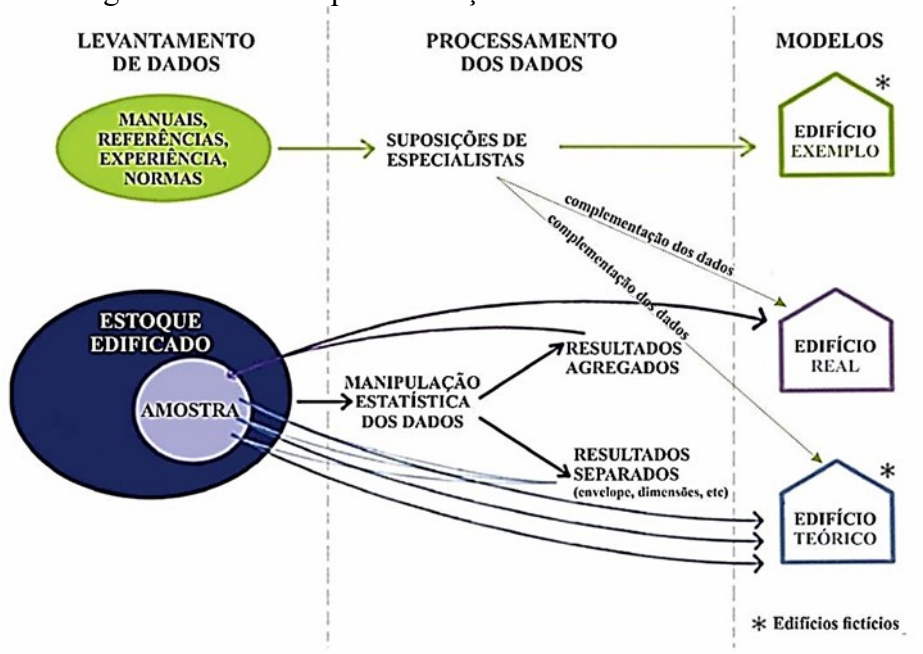
Por causa da eficiência obtida com sua aplicação, diversos governos estão adotando políticas para a criação de modelos de referência. O DOE, Departamento de Energia dos Estados Unidos, disponibiliza em seu site diversos modelos de edificações para aplicação em estudos de desempenho, sendo dois deles para uso residencial. Esses estudos avaliam a qualidade de sistemas de iluminação, ventilação, qualidade do ar interno ou desempenho térmico e energético. Também dão suporte ao desenvolvimento de novas versões da ASHRAE Standard 90.1 (TORCELLINI et al., 2008). Na Europa, também estão sendo desenvolvidos projetos cuja determinação de modelos e sua aplicação em estudos de desempenho têm se tornado primordial. Dentre eles, pode-se citar o projeto TABULA, voltado ao setor residencial. Neste projeto, vários países colaboram para a construção e mapeamento de um banco de dados sobre o estoque edificado de edifícios residenciais tanto unifamiliares quanto multifamiliares, classificados pelo seu ano de construção, tamanho e consumo energético. O banco de dados está disponibilizado no site www.building-typology.eu (LOGA et al., 2008, LOGA et al., 2016).

Os métodos para obtenção dos modelos diferem de estudo para estudo. Alguns partem de uma simples classificação do estoque quanto à sua funcionalidade, tipologia arquitetônica e ano de construção (KOHLER; HASSLER, 2002, SERGHIDES et al., 2016; BHATNAGAR et al., 2019). Outros são obtidos a partir da aplicação de métodos estatísticos, sendo uni ou multivariados (THEODORIDOU et al., 2011b, DASKALAKI et al., 2011, SANDBERG et al., 2016, SOKOL et al., 2017, GEYER et al., 2017). Recentemente, redes neurais artificiais também têm sido aplicadas à obtenção de modelos de referência (SANGIREDDY et al., 2019). No entanto, de forma geral, a sua obtenção parte da conclusão de quatro passos. Primeiramente, define-se um objeto de estudo, conforme funcionalidade, região climática, etc. Obtém-se as variáveis com base nos objetivos do estudo, coletando-as em campo ou em base de dados existente. Essas variáveis devem ser tratadas, organizando os dados para achar as características mais representativas. Por fim, define-se o modelo a partir dessas características (CORGNATI et al., 2013, SCHAEFER; GHISI, 2016).

Na EPBD *recast* (UE, 2010), três formas de definir o modelo são propostas: através do edifício exemplo, do edifício real e do edifício teórico. O Edifício Exemplo (*Example Reference Building*) é obtido quando não há dados de levantamento e estes são encontrados, portanto, na literatura, através de normas, manuais ou até de conhecimento de *experts* na área. É um edifício fictício, construído a partir do que se supõe ser mais característico na amostra. O Edifício Real (*Real Reference Building*), como o próprio nome diz, trata de uma edificação real, não fictícia. É definido a partir do tratamento estatístico de dados coletados em campo, adotando como

modelo a tipologia mais frequente no estoque. Por fim, o Edifício Teórico (*Theoretical Reference Building*) é também uma edificação fictícia, mas baseada em dados reais. A partir de dados coletados em campo, obtém-se as características mais representativas com tratamento estatístico dos dados. A partir dessas características, é criado o modelo. O modelo final não precisa, necessariamente, ser definido exclusivamente a partir de apenas um dos métodos. Ele pode ser obtido a partir da combinação de todos eles, para cada informação. Por exemplo, um modelo pode ter sua forma determinada a partir da literatura, enquanto a operação do edifício pode ser definida a partir da manipulação estatística de dados coletados em campo (CORGINATI et al., 2013, SCHAEFER; GHISI, 2016). A Figura 2, desenvolvida por Corgnati et al. (2013), apresenta um esquema dos métodos mencionados

Figura 2 - Métodos para obtenção de modelos de referência.



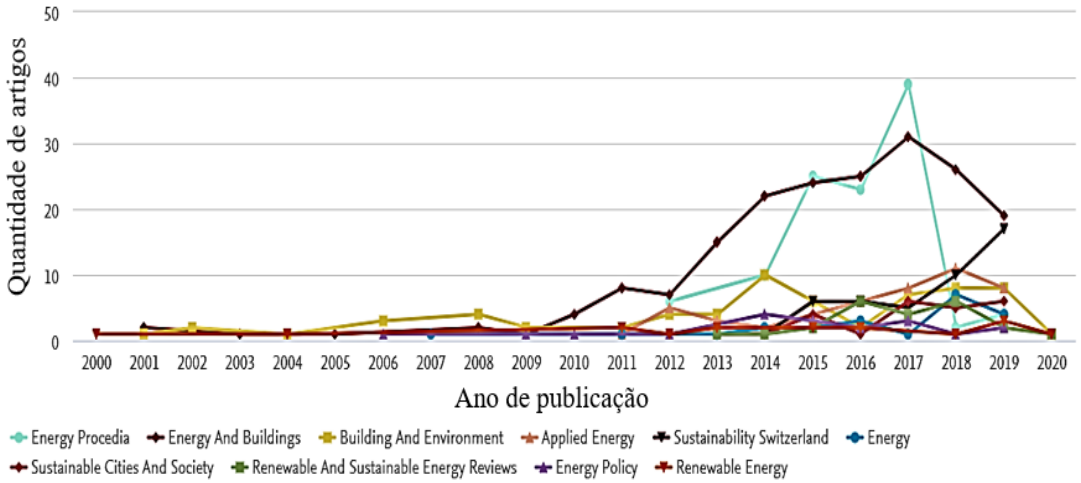
Fonte: Adaptado de Corgnati et al. (2013)

2.1.2. Estudos de desempenho térmico de edificações a partir de modelos de referência

O estoque edificado tem sido objeto de estudo em pesquisas sobre desempenho térmico e energético há algum tempo, tendo crescido sobremaneira ao longo dos últimos quinze anos. A Figura 3 mostra a quantidade de artigos publicados por ano e por fonte acerca do tema, ao se fazer uma a busca rápida em bases como a SCOPUS (total de 5.923 artigos, que possuem os termos "*reference building*" ou "*reference model*" ou "*typology*" ou "*archetype*" e "*building*"

no título ou resumo). Esse cenário mostra a necessidade de ampliar o conhecimento sobre o estoque de edificações existentes, o que pode ser feito através do uso de modelos de referência.

Figura 3 - Quantidade de publicações por ano e fonte com o termo *reference building*.



Fonte: SCOPUS (2019)

Os estudos apresentados a seguir apontam a importância do tema e aplicabilidade dos modelos, forma de obtenção e resultados obtidos a partir dos mesmos.

Em 2002, Kohler e Hassler já apresentam a necessidade de estudar o estoque edificado. Segundo apontam os autores, o foco de pesquisa apenas em novas construções, como usual, não reflete adequadamente as condições existentes. É preciso conhecer as características do estoque para então analisar as necessidades de melhorias no setor. Além disso, projetando um cenário futuro, a necessidade de *retrofit* para adequação de edificações antigas está cada vez mais em evidência, contrapondo o interesse de focar estudos apenas no design de novas edificações. Considerando esse contexto, os autores apresentam uma revisão sobre as diferentes abordagens existentes até então de estudos envolvendo o estoque edificado. Cerca de dez temas foram criados pelos autores para classificar as diferentes abordagens encontradas, dentre elas a energia, predição da demanda de edificações, materiais e seu impacto ambiental e infraestrutura. Estudos que envolviam o tema energia, por exemplo, buscavam acessar o consumo energético geral dos edifícios, determinar como os padrões de consumo evoluíam e determinar de que forma as estratégias para redução do consumo energético poderiam ser implementadas. De forma geral, as edificações eram divididas conforme seu uso e época de construção e caracterizadas pelo consumo anual de energia por metro quadrado de uma

edificação típica do seu grupo. Esse valor era multiplicado pelo somatório das superfícies de todas as edificações do seu grupo, obtendo-se então o consumo predito de todo o estoque.

A partir da revisão, os autores apresentaram uma proposta de modelar de forma integrada o estoque de edificações da Alemanha, cujos objetivos envolviam a criação de dados que pudessem ser utilizados em diferentes aplicações, identificação de correlações entre os modelos e validação dos mesmos. O estudo envolveu tanto edificações residenciais como não residenciais. Por fim, foi proposta uma metodologia integrada para seis temas de pesquisa: ciclo de vida, edificações históricas, gerenciamento de estratégias no estado atual, modelamento do produto, tempo, infraestrutura e uso do solo.

Em vista de uma nova regulamentação para o desempenho energético de edificações na Grécia, Theodoridou et al. (2011a) desenvolveram um método para investigar a eficiência de medidas de conservação de energia no estoque edificado, a partir da proposição de novos cenários. O método desenvolvido baseou-se na análise estatística dos dados que descrevem as edificações. Inicialmente, as edificações do estoque foram classificadas conforme seu período de construção. Segundo os autores, essa classificação permite agrupar edificações com uso e sistemas construtivos similares, além da tipologia construtiva e uso de equipamentos. Para cada categoria, obteve-se informações sobre as edificações em bases de dados existente, tais como: o ano de construção, funcionalidade, área construída, sistemas construtivos e materiais, tipo de cobertura, quantidade de pavimentos e existência de pilotis. Também foram levantados, em base de dados governamentais, informações a respeito do comportamento do usuário, como as rotinas de operação e consumo de água.

Uma amostra de edificações de cada categoria também foi selecionada para realização de auditorias e monitoramento, para posterior validação dos modelos. Foram obtidos com esse levantamento indicadores como qualidade do ar e conforto térmico dos usuários.

Essas edificações foram submetidas à análise térmica com simulação computacional, obtendo-se o consumo médio anual de energia para o condicionamento artificial para aquecimento. O programa escolhido foi o *EnergyPlus*. Os resultados mostraram boa correlação com os dados reais, obtidos a partir do monitoramento. Os pesquisadores concluíram que a classificação das edificações em grupos similares pode auxiliar no estudo de estratégias para melhoria do desempenho a partir cenários de medidas de conservação de energia em estudos futuros, identificando, assim, as propostas mais efetivas. Dessa forma, é possível prever um cenário para todo o estoque edificado, com poucas simulações.

No Egito, modelos de referência também foram foco de estudos de desempenho energético de edificações, como o desenvolvido por Attia et al. (2012), cujo foco foi o setor residencial. Conforme apontam os autores, a busca crescente por conforto térmico no interior das edificações tem levado a um aumento do consumo de energia pelo uso de condicionadores de ar. Este fato evidencia a necessidade de desenvolver modelos capazes de predizer o consumo energético de um estoque edificado, a partir da criação de cenários. Dessa forma, seria possível determinar estratégias de economia de energia mais eficientes aplicadas àquele estoque.

Para desenvolver seu estudo, os autores levantaram dados em banco de dados existentes e também em campo, a fim de determinar um perfil de uso de energia pelos moradores. Como o foco do estudo era o consumo de energia para condicionamento artificial, determinou-se uma amostra de apartamentos em que houvesse tal sistema instalado e em uso. A coleta de dados se deu nas três regiões metropolitanas do país. Nos levantamentos de campo, foram obtidos dados sobre a geometria da edificação, suas dimensões e ano de construção, além de uma descrição dos equipamentos condicionadores de ar instalados. Dados das faturas de energia também foram coletados. Em um levantamento com uma amostra mais restrita, foram ainda levantados os demais equipamentos existentes na residência, suas características, padrão de uso e pico de demanda de energia. Duas tipologias de apartamento foram adotadas, baseando-se na predominância desses modelos na amostra. O comportamento do usuário foi definido por meio da densidade de ocupação média e rotinas de operação. Quanto ao consumo de energia, os dados foram categorizados em demanda para iluminação e eletrodomésticos, demanda para cocção e demanda para aquecimento da água. Os equipamentos para condicionamento artificial (ventiladores e condicionadores de ar) foram caracterizados em um item a parte, através da intensidade de carga para resfriamento.

Em posse dos dados, os modelos foram submetidos a simulações, obtendo-se a partir delas o consumo de energia. Ao comparar os dados obtidos com as simulações com as médias mensais estimadas, os autores encontraram boa concordância, apresentando uma diferença de apenas 2% acima para o consumo obtido a partir das simulações.

O estudo realizado por Ballarini et al. (2014) busca analisar o potencial de economia de energia e redução das emissões de CO_2 do estoque edificado, através do uso de edificações de referência. Com esse intuito, as tipologias desenvolvidas para o projeto TABULA foram utilizadas pelo governo italiano para verificar as medidas com o melhor custo benefício quanto ao desempenho energético, conforme regulamentado nas metas da EPBD *recast* (UE, 2010).

Como apontam os autores, o desenvolvimento de edificações de referência para representar o estoque edificado é um requisito que todos os países membros do projeto devem cumprir, como etapa preliminar do desenvolvimento de estudos de análise de custo-benefício. De acordo com as diretrizes da Comissão Europeia, a edificação de referência deve ser classificada de acordo com três parâmetros: (1) localização, conforme sua região climática, (2) período construtivo, que está relacionado com as técnicas e materiais construtivos, e (3) forma e tamanho da edificação, que no projeto são classificadas em quatro categorias (casas unifamiliares, geminadas, edificações multifamiliares e blocos de apartamento). Esses três parâmetros devem formar a "Matriz de Tipologias Edilícias". Para cada uma das tipologias, é necessário desenvolver uma edificação de referência, que represente de forma geral as edificações daquela tipologia. A definição desse modelo pode se dar a partir de modelos existentes na literatura, pela seleção de uma edificação real daquele grupo ou pela construção de um modelo virtual, baseado nas características mais frequentes daquele grupo. Na Itália, foram identificadas três classes de regiões climáticas, oito classes de períodos construtivos e as quatro classes previstas no Projeto Tabula quanto ao tamanho e forma das edificações. O estudo de Ballarini et al. (2014) foi desenvolvido especificamente para a região de Piemont, cuja Matriz de Tipologias Edilícias era composta por três classes de tamanho e forma (edificações unifamiliares, multifamiliares e blocos de apartamentos) e seis classes de períodos, resultando em um total de dezoito modelos de referência.

Como definido no Projeto TABULA, dois conjuntos de medidas de eficiência foram aplicadas a cada um dos modelos: medidas padrão (considerando a legislação vigente) e medidas avançadas (mais restritivas). Essas medidas incluíam maior isolamento térmico das paredes, coberturas e pisos, substituição das aberturas, substituição do sistema de aquecimento artificial e substituição do sistema de aquecimento de água. O desempenho das edificações foi analisado através de dois indicadores: consumo por área, que corresponde ao desempenho energético por metro quadrado aquecido (expresso em kWh/m^2) e a emissão de CO_2 por metro quadrado aquecido (expresso em kg/m^2). Como os modelos são considerados representativos do estoque edificado, os resultados obtidos com os modelos foram multiplicados pela área de piso aquecido estimada para todo o estoque.

Os resultados encontrados mostraram o grande potencial de redução do consumo de energia se as medidas propostas fossem adotadas. O pior desempenho foi encontrado para edificações anteriores a 1976, que englobam a maior parte dos edifícios de Piemont. Assim sendo, mesmo aplicando medidas básicas nessas edificações, obter-se-ia uma grande melhora

do desempenho energético e redução dos níveis de emissão de CO_2. As medidas padrão e avançadas apresentaram cerca de 77% e 85% de economia de energia, respectivamente. Como a diferença é pequena, os autores sugerem que uma análise de custo seja realizada para verificar qual das alternativas oferece o melhor custo-benefício. Os estudos mostraram que as edificações de referência podem ser instrumentos interessantes para acessar o consumo de energia de todo um estoque e para analisar o desempenho a partir da criação de cenários.

No estudo desenvolvido por Kragh e Wittchen (2014), dois métodos para obter modelos de referência do estoque edificado residencial da Dinamarca foram adotados: um a partir de um modelo real e outro a partir de um modelo teórico. O principal objetivo dos pesquisadores era desenvolver uma ferramenta capaz de analisar os efeitos de diferentes medidas de conservação de energia para todo o estoque residencial. O primeiro método, o modelo real, foi desenvolvido a partir da seleção de uma edificação real, considerada aquela cujas características são mais frequentes na amostra para o seu período construtivo, dentre elas: área de piso aquecida, número de pavimentos, tipo de isolamento térmico e características do morador. O segundo método, o modelo teórico, foi determinado a partir da média aritmética das características das edificações levantadas, tais como área de piso aquecido e número de pavimentos, enquanto a geometria do modelo foi adotada baseando-se em algumas suposições, como a forma, o pé direito, área de parede externa e área de janela. Nos dois casos, os dados foram obtidos a partir de dois bancos de dados existentes. Para cada método, as edificações do estoque foram classificadas em três tipologias residenciais (casa unifamiliar, geminada e bloco de apartamentos) e nove períodos construtivos.

Para estimar o consumo de energia dos modelos, uma ferramenta de cálculo do desempenho energético foi desenvolvida baseando-se na norma ISO 13790 (2008), a partir da qual obteve-se a demanda energética necessária para aquecimento dos modelos. Os pesquisadores observaram que, ao comparar os resultados obtidos com os cálculos com os dados estatísticos, obteve-se uma pequena diferença dos valores de aproximadamente 3% a mais para o método do modelo. Os resultados também mostraram que há grande potencial de economia de energia para as edificações classificadas nos períodos de 1851 a 1930 aproximadamente 11TWh/ano) e de 1961 a 1972 (aproximadamente 9TWh/ano).

Os autores concluíram que os modelos podem ser utilizados para a proposição de estratégias políticas de planejamento para melhora do desempenho térmico de todo o estoque edificado, e que a aplicação de modelos de referência para estudos de cenários futuros é adequada e eficiente.

Uma pesquisa realizada em Chipre (SERGHIDES et al., 2016) usou modelos de referência para verificar se as medidas de conservação de energia propostas pelas normas europeias para o estoque edificado existente e novas edificações seriam suficientes para alcançar as metas desejadas.

Cerca de 2480 edificações foram levantadas a fim de criar um perfil de consumo de energia. Foram obtidos dados sobre a sua geometria (área total e volume dos ambientes aquecidos), sistemas construtivos (capacidade e transmitância térmica das paredes, piso, cobertura e aberturas), sistemas de condicionamento artificial (para aquecimento e resfriamento) e consumo de energia para aquecimento e resfriamento, além da contribuição do uso de energias renováveis. As edificações levantadas foram separadas em três categorias tipológicas (multifamiliar, geminadas e unifamiliar) e em quatro períodos (antes de 1980, entre 1980 até 2007, de 2007 a 2013 e após 2014). Foram então determinados doze modelos de referência (um para cada tipologia e período), considerados representativos do estoque nacional.

Cada um dos modelos foi submetido a simulações de desempenho, com uso da ferramenta iSBEM-cy, de forma a obter-se três indicadores de desempenho: o consumo de energia para aquecimento e resfriamento, o nível de emissão de CO_2 e a contribuição de sistemas renováveis de energia. Esses indicadores foram obtidos para o cenário atual (referência 2005, ano dos levantamentos) e cenários futuros: 2020, 2030 e 2050. Para isso, o programa compara a energia global consumida pela edificação em questão (dada em kWh/m^2 anual) com o desempenho de uma edificação de referência padrão. Essa edificação é exposta a mesma condição climática da edificação em análise, com mesma orientação, geometria e uso, mas com tipo e área envidraçada, sistemas construtivos e condicionamento artificial padrão, assim como o padrão de operação. Os autores observaram discrepâncias entre os resultados monitorados e aqueles obtidos através de simulação. O principal motivo foi devido ao algoritmo de simulação que controla o aquecimento e resfriamento, quando ultrapassada a temperatura de set point, liga o condicionamento na casa inteira, enquanto na realidade esses sistemas são ligados apenas nos ambientes ocupados. Os autores sugerem que o programa seja atualizado, em especial com a criação de um padrão de ocupação mais real.

Os autores também verificaram que, com base nas tendências atuais, as metas de conservação de energia almejadas pela EPBD *recast* (2010) são inatingíveis. Um dos principais motivos seria a inadequação das medidas de reforma dos edifícios antigos, que acontecem em quantidade abaixo do desejado e cuja manutenção tem baixo impacto no desempenho do

estoque edificado nos anos futuros. Outro motivo estaria relacionado com a necessidade de resfriamento. Em Chipre, os esforços voltados a minimizar a produção de CO_2 deveriam focar em redução do consumo de energia tanto para aquecimento quanto resfriamento, pois o resfriamento representou mais de 60% da emissão de CO_2 em 2015.

2.2. Análise de agrupamentos

2.2.1. Conceito de análise de agrupamentos

Para entender o que é uma análise de agrupamentos ou análise de *cluster*, deve-se, anteriormente, entender o que é um *cluster*. *Cluster* é um conjunto de dados, comumente chamados de "objetos", cujas características se assemelham. Isto quer dizer que, embora sejam objetos independentes, que apresentam diferenças no somatório de suas características, estão ainda assim correlacionados de alguma forma, assumindo juntos uma mesma identidade. Assim, ao considerar, por exemplo, as cartas de um baralho, pode-se identificar ali um grupo de cartas de cor vermelha. Esse grupo é constituído por cartas com diferentes números e símbolos (naipes), mas que possuem em comum a característica de serem cartas vermelhas. A sua cor, portanto, é a característica que possuem em comum e as difere das demais cartas. É o que define a identidade daquele grupo.

A análise de *cluster* é, portanto, uma análise cuja finalidade centra-se na diferenciação ou delimitação de grupos de objetos dentro de uma amostra. Essa delimitação baseia-se na similaridade (ou dissimilaridade) entre os objetos, de forma a unir em um mesmo agrupamento aqueles que possuem características similares e a designar a agrupamentos diferentes os objetos que possuem características diferentes. Por isso, diz-se que um bom resultado da análise de *cluster* é uma divisão da amostra em grupos com alta homogeneidade interna e alta heterogeneidade entre os grupos (HAIR et al., 2009). Dessa forma, cada grupo pode ser descrito pelas características semelhantes entre seus objetos. É o que define sua identidade.

Há uma infinidade de aplicações para a análise de *cluster*. Tem sido usada, por exemplo, como pré-processamento ou etapa intermediária de estudos envolvendo mineração de dados. Pode-se aplicá-la para criar um índice para classificação (como, por exemplo, indicadores de eficiência), descobrir padrões, criar hipóteses, para detecção de objetos atípicos, redução e simplificação de dados, indicação de tendências, dentre muitas outras.

A análise de *cluster* é uma técnica exploratória, não teórica e não inferencial (BUSSAB et al., 1990). Isso quer dizer, em outras palavras, que não existe um valor de referência para classificar ou predizer o resultado de uma análise de *cluster* como adequado ou inadequado. O resultado depende da combinação entre diversas medidas e algoritmos, portanto adotando algumas combinações pode-se obter resultados bem diferentes de quando adotam-se outras. É também muito sensível às variáveis envolvidas na análise, por isso a seleção destas e o tratamento dos dados devem ser realizados com muito cuidado. O pesquisador certo da influência que cada variável exerce sobre o objeto de pesquisa. Por exemplo, em uma pesquisa na área de eficiência energética em edificações, as variáveis envolvidas na análise devem ter impacto sobre o desempenho energético das edificações. Se o objetivo da pesquisa for desempenho térmico ou lumínico, as variáveis selecionadas aqui devem corresponder aos indicadores de desempenho térmico e lumínico. E assim por diante. As variáveis devem ser capazes de descrever os fenômenos relacionados aos objetivos da pesquisa.

2.2.2 Procedimentos

A análise de *cluster* é complexa e envolve uma sequência de tomadas de decisão. A fim de auxiliar esse processo, Hair et al. (2009) elaboraram um diagrama de decisão, organizado em seis estágios. A Figura 4 apresenta os primeiros 3 estágios e a Figura 5, os estágios 4, 5 e 6.

No primeiro estágio, são definidos os objetivos da análise de agrupamento. O pesquisador deve ter em mente o que deseja obter com a aplicação da análise: é para classificação? Ou simplificação dos dados? O pesquisador deve também selecionar os objetos baseando-se nessa premissa e definir quais variáveis vão descrever os objetos da análise.

No segundo estágio, o pesquisador deve atentar para o tratamento estatístico das variáveis e também selecionar uma medida de similaridade baseada no tipo de variável envolvida na análise.

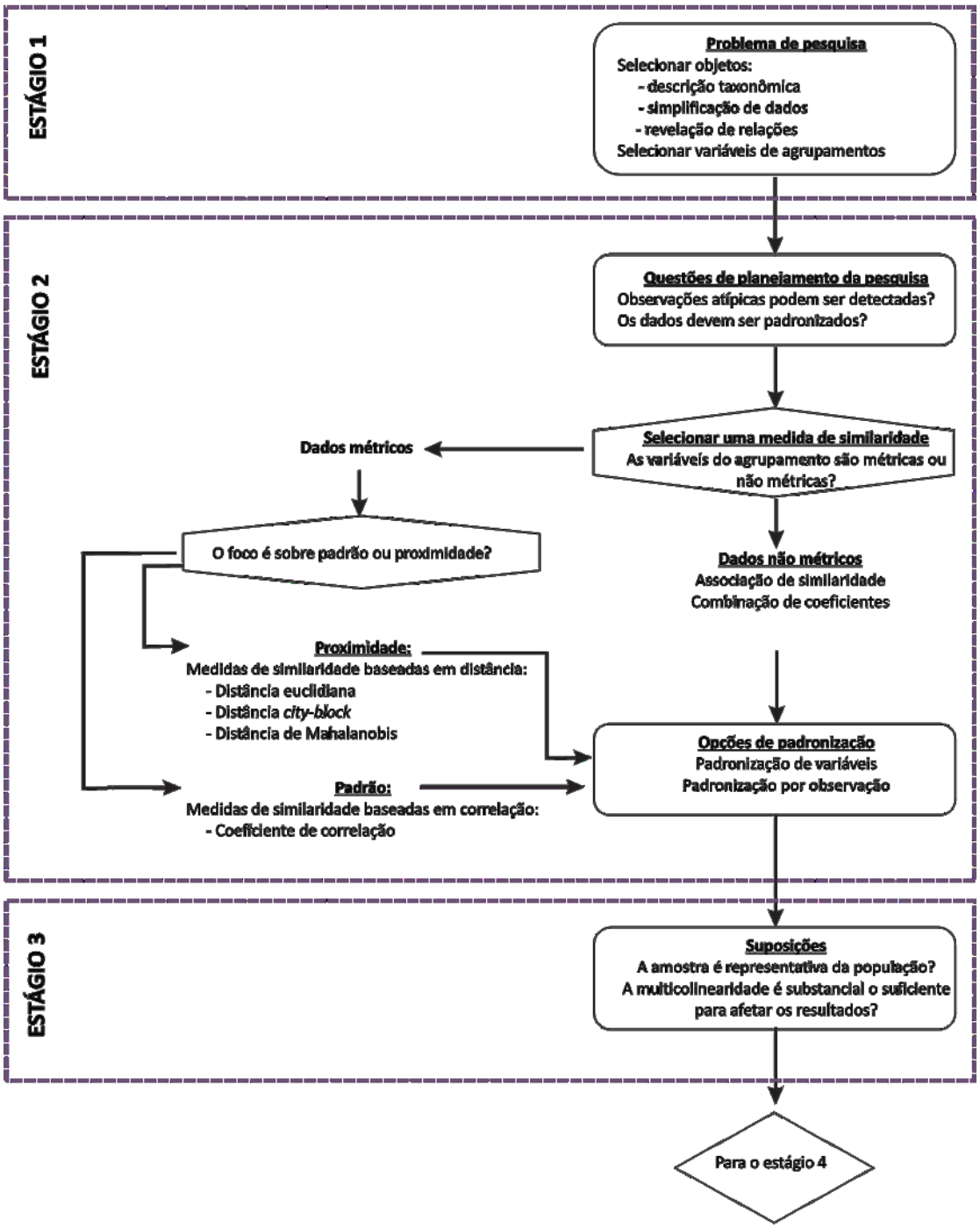
No terceiro estágio, suposições a respeito das características da amostra podem ser feitas. Uma breve análise dos dados pode indicar necessidade de tratamentos extras e expectativas que o pesquisador pode ter ao realizar a análise.

A análise de *cluster* é propriamente aplicada no quarto estágio, onde define-se as técnicas de partição como hierárquicas, não hierárquicas ou combinadas, e seleciona-se os algoritmos de partição. O produto desse estágio é a separação da amostra em grupos.

No quinto estágio, os agrupamentos formados devem ser interpretados, seja a partir do perfil do centroide ou da distribuição dos valores ao longo das variáveis envolvidas na análise.

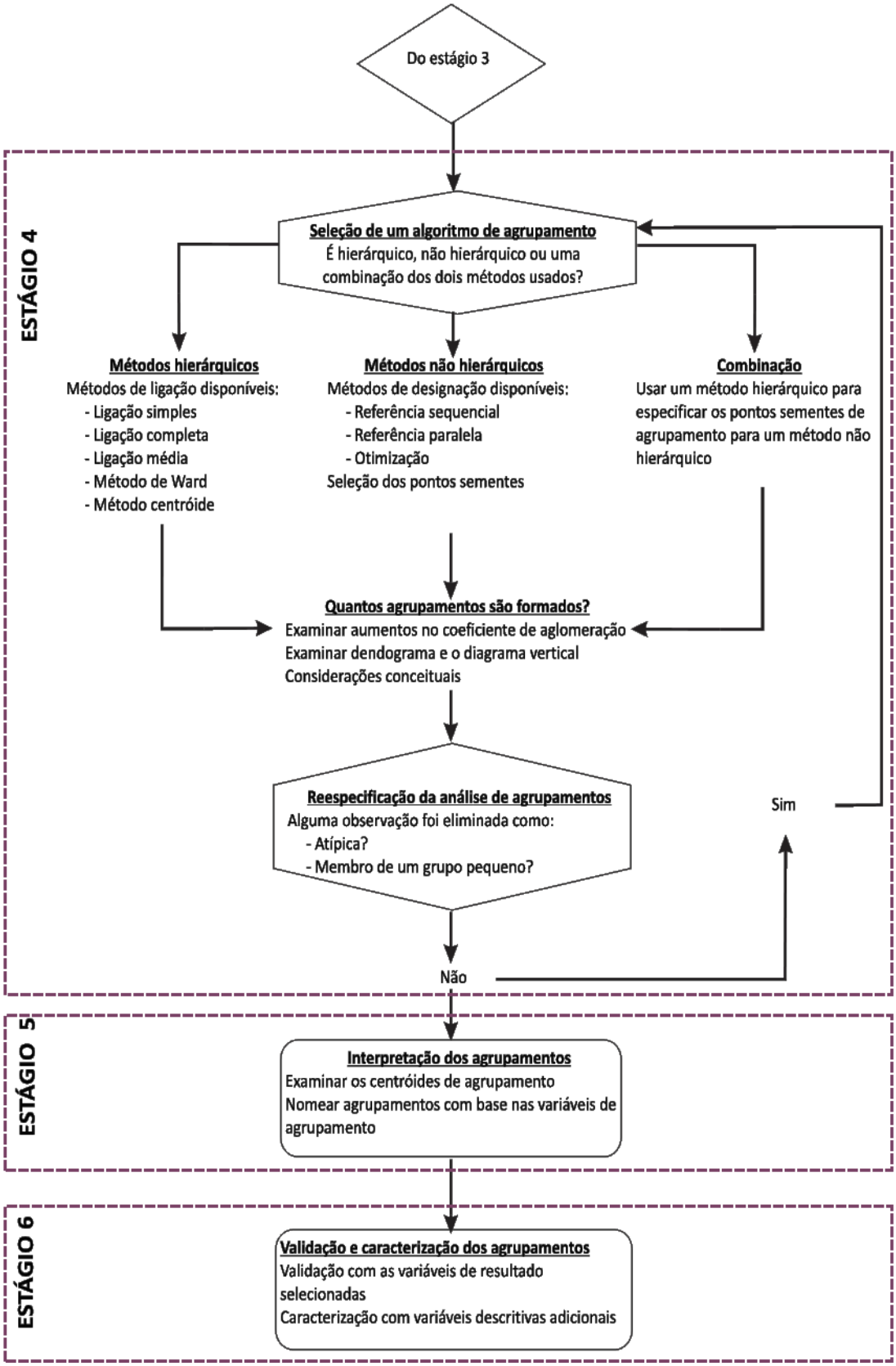
Por fim, no sexto e último estágio, medidas de validação devem ser aplicadas de forma a verificar a qualidade dos agrupamentos. O processo de validação pode ser realizado tanto a partir das variáveis utilizadas no processo de clusterização como a partir de variáveis adicionais.

Figura 4 - Estágios 1-3 do diagrama de decisão segundo Hair et al. (2009).



Fonte: Adaptado de Hair et al. (2009).

Figura 5 - Estágios 4-6 do diagrama de decisão segundo Hair et al. (2009).



Fonte: Adaptado de Hair et al. (2009).

Há três decisões importantes a serem tomadas ao se realizar uma análise de *cluster*: (1) a seleção e tratamento dos dados envolvidos na análise, (2) a escolha de uma medida de similaridade e (3) a determinação de um algoritmo de partição.

Quanto às variáveis, Bussab et al. (1990) afirmam que o produto final de uma boa análise de *cluster* é um conjunto de grupos que podem ser consistentemente descritos pelas características dos seus objetos, assumindo dessa forma uma identidade diferente para cada agrupamento. As variáveis que compõem o banco de dados da pesquisa são, portanto, um dos fatores de maior impacto sobre os resultados. As variáveis selecionadas devem: (1) caracterizar adequadamente os objetos da amostra e (2) estar relacionadas diretamente com os objetivos de pesquisa (HAIR et al., 2009).

As variáveis podem impactar os resultados pela diferença de magnitude entre elas (se uma análise envolver variáveis como área e quantidade de cômodos, é provável que a variável área tenha maior impacto sobre o resultado final). Por isso, é muito aconselhado padronizar os dados antes do início das análises. Isso significa dar um tratamento aos dados de forma a criar variâncias mais homogêneas na base de dados. Há várias formas de padronizar as variáveis. A mais comum é a padronização estatística (mais conhecida como *z-scores*), que redimensiona os valores de cada variável de forma a manter média zero e valores em termos de desvio padrão. Outra forma bem conhecida é a padronização min-max, que redimensiona os valores de forma a organizá-los em uma escala que vai de -1 a 1. Outras formas de padronização de dados podem ser encontradas em Bussab (1990) and Jain et al. (1999). Outra análise importante é a verificação quanto à presença de objetos atípicos (espúrios), que podem impactar negativamente os resultados. Esses objetos podem ser identificados na amostra através de gráficos de caixa, aplicando o teste *t de Student* ou através da medida de Mahalanobis, mais indicada a análises multivariadas.

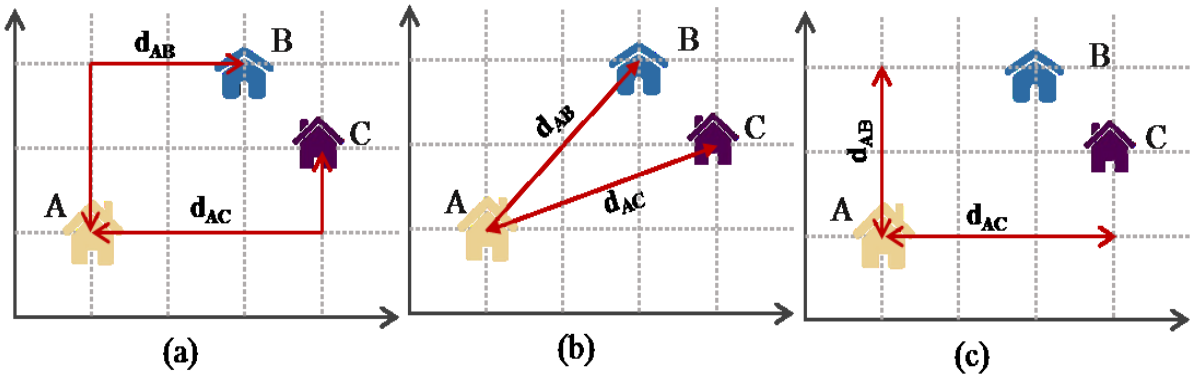
A medida de similaridade (ou dissimilaridade), por sua vez, representa as semelhanças (ou diferenças) entre os objetos da amostra através de uma medida matemática (MINGOTI, 2007). É uma consideração muito importante na análise de *cluster*, pois é a partir dessa medida que dois objetos serão quantificados como semelhantes ou diferentes e, assim, designados como pertencentes ao mesmo grupo ou a grupos distintos.

Há inúmeras medidas de similaridade. Cada uma delas representa uma forma diferente de quantificar a semelhança entre os objetos. Elas são medidas entre dois objetos da amostra ou entre um objeto e o centro do agrupamento, dependendo do tipo de medida de distância e do método de partição utilizados. Podem tratar-se de medidas de distância ou de correlação.

Medidas de distância são utilizadas quando o que importa é a proximidade entre os objetos. A correlação é preferida quando o que se deseja é muito mais verificar se há alguma associação entre os objetos do que proximidade, isto é, se existe algum padrão ao longo dos seus atributos, mais do que se há valores próximos. Diferentes medidas de similaridade também são aplicadas a diferentes tipos de variáveis: há medidas que usam apenas valores numéricos, outras apenas dados binários, e ainda medidas aplicáveis preferencialmente a dados qualitativos e ordinais (HAN et al., 2011).

A Figura 6 ilustra como a aplicação de diferentes medidas de similaridade a dois objetos resulta em diferentes valores de similaridade.

Figura 6 – Medidas de similaridade: (a) Distância City-block. (b) Distância Euclidiana. (c) Distância Chebyshev.



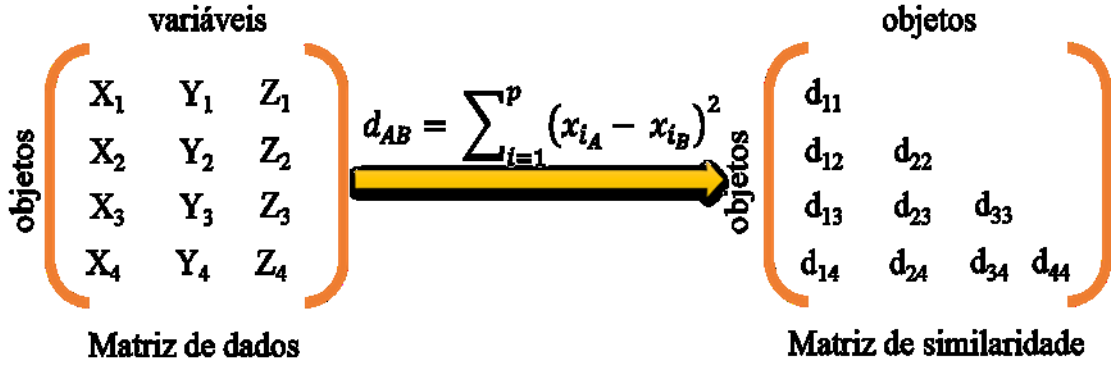
Outras medidas de similaridade são apresentadas em Bussab et al. (1990), Johnson e Wichern (1998) e Mingoti (2007).

O produto da aplicação da medida de similaridade é a transformação da matriz de dados em uma matriz de similaridade. A matriz de dados diz respeito à base de dados propriamente dita, onde são apresentados os *n* objetos nas linhas e seus *m* atributos nas colunas (ou vice-versa). Ao aplicar uma medida de distância a cada par de objetos (ou a um objeto e o centro do agrupamento), obtém-se a matriz de similaridade. A matriz de similaridade é uma matriz de referência triangular, pois registra a distância entre cada par de objetos (BUSSAB et al., 1990).

A Figura 7 ilustra a construção de uma matriz de similaridade a partir de uma matriz de dados, aplicando a distância euclidiana quadrada. À esquerda, é apresentada uma matriz de dados formada por quatro objetos descritos por três variáveis. Aplicando-se a distância

euclidiana quadrada a cada par de objetos, obtém-se a matriz de similaridade, apresentada na direita da Figura 7. Assim, a distância entre os objetos 1 e 2 é dada pela distância d_{12}.

Figura 7 - Construção da matriz de similaridade com a distância Euclidiana Quadrada.



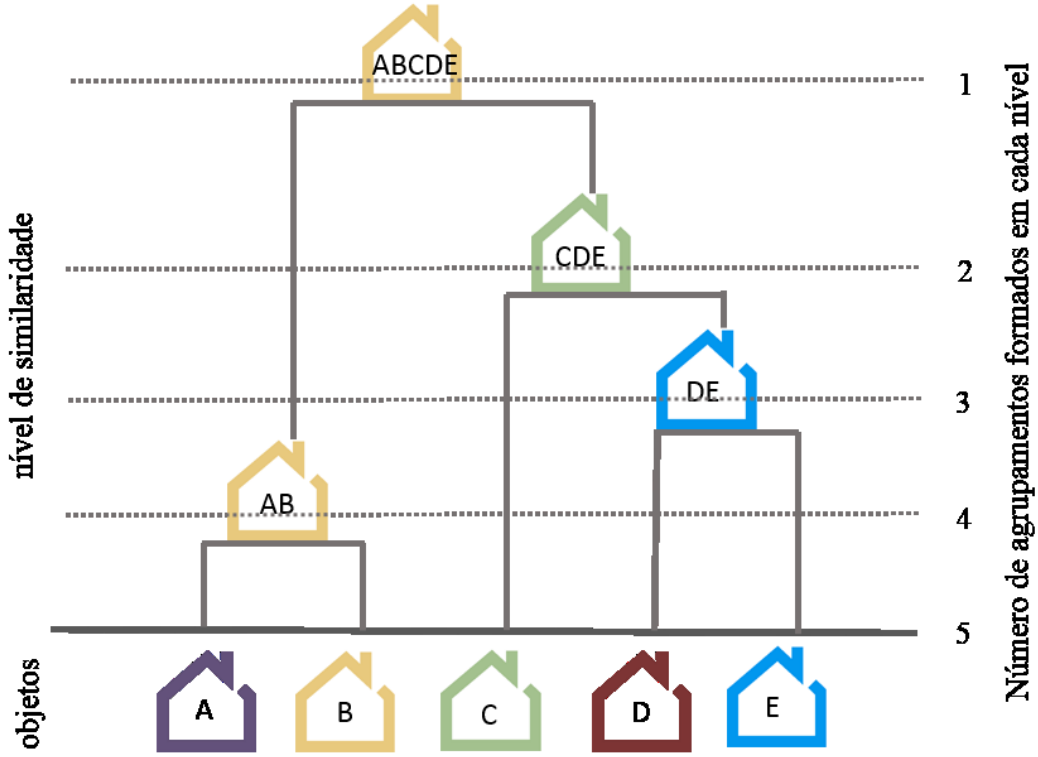
Fonte: Rosa (2014)

A matriz de similaridade apresenta o valor das distâncias entre os objetos. A partir dela, são aplicados os algoritmos de partição para a formação dos *clusters*. Os algoritmos de partição definem quais objetos, a partir da distância entre eles, serão unidos em um mesmo agrupamento. Diferentes algoritmos de partição possuem regras diferentes para designar os objetos aos agrupamentos. Há duas principais técnicas de partição: hierárquica e não hierárquica (BUSSAB et al., 1990; MINGOTI et al., 2007; HAIR et al., 2009).

A técnica hierárquica de partição caracteriza-se pela formação em árvore: o processo de agrupamento dá-se por etapas, e em cada nova etapa um par de objetos é unido conforme as regras do algoritmo de partição selecionado. Esse processo permite a construção de um gráfico chamado dendograma, que indica o nível de similaridade obtido com cada nova união de dois *clusters*. A Figura 8 representa o processo de agrupamento a partir do método hierárquico. Na parte inferior, encontram-se todos os objetos da base de dados. Cada um deles representa um agrupamento unitário, ou seja, formado por apenas um único elemento. Na primeira etapa, os dois agrupamentos com menor distância (mais similares) vão se unir, formando um agrupamento composto por dois objetos. As medidas de distância entre os demais agrupamentos e o novo agrupamento são recalculadas, e novamente os dois agrupamentos com menor distância entre si são unidos. Esse processo se repete até que todos os objetos estejam unidos em um mesmo agrupamento (parte superior da Figura 8). O nível de similaridade, representado à esquerda da Figura 8, indica a distância que possuíam os agrupamentos no momento em que

foram unidos. Essa medida é importante, pois serve como um indicador do momento ideal em que se deve parar o processo de agrupamento e obter-se os grupos.

Figura 8 - Representação do processo de agrupamento com o método hierárquico.



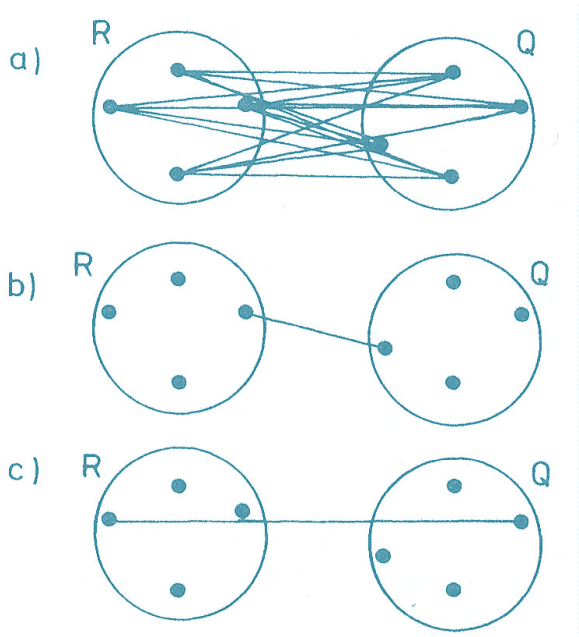
Fonte: Rosa (2014)

Quando os agrupamentos são unitários, a medida de similaridade é obtida para cada par de objetos. Mas como fazer para obter a medida de similaridade entre agrupamentos formados por mais de um objeto? Essa resposta é dada pelo algoritmo de partição. O algoritmo de partição define de que forma será aplicada a medida de similaridade para obter as distâncias entre dois agrupamentos.

A Figura 9 ilustra as formas diferentes de obter a distância entre dois grupos a partir de três dos algoritmos mais utilizados: O algoritmo da Ligação Média (*Average Linkage*), o algoritmo da Ligação Completa (*Complete Linkage*) e o algoritmo da Ligação Simples (*Single Linkage*). O primeiro mede a distância entre dois grupos a partir da média da distância entre todos os seus objetos. Com o segundo, a distância é medida a partir dos dois objetos mais próximos. O terceiro algoritmo define a distância entre dois grupos a partir da distância entre seus objetos mais distantes.

Detalhes sobre esses e outros algoritmos podem ser encontrados em Bussab et al. (1990), Kaufman e Rousseeuw (2005) e Mingoti (2007).

Figura 9 – Representação de alguns algoritmos de partição: (a) Ligação Média. (b) Ligação Simples. (c) Ligação Completa.



Fonte: Kaufman e Rousseeuw (2005).

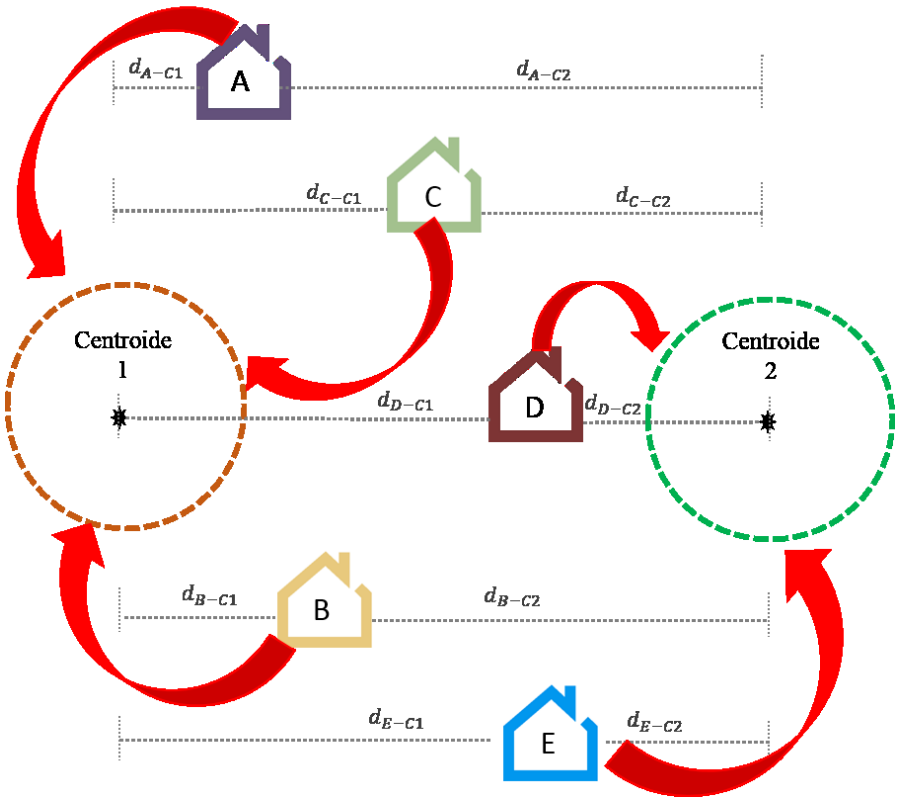
Uma característica interessante do método hierárquico é a possibilidade de acompanhar a construção dos agrupamentos etapa por etapa, através do dendograma. A partir do dendograma é possível também ter uma indicação da melhor quantidade de agrupamentos a serem formados. Entretanto, isso também impede que algum objeto seja redesignado a outro agrupamento no final do processo, ou seja, uma vez que um objeto é designado a um determinado agrupamento, ele nunca mais é desagregado no processo. No método não hierárquico, por sua vez, os agrupamentos são formados de forma interativa e qualquer objeto pode ser redesignado a outro agrupamento ao final do processo, o que garante formações mais homogêneas. Nesse método, é preciso, inicialmente, sugerir uma quantidade de agrupamentos a serem formados. São determinados também pontos sementes, que representam um valor abstrato para o centro de cada grupo (centroide). Assim, em vez de criar uma matriz de similaridade a partir da combinação de todos os objetos, as distâncias são obtidas entre cada objeto e os pontos sementes, e os objetos designados ao agrupamento cuja distância ao ponto semente é menor. Os centroides são recalculados e assumem um novo valor de referência do centro do grupo. As distâncias são novamente calculadas entre os objetos e os novos centros dos grupos, e os objetos que se aproximarem mais do centro de outro grupo, são redesignados.

Esse processo se repete até se alcançar a convergência. Dessa forma, reduz-se consideravelmente o custo de operação, em relação ao método hierárquico.

Os algoritmos baseados em técnicas não hierárquicas mais conhecidos são o algoritmo K-means e o algoritmo k-medoids. O primeiro computa o centroide do grupo como a média multivariada entre seus elementos, enquanto o segundo utiliza a mediana. Esses e outros algoritmos podem ser aprofundados em Jain et al. (1999), Hair et al. (2009), e Han et al. (2011).

A Figura 10 representa esquematicamente como se dá a formação dos agrupamentos a partir de métodos não hierárquicos.

Figura 10 - Representação do processo de agrupamento com o método não hierárquico.



Fonte: Rosa (2014)

A análise de *cluster* é um método de análise de dados não supervisionado e não inferencial. Em outras palavras, não há, de início, conhecimento para julgar uma formação como adequada ou mesmo o que deveria estar sendo agrupado. Por esse motivo, adicionar uma etapa de validação dos resultados, muitas vezes negligenciada em estudos, é um passo muito importante enquanto forma de assegurar a adequada significância dos grupos encontrados, dentro contexto do problema de pesquisa.

Conforme apontado por Jain et al. (1999), a necessidade de validação dos resultados de uma análise de *cluster* tem por base duas questões; a primeira é devido ao fato de que, independentemente de haver ou não uma estrutura natural dentro de uma amostra de dados, a análise sempre resultará em uma formação de agrupamentos. A segunda trata do fato de que as propriedades de cada algoritmo se adequam melhor a um ou outro problema de pesquisa.

A validação nada mais é do que a adoção de procedimentos de análise da qualidade do *cluster*, visando identificar o processo de formação de grupos que melhor se ajusta aos dados amostrados, de forma quantitativa e objetiva (HALKIDI et al., 2002a, HALKIDI et al., 2002b, SELVI; PARILAMA, 2018). Infelizmente, ainda não há uma medida de validação reconhecida como mais adequada. Há, entretanto, três categorias de medidas de validação: medidas internas, medidas externas e medidas relativas.

As medidas internas de validação são medidas não supervisionadas, ou seja, derivadas do próprio banco de dados. O objetivo é identificar se a formação encontrada é intrinsecamente apropriada para os dados em análise. Nesse caso, a análise de qualidade da clusterização considera a separação dos agrupamentos e a compacidade dos *clusters* formados. As medidas externas baseiam-se em análise supervisionada. Isso implica no conhecimento a priori de classes existentes ou fundamentação teórica que possa embasar a capacidade de formar agrupamentos correspondentes a essas classes. Por fim, há as medidas de validação relativas. Essa categoria remete a medidas que utilizam dados a partir de diferentes parâmetros, que estão relacionados aos objetos, mas que não foram envolvidos na análise de *cluster*. "Os critérios relativos comparam duas ou mais estruturas e medem o seu mérito relativo em determinado aspecto" (OLIVEIRA, 2010).

2.2.3. Estudos de desempenho térmico e energético de edificações envolvendo a análise de agrupamentos

Gaitani et al. (2010) utilizaram a análise de agrupamentos para determinar modelos de referência e uma ferramenta para classificação energética de edificações educacionais, na Grécia. A partir de uma base de dados sobre consumo de energia para aquecimento e iluminação, padronizada com o método *z-scores*, a análise foi aplicada dividindo a amostra em cinco grupos distintos, cada um considerado como uma classe de energia. A análise utilizou o algoritmo K-means para a partição dos objetos. Foram consideradas sete variáveis: superfície de aquecimento (m^2), idade de construção (anos), isolamento térmico (sim ou não), número de

salas, número de alunos, número de horas operadas por dia e idade do sistema de aquecimento. Os pesquisadores adotaram como modelo de referência uma edificação real, cuja distância ao centro do agrupamento mais se aproximava, obtendo, portanto, cinco modelos distintos. Os resultados obtidos usando esses modelos podem ser usados para identificar as melhores estratégias e assim determinar metas de desempenho energético em edifícios.

Yu et al. (2011) aplicaram a análise de agrupamento em cerca de 80 edifícios residenciais no Japão, em seis diferentes localidades, com o objetivo de criar um método para avaliar a influência do usuário no consumo de energia em edificações. Segundo os autores, o método se baseia na segregação da amostra de edificações em grupos onde diversos fatores externos ao usuário impactam o consumo energético sejam semelhantes em cada grupo. Dessa forma, a diferença entre o consumo de energia das edificações do mesmo grupo se daria, exclusivamente, pelo usuário. Yu et al. (2011) aplicaram a padronização min-max para normalizar os dados, além da GRA (*Grey Relational Analysis*) para dar pesos a cada variável de acordo com o impacto que exercem sobre a variável resposta. A análise de *cluster* foi realizada combinando a distância Euclidiana ao algoritmo K-means, dando origem a quatro grupos. O potencial de economia de energia foi avaliado a partir de um modelo de referência, representado pelo objeto com menor distância ao centroide do agrupamento. A comparação do maior consumo com o modelo de cada grupo comprovou que poderia haver economia de 281MJ/m^2, 250 MJ/m^2, 198MJ/m^2 e 220 MJ/m^2 em cada um dos quatro grupos, também indicando o uso do ar-condicionado como maior potencial de economia. Os autores concluíram que os resultados permitirão priorizar os esforços para reduzir o consumo de energia, além de ajudar na modelagem do comportamento do usuário nas simulações.

Um estudo em edificações comerciais na Tailândia, desenvolvido pelos pesquisadores Petcharat et al. (2012), buscou estimar o potencial de economia com iluminação dessas edificações. Para isso, os autores aplicaram três diferentes métodos de análise: o método utilizado atualmente, uma abordagem tradicional e uma análise através da análise de *cluster*. As edificações levantadas foram classificadas em três tipologias, conforme o seu uso: escolas, hospitais e hotéis e lojas de departamento. Foram obtidos, através de levantamentos em banco de dados existente, para cada uma das edificações, a densidade de potência com iluminação instalada.

Primeiramente, aplicou-se a análise atual. Nessa análise, os dados do valor da densidade de potência instalada em iluminação de cada edificação foram comparados com aqueles determinados como meta pelo ato Promoção de Conservação de Energia (*Energy*

Conservation Promotion). Na análise pela abordagem tradicional, os dados de densidade de potência instalada em iluminação obtidos também são comparados com o valor estipulado pela meta, mas, dessa vez, utilizando o valor da média aritmética. Por fim, a análise de *cluster* foi aplicada. Para a realização dos agrupamentos, adotou-se o algoritmo *Expectation-Maxization*, que calcula a probabilidade de determinado objeto pertencer a um agrupamento ou outro. A partir da formação dos agrupamentos, o centroide de cada grupo, que representa a média de todos os atributos, foi selecionado como valor de referência de cada grupo. Esses valores também foram comparados com o valor da meta.

A partir dos resultados, foi possível estimar o potencial de economia de energia em iluminação dessas edificações. Os autores verificaram que os resultados mais precisos foram obtidos através da análise de *cluster*, evidenciando a aplicabilidade do método para estudos de eficiência energética. Adicionalmente, os autores sugerem que outras variáveis sejam adicionadas em estudos futuros, como as propriedades térmicas do envelope e a densidade de potência instalada dos demais equipamentos, obtendo-se, assim, um cenário mais completo.

A análise de *cluster* também foi aplicada em um estudo em habitações de baixa renda em Londrina, PR, desenvolvido por Giglio et al. (2014). O objetivo principal era verificar o potencial de economia de energia com o uso de aquecedores solares para o aquecimento de água. Ao todo, 27 atributos foram utilizados pelos autores para caracterizar as famílias de cada habitação. Eles descreviam as características socioeconômicas das famílias e dados sobre o consumo de água e energia, como hábitos rotineiros e fatura. Também foram levantados dados subjetivos, como o nível de satisfação do usuário em relação ao uso do novo equipamento (aquecimento solar). As variáveis foram tratadas de forma a transformar todas em valores numéricos. Assim, as análises poderiam ser realizadas conjuntamente. Para isso, todos os dados foram registrados como variáveis binárias ou descritas em *ranks*. Outro tratamento adotado pelos pesquisadores foi a padronização estatística (*z-score*), que normaliza os dados em função da média e desvio padrão da amostra, para cada variável. Para detecção de objetos atípicos, adotou-se a medida D^2 de Mahalanobis, com p_{valor} menor que 0,001 (são considerados atípicos os objetos com probabilidade menor do que a adotada). Organizados os dados, deu-se início à análise de *cluster*. A análise foi rodada com o programa SPSS. O algoritmo não hierárquico K-means foi selecionado para a partição dos grupos e a medida de distância adotada foi a distância euclidiana. Com o método não hierárquico, é necessário definir previamente o número de agrupamentos desejados. Como não havia um valor predefinido, foram realizados testes com diferentes quantidades de grupos. Ao final, adotou-se a solução com cinco grupos. Seis objetos

foram detectados como atípicos, sendo excluídos do estudo. Os agrupamentos foram descritos conforme o potencial de economia de energia que os caracterizava. Os pesquisadores observaram que dos cinco, apenas dois grupos apresentaram bom potencial para economia de energia (representando quase metade da amostra). Os autores concluíram que a aplicação da análise de agrupamentos foi uma ferramenta imprescindível para a realização do estudo, permitindo a identificação de resultados adequados a cada perfil de usuário.

Schaefer e Ghisi (2016) determinaram dois modelos de referência para habitação de baixa renda na região de Florianópolis, no sul do Brasil, para realizar estudos de eficiência energética para este grupo de edifícios. Os dados foram coletados em 100 habitações, ao longo de um ano, referentes à sua geometria, tais como: dimensões internas e externas, distribuição espacial de ambientes, orientação solar, bem como informações sobre aberturas como manobra de operação, dimensões e sombreamento. Os dados coletados foram submetidos à análise de *cluster*, cujo objetivo era encontrar subgrupos dentro de uma amostra com alta homogeneidade interna e alta heterogeneidade entre os grupos. Para isso, utilizou-se a combinação de técnicas hierárquicas (algoritmo Ward) e não hierárquicas (K-means). Testes de hipótese também foram realizados para verificar a independência estatística entre os grupos encontrados. Dois modelos de referência foram determinados a partir dos prédios mais próximos do centro de cada grupo: um edifício menor, com dois quartos, sala de estar combinada e cozinha com aproximadamente $37m^2$, e outro com três quartos, sala de estar independente e cozinha, com $76m^2$. Finalmente, os modelos foram submetidos à simulação computacional. Os modelos que representam as características reais de cada edifício a partir da amostra também foram submetidos à simulação, nas mesmas condições externas. Os resultados obtidos com as simulações dessas casas foram comparados com os resultados obtidos usando os modelos de referência para verificar se o modelo poderia representar o desempenho térmico do grupo. No final, verificou-se que o valor de grau-hora dos modelos, um indicador de desempenho utilizado na pesquisa, estava muito próximo da mediana do grupo, sugerindo boa representatividade dos modelos.

No estudo desenvolvido por Geyer et al. (2017), a análise de *cluster* foi aplicada com o objetivo de identificar edificações que reagem similarmente ao serem aplicadas medidas de retrofit. Dessa forma, diferentes estratégias de eficiência poderiam ser determinadas para diferentes grupos de edificações, baseadas no custo-benefício de sua implantação. Conforme apontam os autores, o desenvolvimento de estratégias de eficiência voltadas a atender grupos de edificações permite o desenvolvimento de estratégias de forma mais eficiente, quando comparado a estudos utilizando edifícios individualmente.

Para isso, os autores aplicaram medidas de eficiência a um conjunto de edificações residenciais na Suíça, obtendo a redução na emissão de CO_2 a partir de cada medida adotada. Com esses dados, as edificações foram separadas em grupos similares a partir da análise de *cluster*, utilizando diferentes algoritmos de partição, baseados em métodos de distância e cujo agrupamento segue técnicas hierárquicas aglomerativas. Também foi adotado o algoritmo K-means, com propósito de comparação. Quanto às técnicas hierárquicas, cinco diferentes algoritmos foram aplicados: (1) *Unweighted Pair Group with Arithmetic Mean* (Método da média aritmética não ponderada), (2) *Unweighted Pair Group using Centroid* (Método do centroide não ponderado), (3) *Weighted Pair Group with Arithmetic Mean* (Método da média aritmética ponderada), (4) *Weighted Pair Group using Centroid* (Método do centroide ponderado), (5) *Shortest distance* (Método da menor distância). O número de *clusters* a serem formados foi fixado em sete e o Método da média aritmética não ponderada foi adotado como referência. O critério de comparação utilizado baseou-se nos desvios obtidos com a designação das edificações para cada *cluster*, através dos diferentes algoritmos.

Os resultados da comparação mostraram que houve pouca diferença para os resultados obtidos entre o caso de referência e o algoritmo *K-means*. Entretanto, ao comparar os resultados obtidos através dos demais métodos hierárquicos, encontrou-se diferenças mais significativas. O fator de ponderação foi o que levou às maiores diferenças (ponderar ou não aqui significa, ao fundir dois *clusters*, adotar a nova média ou centroide considerando apenas a média ou centroide dos *clusters* que se unem ou então recalcular a média e centroide considerando todos os objetos pertencentes aos dois grupos). Essa diferença é obtida pois, quando não ponderado, o *cluster* que possuía previamente número maior de objetos exerce influência maior que aquele com menos objetos. O método com os maiores desvios foi o Método da menor distância. Em função disso, os autores recomendam que o foco seja não apenas no algoritmo de agrupamento, mas no método de agregação utilizado por esses algoritmos.

Quanto à formação dos agrupamentos, os autores encontraram cinco grupos de edificações que reagiram similarmente às medidas. O *cluster* 1 era composto por edificações com alta emissão de CO_2. O *cluster* 2 compreendia edificações com menor isolamento, mas menor demanda energética para aquecimento, devido ao seu formato compacto e usuários econômicos. O *cluster* 3 era composto por um grupo de edificações menores, com média demanda de aquecimento e emissões de CO_2. O *cluster* 4 era formado por edificações de baixo ou zero consumo de energia, por ter baixa demanda ou possuir sistema de aquecimento eficiente. Por fim, o último grupo, *cluster* 5, é composto por edificações com alto consumo de

energia e emissão de CO_2. As principais estratégias adotadas foram o isolamento térmico para os *clusters* 3 e 5, substituição do sistema de aquecimento para os *clusters* 1, 2, 3 e 5. Nenhuma medida apresentou bom custo benefício para o *cluster* 4 e todas as medidas foram eficientes para as edificações do *cluster* 5. Com essas informações, os autores desenvolveram uma estratégia de transformação da vila a longo prazo, onde seriam transformadas, primeiramente, as edificações dos *clusters* com maior emissão, e, ao longo do tempo, as edificações dos demais *clusters*.

Os autores demonstram com esse estudo a aplicabilidade do uso da análise de *cluster*, principalmente se aplicada aos resultados de medidas de eficiência. Os resultados mostraram que essa abordagem provê uma classificação para ações de retrofit em edificações muito melhor que a abordagem tradicional de classificar as edificações por tipologia e idade. Os pesquisadores também destacam que o método se direciona à descrição das características de grupos de edificações com propriedades similares e não à descrição detalhada de edificações individuais. Isso significa que o resultado do *cluster* não é necessariamente a melhor medida ao nível individual de cada edificação, e sim a um grupo de edificações e seu potencial com respeito às medidas combinadas.

2.3. Síntese do capítulo

O capítulo de revisão de literatura serviu para dar base conceitual ao estudo, identificando objetos potenciais para análise e investigação e lacunas nos estudos desenvolvidos até então.

Na seção sobre os modelos de referência, apresentou-se o conceito e a importância dessa ferramenta e como ela pode auxiliar em estudos de desempenho térmico e energético de todo um estoque de edificações, obtendo indicadores em larga escala (DASKALAKI et al., 2010). Viu-se, ainda, que não há um método claro e definido para sua obtenção, levando os pesquisadores a adotar métodos univariados e bastante aleatórios para sua determinação. Foram apresentadas também, conforme Loga (2008), três diferentes formas de determinar um modelo a partir de um grupo de dados: edifício exemplo, edifício real e edifício teórico. O primeiro baseia-se na literatura, enquanto os demais em dados reais. A maioria dos estudos parece basear-se no método do Edifício Real (GAITANI et al., 2010; ATTIA et al., 2012; SCHAEFER; GHISI, 2016), onde uma edificação é selecionada na amostra como modelo.

Na sequência, apresentou-se a análise de agrupamento como um possível método para obtenção desses modelos. A análise de agrupamentos é uma análise baseada em estatística

multivariada, capaz de identificar subgrupos dentro de uma amostra, agregando edificações similares e facilitando o processo de definição de um modelo (BUSSAB et al., 1999). Esse método é bastante antigo e ainda pouco usado na área de edificações, embora a pesquisa em base de dados de artigos mostrou a adoção crescente desse método em pesquisas de desempenho de edificações (PETCHARAT et al.; 2012, GIGLIO et al., 2014; LI et al., 2018, HUANG et al., 2019). A análise de agrupamentos, conforme apontado por Hair et al. (2009), é realizada em três etapas principais, que consistem no tratamento dos dados e formação de uma matriz de dados, a adoção de uma medida de similaridade e a aplicação de um algoritmo de partição. Vários são os métodos para tratamento dos dados, as medidas de similaridade e os algoritmos de partição existentes. Entretanto, os estudos atuais parecem não averiguar quais as implicações no uso das diferentes combinações entre esses fatores, não tendo sido encontrado nenhum estudo que identifique qual melhor método a ser adotado. A maioria dos estudos têm adotado a distância euclidiana (GIGLIO et al., 2014; SCHAEFER; GHISI, 2016) e o algoritmo k-mean (GAITANI et al., 2010; YU et al., 2011; SANGIREDDY et al., 2019) de forma indiscriminada, argumentando que essa combinação já foi utilizada em outros estudos, mas sem expressar de fato qualquer convicção sobre um possível melhor resultado com esses fatores do que com outros existentes.

Baseando-se nessa revisão de literatura, verifica-se a necessidade de encontrar um método eficiente para analisar o estoque edificado, o que pode ser feito usando modelos de referência. Esses modelos podem ser encontrados a partir da divisão das edificações existentes em grupos com edificações semelhantes, definindo-se para cada grupo um modelo que os represente. Esse processo pode ser realizado com a análise de agrupamentos. Entretanto, é preciso investigar os efeitos das diferentes configurações da análise de *cluster*, por exemplo, ao combinar tratamento de dados, medidas de similaridade e algoritmos de partição. Conforme apontado por Geyer et al. (2017), alguns algoritmos podem apresentar melhor habilidade para estruturação de dados de uma amostra do que outros.

Considerando esse contexto, esta tese visa encontrar um método que combine diferentes medidas e algoritmos para obtenção de modelos de referência. Esses modelos poderão ser utilizados para estudos de desempenho em larga escala, dando base a ações mitigadoras e elaboração de normas e diretrizes que orientem a construção de novas edificações deste tipo ou revitalização das existentes.

3. Método

Neste capítulo, apresenta-se o método elaborado para obtenção de modelos de referência a partir da aplicação da análise de agrupamento. Diversas configurações combinando diferentes tratamentos de dados, medidas de distância e algoritmos de partição foram utilizadas. Para isso, seis etapas sequenciais foram realizadas:

- Composição do banco de dados inicial;
- Formação das matrizes de dados;
- Aplicação da análise de agrupamento;
- Definição dos modelos de referência;
- Validação dos métodos e modelos;
- Interpretação e caracterização dos agrupamentos.

O estudo de caso deste trabalho são as habitações de interesse social da Grande Florianópolis, por representar um tema importante de pesquisa sobre eficiência energética e desempenho térmico. Na primeira etapa, informações sobre a geometria dessas edificações foram obtidas a partir de um banco de dados pré-existente. Essas informações foram utilizadas para compor a matriz de dados usada como base para todo o estudo. Paralelamente, informações sobre o desempenho térmico das mesmas habitações foram obtidas por meio de simulação computacional. Essas informações foram utilizadas para a realização de alguns procedimentos na segunda e quinta etapas.

Na segunda etapa, os dados obtidos foram tratados a partir de três métodos estatísticos: a padronização estatística, a detecção de objetos atípicos e a ponderação dos fatores. A diferente combinação desses tratamentos deu origem a cinco diferentes matrizes de dados. O estudo foi conduzido separadamente para cada uma dessas matrizes. Essa etapa foi realizada a fim de se identificar o impacto que cada um desses tratamentos tem sobre o resultado final.

Na terceira etapa, realizou-se a análise de agrupamentos. Nessa etapa, a combinação de diferentes medidas de similaridade e algoritmos de partição, baseados em técnicas hierárquicas e não hierárquicas, deu origem a 21 métodos de aplicação da análise de *cluster*. Cada uma das matrizes obtidas na segunda etapa foi submetida a todos os métodos (combinações entre medida de similaridade e algoritmo de partição), resultando em um total de 105 métodos de clusterização.

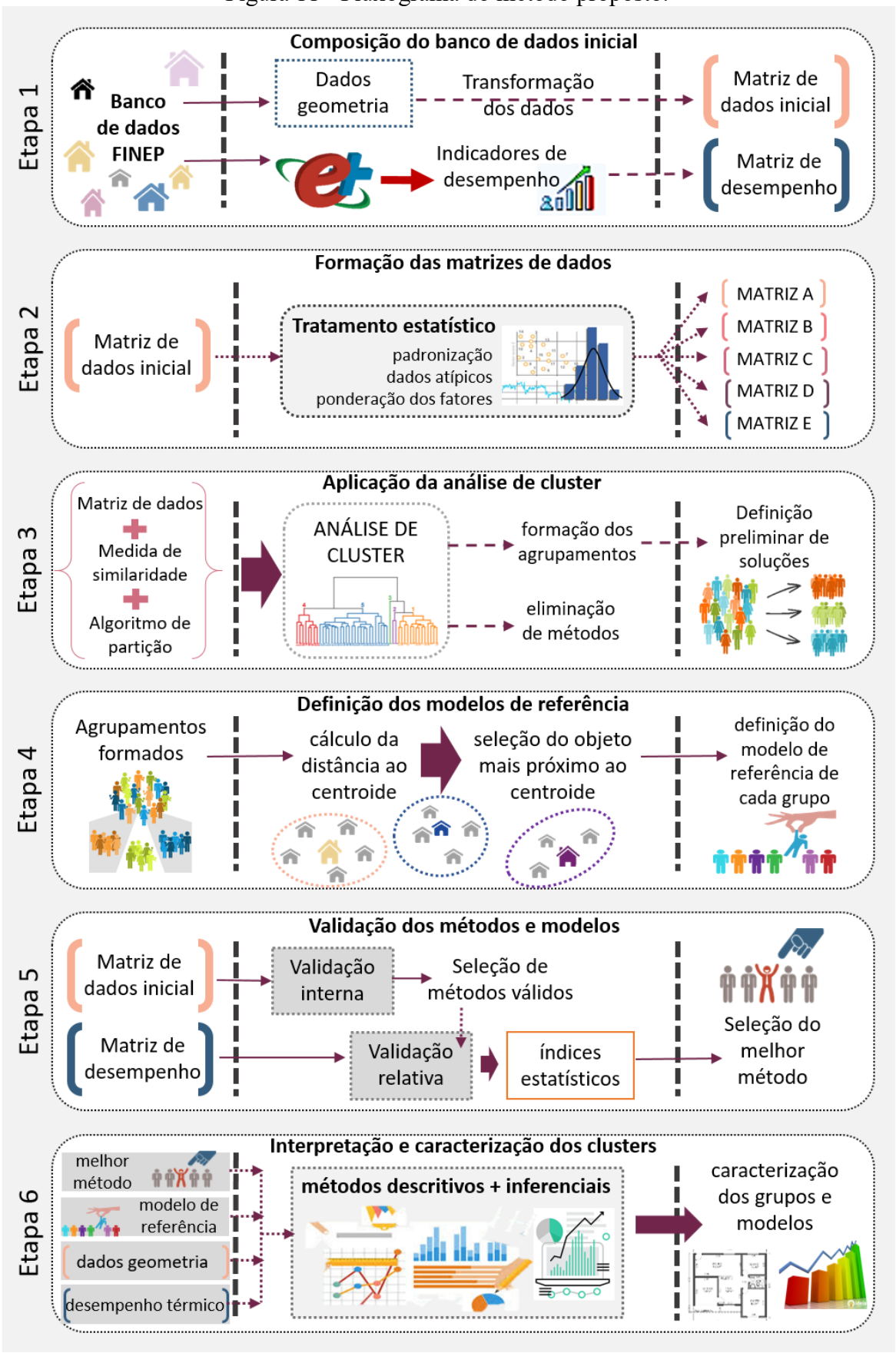
Na quarta etapa, foi atribuído, para cada método de clusterização, modelos de referência, ou seja, objetos que supostamente seriam capazes de representar de forma aproximada todas as edificações do seu grupo em estudos de desempenho, considerando os conceitos apresentados por Loga (2008).

Na quinta etapa, a validação dos métodos de clusterização e modelos obtidos foi realizada a partir de duas medidas: interna e relativa. A medida interna considera a qualidade da separação dos grupos a partir dos mesmos dados utilizados na análise de *cluster*, enquanto a medida de validação relativa utiliza dados não envolvidos na análise. Para a validação interna, foi calculada a inércia, uma medida que relaciona a compacidade dos grupos em relação à nuvem de dados. Os métodos classificados como válidos a partir do cálculo da inércia foram submetidos à validação relativa, através de seis índices estatísticos. Para essa validação, foram utilizados os indicadores de desempenho térmico obtidos através de simulação computacional, na primeira etapa. Os índices estatísticos foram aplicados de forma a se obter o erro quanto ao desempenho térmico de cada objeto em relação ao modelo de referência e a média do grupo em que está alocado. O método que apresentou os menores índices de erro (ou seja, menores desvios) foi adotado como melhor método para esse estudo.

Na sexta e última etapa, foram apresentados os grupos formados a partir do método escolhido na quinta etapa, assim como os modelos de referência definidos como característicos de cada grupo. Os grupos e os modelos foram descritos e comparados a partir das características geométricas de seus objetos e também quanto ao desempenho térmico.

A Figura 11 apresenta o fluxograma do método proposto e, na sequência, cada uma dessas etapas foi apresentada com maiores detalhes.

Figura 11 - Fluxograma do método proposto.



3.1 Composição do banco de dados inicial

O método em que se baseia esse trabalho foi desenvolvido a partir de sua aplicação em um estudo de caso de habitações de interesse social de Florianópolis. Foram criados dois bancos de dados iniciais: um refere-se à caracterização da amostra de habitações quanto à sua geometria e o outro, quanto ao desempenho térmico.

Os dados utilizados para caracterizar a geometria foram obtidos em um banco de dados existente, construído com levantamentos em campo durante o projeto de pesquisa "Uso Racional de Água e Eficiência Energética em Habitações de Interesse Social", financiado pela FINEP (Chamada Pública 07/2009). Os levantamentos foram realizados entre os anos de 2011 e 2013, em aproximadamente 120 habitações de baixa renda da região da Grande Florianópolis. Mais informações sobre esse projeto podem ser encontradas em Ghisi et al. (2015).

Os dados utilizados para caracterizar o desempenho térmico foram obtidos a partir da simulação computacional de cada habitação.

Ressalta-se que as variáveis que compuseram o banco de dados, seja quanto à geometria ou quanto ao desempenho, foram selecionadas considerando o estudo de caso proposto, cujo foco direcionou-se ao desempenho térmico de habitações de interesse social. O método apresentado nesta tese pode ser aplicado a qualquer outro estudo, como de edificações multifamiliares e comerciais, ou mesmo comportamento do usuário, que avaliem a qualidade ambiental a partir do desempenho lumínio ou eficiência energética, etc. Entretanto, é importante que, ao mudar o foco de pesquisa, as variáveis que descrevem o objeto de estudo e os indicadores de desempenho correspondam àqueles para os quais se busca respostas. Em estudos envolvendo desempenho energético, por exemplo, pode ser interessante considerar a potência instalada e padrão de uso, e adotar o consumo de energia como indicador. Nesse sentido, essa primeira etapa (composição do banco de dados) deve ser reelaborada sempre que se deseja aplicar o método proposto a outros estudos.

3.1.1 Dados referentes à geometria das edificações

Os dados levantados no projeto de pesquisa financiado pela FINEP referem-se à geometria das edificações e distribuição espacial dos seus cômodos. Para obtenção desses dados, foram levantadas em campo informações das edificações quanto às suas dimensões internas e externas, o *layout* de distribuição dos cômodos, a orientação solar da edificação e sua

inserção no contexto urbano. Também foram obtidas informações a respeito das aberturas (tais como suas dimensões, manobra de abertura e existência de proteção solar). Esses dados foram levantados com uso de trena, bússola e GPS, e anotados em um questionário semiestruturado. Informações mais detalhadas sobre o levantamento desses dados podem ser encontradas em Rosa (2014) e Ghisi et al. (2015).

A partir da base de dados levantada no projeto de pesquisa, foi criada uma matriz de dados com as informações referentes à forma das edificações, denominada aqui matriz de dados inicial. Ao todo, foram obtidas informações a respeito de 102 habitações (as demais habitações foram descartadas por possuírem dados incompletos ou inconsistentes). Essas habitações foram descritas na matriz de dados inicial por 35 variáveis, apresentadas no Quadro 1, juntamente com o termo pelo qual foram referenciadas ao longo do trabalho.

Quadro 1 - Variáveis que compõem a matriz de dados inicial.

Variáveis referentes a edificação como um todo		
Variáveis	**Unidade**	**Abreviação**
Existência de sala e cozinha conjugadas	sim/não	Slcoz
Quantidade de dormitórios	unid	N_dorm
Quantidade de ambientes de ocupação transitória	unid	N_AmbT
Quantidade de ambientes de permanência prolongada	unid	N_AmbPP
Quantidade de pavimentos	unid	N_pavto
Área total	m^2	A_tot
Área de cobertura	m^2	A_cob
Pé direito	m	PD
Volume total	m^3	V_tot
Área de parede exposta	m^2	A_par_tot
Área de janela	m^2	A_jan_tot
Proporção entre as dimensões de fachada norte e leste	adim	A_parN/A_parL
Razão entre a soma das áreas de parede exposta e cobertura, e volume total	adim	A_par_cob/ V
Razão entre a área de janela e área de parede exposta	adim	A_jan_tot/ A_par_tot
Variáveis referentes aos ambientes de permanência prolongada		
Variáveis	**Unidade**	**Abreviação**
Soma das áreas úteis dos ambientes de permanência prolongada	m^2	AU
Área útil dos ambientes de convivência social	m^2	AU_social
Área útil dos ambientes de convivência íntima	m^2	AU_íntimo
Soma dos volumes dos ambientes de permanência prolongada	m^3	Vol

Quadro 1 - Variáveis que compõem a matriz de dados inicial (continuação).

Variáveis referentes aos ambientes de permanência prolongada		
Variáveis	**Unidade**	**Abreviação**
Volume dos ambientes de convivência social	m^3	Vol_social
Volume dos ambientes de convivência íntima	m^3	V_íntimo
Área de parede exposta	m^2	A_par
Área de janela	m^2	A_jan
Razão entre a área de janela e área de parede exposta	adim	A_jan/ A_par
Área de parede exposta na fachada norte	m^2	A_parN
Área de janela na fachada norte	m^2	A_janN
Razão entre a área de janela e área de parede exposta na fachada norte	adim	A_janN/ A_parN
Área de parede exposta na fachada leste	m^2	A_parL
Área de janela na fachada leste	m^2	A_janL
Razão entre a área de janela e área de parede exposta na fachada leste	adim	A_janL/ A_parL
Área de parede exposta na fachada sul	m^2	A_parS
Área de janela na fachada sul	m^2	A_janS
Razão entre a área de janela e área de parede exposta na fachada sul	adim	A_janS/ A_parS
Área de parede exposta na fachada oeste	m^2	A_parO
Área de janela na fachada oeste	m^2	A_janO
Razão entre a área de janela e área de parede exposta na fachada oeste	adim	A_janO/ A_parO

3.1.2. Dados referentes ao desempenho térmico das habitações

Para acessar o desempenho térmico das 102 habitações da amostra, cada uma delas foi submetida à simulação computacional, resultando em 102 arquivos. Os dados obtidos por meio das simulações de cada uma das habitações foram tratados e compuseram a matriz de dados de desempenho. Os procedimentos adotados estão detalhados a seguir.

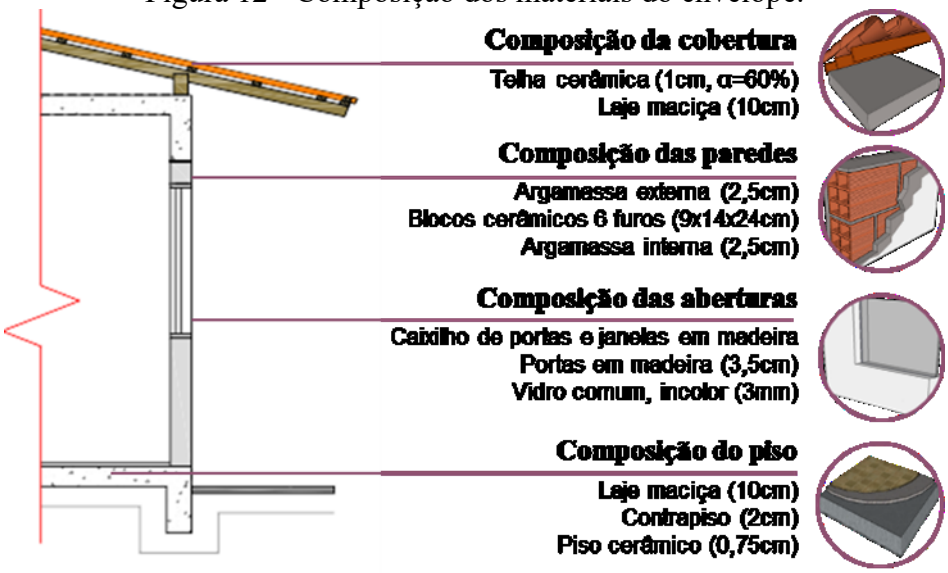
3.1.2.1. Configurações gerais dos arquivos de simulação

As configurações gerais dizem respeito às configurações dos arquivos de simulação comuns a todos os modelos. O programa *EnergyPlus*, versão 8.9 (DOE, 2018) foi selecionado para realizar as simulações computacionais. Optou-se por utilizar esse programa pois trata-se de um programa livre, confiável e com interface amigável.

As simulações foram realizadas considerando o clima de Florianópolis, SC, que se encontra na Zona Bioclimática 3 (NBR 15220-3). As características climáticas foram representadas nas simulações com o arquivo climático TRY, em formato epw, obtido no site do Laboratório de Eficiência Energética de Edificações da UFSC (www.labeee.ufsc.br).

A configuração dos materiais e sistemas construtivos foi obtida em Schaefer e Ghisi (2016). A Figura 12 apresenta a descrição dos materiais que compõem a envoltória. As propriedades térmicas dos sistemas construtivos foram obtidas no sítio do projeto Projetando Edificações Energeticamente Eficientes (MMA, 2018). As propriedades térmicas dos materiais da envoltória estão apresentadas na Tabela 1.

Figura 12 - Composição dos materiais do envelope.



Fonte: Adaptado de Schaefer e Ghisi (2016).

Tabela 1 – Propriedades térmicas dos sistemas construtivos.

Sistema construtivo	**Transmitância térmica (W/(m²K))**	**Capacidade térmica (kJ/(m²K))**	**Absortância da camada externa**	**Fator solar**
Paredes	2,39	152	50%	-
Cobertura	2,05	238	60%	-
Piso	4,00	294	-	-
Janelas	5,70	-	-	0,87

Fonte: MMA (2018)

As configurações quanto às cargas internas dizem respeito às rotinas de ocupação dos ambientes e operação de equipamentos e iluminação. As cargas internas e o padrão de operação

foram configurados conforme sugerido em CB3E (2018). A Tabela 2 apresenta o padrão de uso e densidade de carga de equipamentos. A Tabela 3 apresenta a taxa metabólica conforme a atividade principal realizada em cada ambiente. O padrão de ocupação está apresentado na Tabela 4 e o padrão de iluminação está apresentado na Tabela 5.

Nas habitações levantadas, a existência de ambientes com sala e cozinha conjugadas é frequente. Por esse motivo, considerou-se nesse estudo taxas metabólicas para esse tipo de ambiente diferente daquelas que possuem apenas sala de estar. O valor adotado foi obtido na ASHRAE (2010). Também se adotou padrão de uso e densidade de cargas para os dormitórios. Esses valores foram obtidos em Schaefer e Ghisi (2016). Foram consideradas duas pessoas por dormitório, para habitações com até dois dormitórios, e uma pessoa por dormitório adicional. Nas salas, considerou-se a ocupação correspondente à soma da quantidade de pessoas que ocupam os dormitórios.

Tabela 2 – Padrão de uso e densidade de carga interna de equipamentos.

Ambiente	**Período de uso**	**Potência (W)**
Sala	0h00 às 13h59	0
	14h00 às 21h59	120
	22h00 às 23h59	0
Dormitórios*	0h00 às 6h59	0
	7h00 às 21h59	30
	22h00 às 23h59	0

* Valores adotados de Schaefer e Ghisi (2016).

Fonte: Adaptado de CB3E (2018).

Tabela 3 – Taxas metabólicas adotadas por ambiente.

Ambiente (zona térmica)	**Atividade realizada**	**Taxa metabólica (W/m^2)**	**Taxa metabólica para área de pele = 1,80m^2 (W)**
Sala	Sentado, lendo, quieto	60	108
Cozinha	Cozinhando, limpando	110	198
Dormitórios	Dormindo, relaxando	45	81

Fonte: ASHRAE (2010).

Tabela 4 – Padrão de ocupação.

Ambiente	Período de ocupação	Taxa de ocupação (W)
Sala	0h00 às 12h59	0
	13h00 às 17h59	50
	18h00 às 20h59	100
	21h00 às 23h59	0
Dormitórios	0h00 às 6h59	100
	7h00 às 20h59	0
	21h00 às 23h59	100

Fonte: Adaptado de CB3E (2018).

Tabela 5 – Padrão de uso e densidade de carga interna de iluminação.

Ambiente	Período de acionamento	Potência instalada (W)
Sala	00h00 às 14h59	0
	15h00 às 20h59	60*
	21h00 às 23h59	0
Dormitórios	00h00 às 4h59	0
	5h00 às 6h59	20*
	7h00 às 20h59	0
	21h00 às 23h59	20*

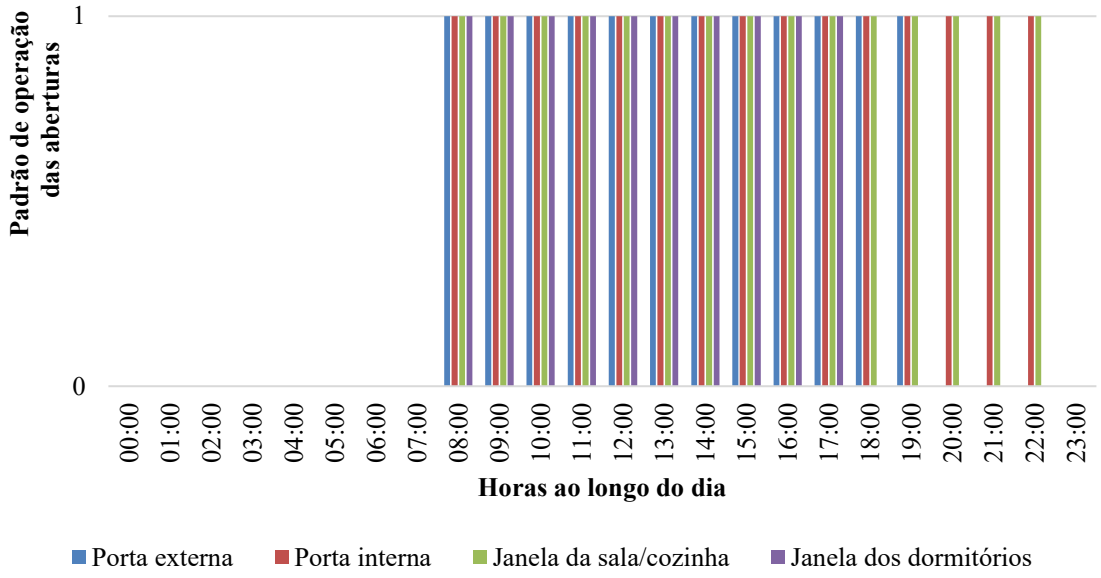
* Valores adotados de Schaefer e Ghisi (2016).

Fonte: Adaptado de CB3E (2018).

As trocas de ar entre a edificação e o meio foram realizadas com a configuração do objeto *Airflow Network*. A cada superfície por onde ocorrem as trocas de ar (portas e janelas) foram configuradas suas propriedades e detalhados seus componentes, atribuindo-lhes os coeficientes e expoentes de fluxo de ar conforme Liddament (1986). Também foi atribuída a cada superfície uma rotina de operação, apresentada na Figura 13 (SCHAEFER; GHISI, 2016). O controle da ventilação se deu a partir da combinação das rotinas adotadas e da adoção de temperatura de *setpoint*, estipulada em 19^{o}C (CB3E, 2018). A operação das aberturas obedeceu ao padrão de operação adotado desde que respeitada as condições: (1) temperatura interna maior ou igual a temperatura externa e (2) temperatura interna igual ou maior que a temperatura de *setpoint*. A adoção de um cenário considerando apenas a ventilação natural faz-se importante pois é o tipo de condicionamento mais condizente com a realidade encontrada, visto que as

habitações levantadas não apresentavam, em nenhum dos casos, sistema de ar-condicionado instalado.

Figura 13 - Padrão de operação das aberturas.



Fonte: Schaefer e Ghisi (2016).

As trocas térmicas entre a edificação e o solo foram configuradas com o objeto *GroundDomain: Slab*, do *EnergyPlus*.

3.1.2.2. Modelagem da geometria

A geometria de cada habitação foi modelada com auxílio da ferramenta *Euclid*, que funciona como um *plug-in* do programa *Sketch Up*. Esse programa facilita a modelagem das geometrias, pois tem uma interface amigável, que permite a visualização 3D dos modelos e fácil inserção dos dados. As geometrias foram modeladas conforme as características levantadas das edificações existentes. A planta baixa de cada habitação foi desenhada no programa AutoCAD e exportada para o *Sketch Up*, para facilitar a construção dos modelos. Foram mantidas suas dimensões gerais, distribuição espacial dos ambientes, dimensionamento e posicionamento das aberturas e orientação solar. Não foram considerados os beirais na modelagem da geometria devido a inconsistências observadas no banco de dados. Também não foram consideradas as inclinações das águas do telhado, adotando-as como uma superfície plana. Cada ambiente da edificação foi configurado como uma zona térmica independente,

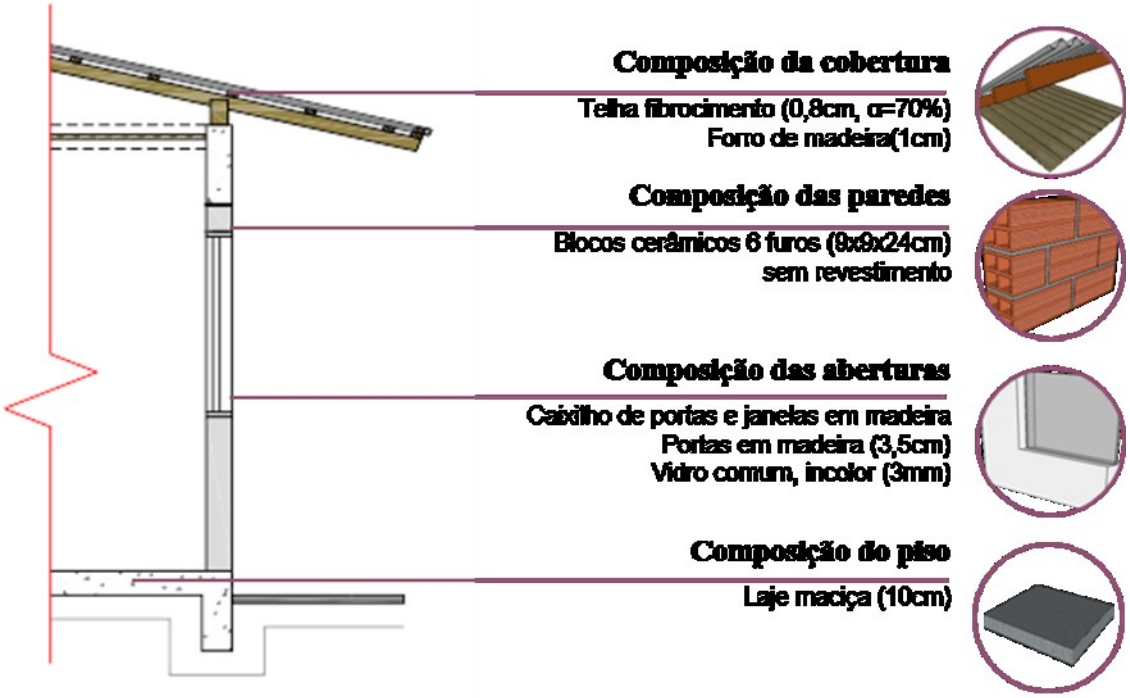
exceto nos casos em que esses ambientes configuram uma única unidade nos casos levantados, como aqueles que possuem sala e cozinha conjugadas.

3.1.2.3. Variações na configuração dos arquivos de simulação

Além da configuração base, apresentada anteriormente, foram propostas duas configurações alternativas nos arquivos de simulação: uma delas alterando os materiais da envoltória e outra alterando os padrões de uso. Essas variações foram motivadas por possibilitar o acesso ao desempenho de cada habitação também quando submetidas a diferentes cenários e foram identificadas como configurações A, B e C representando, respectivamente, o arquivo base, variação nos materiais e variação no padrão de uso.

Quanto aos materiais, foram alteradas a composição da cobertura, piso e paredes. As configurações e propriedades térmicas das portas e janelas permaneceram inalteradas. A Figura 14 apresenta a composição dos sistemas construtivos adotados e a Tabela 6, as propriedades térmicas dos materiais.

Figura 14 - Composição dos materiais do envelope.



Fonte: Adaptado de Schaefer e Ghisi (2016).

Tabela 6 – Propriedades térmicas dos novos sistemas construtivos.

Sistema construtivo	**Transmitância térmica (W/(m²K))**	**Capacidade térmica (kJ/(m²K))**	**Absortância da camada externa**	**Fator solar**
Paredes	2,93	42	60%	-
Cobertura	2,02	21	70%	-
Piso	4,40	240	-	-
Janelas	5,70	-	-	0,87

Fonte: MMA (2018)

Quanto aos padrões de uso, a nova configuração considerou a inexistência de cargas internas. Dessa forma, não foram configurados os objetos que descrevem a ocupação, operação de equipamentos e iluminação nem potência instalada com equipamentos e iluminação. As aberturas foram configuradas para manter-se fechadas por todo o período.

Essas duas variações foram propostas com a intenção de ressaltar a influência das características geométricas das habitações no desempenho térmico das mesmas. O sistema construtivo proposto como variação deixa a envoltória mais sensível às condições externas, enquanto a variação quanto ao padrão de uso elimina a influência do usuário e da ventilação natural.

4.1.2.4. Composição da matriz de desempenho

A matriz de desempenho foi formada a partir dos resultados das simulações do caso base e variações, e tratados com os procedimentos descritos a seguir.

Inicialmente, foram solicitadas como variáveis de saída as temperaturas operativas de cada ambiente de permanência prolongada (salas, sala e cozinha conjugadas, e dormitórios), em base horária, para o período de um ano. Os valores de temperatura operativa foram tratados com as Equações 1 e 2 a fim de se obter um indicador de desempenho considerando as horas em que a temperatura operativa esteve acima ou abaixo de uma faixa de temperatura limite, indicando necessidade de resfriamento ou aquecimento, respectivamente.

Comumente, adota-se os valores de 18ºC e 26ºC como limites para o cálculo de graus-hora de aquecimento e resfriamento (como sugerido em CB3E, 2018, para edificações naturalmente ventiladas). Entretanto, como o objetivo do estudo não é analisar o conforto térmico e sim identificar diferenças no desempenho térmico das habitações, considerou-se que manter esse intervalo seria inadequado, pois poderia igualar em termos de graus-hora ambientes

com diferenças de até de 8ºC na temperatura operativa. De forma a minimizar esse fator e potencializar as diferenças de desempenho entre as habitações, reduziu-se 2ºC dos limites inferior e superior, passando a adotar faixa de temperatura limite de 20ºC a 24ºC.

$$Id_{resf} = \sum_{i=1}^{8760} se \begin{cases} T_o > 24\,, (T_o - 24) \\ T_o < 24, 0 \end{cases} \quad (1)$$

$$Id_{aquec} = \sum_{i=1}^{8760} se \begin{cases} T_o < 20, (20 - T_o) \\ T_o > 20, 0 \end{cases} \quad (2)$$

Onde:
Id_{resf} é o indicador de desempenho para resfriamento (ºCh);
Id_{aquec} é o indicador de desempenho para aquecimento (ºCh);
T_o é a temperatura operativa interna de cada ambiente em cada hora do ano (ºC).

Habitações com diferentes quantidades de ambientes e distribuição espacial (*layout*) foram comparadas nas etapas seguintes desse estudo. Para estabelecer uma medida de comparação, ponderou-se a variável de saída pelo volume de cada ambiente, obtendo-se dois indicadores (de resfriamento e aquecimento) para a área social (salas) e dois para a área íntima (dormitórios). As Equações 3 e 4 apresentam a expressão numérica utilizada para calcular o indicador de desempenho ponderado de resfriamento e aquecimento.

$$ID_R = \sum \frac{(Id_{resf}\ x\ V_{amb})}{V} \quad (3)$$

$$ID_A = \sum \frac{(Id_{aquec}\ x\ V_{amb})}{V} \quad (4)$$

Onde:
ID_R é a média ponderada do indicador de desempenho de resfriamento de cada ambiente pelo volume (ºCh);
ID_A é a média ponderada do indicador de desempenho de aquecimento de cada ambiente pelo volume (ºCh);
Id_{resf} é o indicador de desempenho para resfriamento (ºCh);
Id_{aquec} é o indicador de desempenho para aquecimento (ºCh);
V_{amb} é o volume do ambiente(m^2);
V é a soma dos volumes dos ambientes (m^2).

Esses procedimentos resultaram na criação de um banco de dados com cada habitação sendo descrita por doze variáveis, referentes aos indicadores de de desempenho, e listadas no Quadro 2.

Quadro 2 - Variáveis submetidas à medida de validação relativa.

Nome da variável	Indicador de desempenho	Ambiente	Variação na configuração
ID_R_dorm_A	Resfriamento	Dormitórios	Arquivo base
ID_A_dorm_A	Aquecimento	Dormitórios	Arquivo base
ID_R_sala_A	Resfriamento	Salas	Arquivo base
ID_A_sala_A	Aquecimento	Salas	Arquivo base
ID_R_dorm_B	Resfriamento	Dormitórios	Sistemas construtivos
ID_A_dorm_B	Aquecimento	Dormitórios	Sistemas construtivos
ID_R_sala_B	Resfriamento	Salas	Sistemas construtivos
ID_A_sala_B	Aquecimento	Salas	Sistemas construtivos
ID_R_dorm_C	Resfriamento	Dormitórios	Ocupação
ID_A_dorm_C	Aquecimento	Dormitórios	Ocupação
ID_R_sala_C	Resfriamento	Salas	Ocupação
ID_A_sala_C	Aquecimento	Salas	Ocupação

3.2. Formação das matrizes de dados

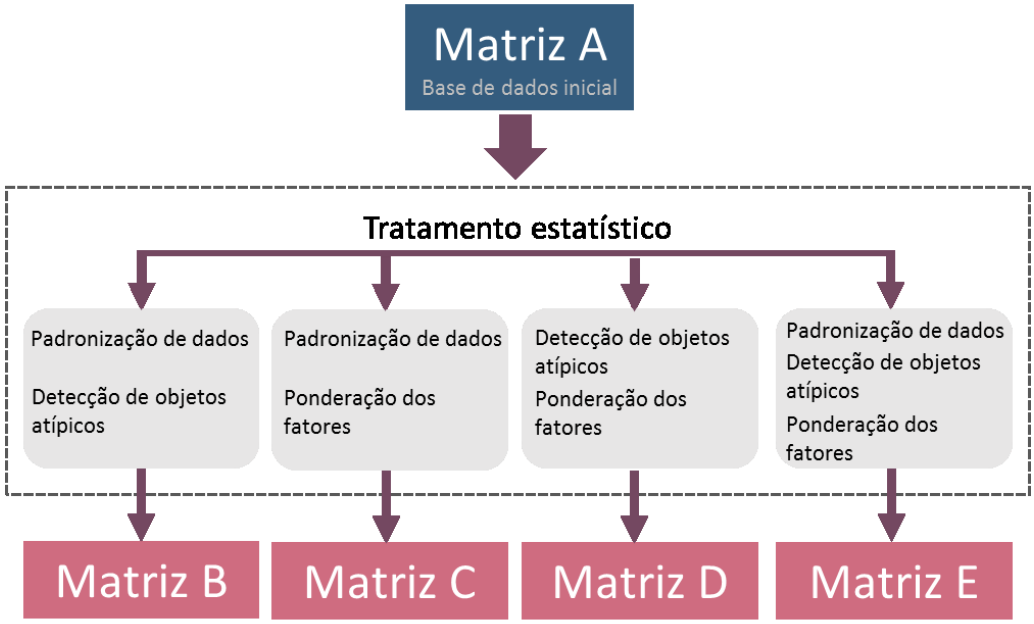
Uma importante etapa antes da análise de *cluster* é o tratamento dos dados, ou seja, a forma como são organizados e descritos. Essa etapa é fundamental para a obtenção de bons resultados, pois a forma como os dados são dispostos pode impactar sobremaneira e negativamente os resultados.

A formação das matrizes de dados corresponde, portanto, à etapa onde os dados obtidos na seção 3.1.1 foram tratados antes de serem submetidos à análise de *cluster*. Essa etapa foi proposta com o intuito de verificar se e como diferentes formas de organizar e tratar os dados impactam o resultado da análise. O tratamento dos dados consiste em aplicar procedimentos estatísticos de forma a eliminar vieses como a presença de dados pouco comuns na amostra ou a diferença de amplitude entre valores de variáveis com unidades de medida diferentes.

A partir da matriz de dados inicial, os dados foram submetidos a diferentes tratamentos: padronização dos dados, detecção de objetos atípicos e ponderação dos fatores. Estes três tratamentos de dados foram combinados de formas diferentes, dando origem a cinco matrizes de dados, identificadas de A a E. Cada uma delas foi submetida separadamente à análise de *cluster* nas etapas seguintes (foram conduzidos cinco estudos paralelos, cada um a partir de uma das matrizes).

A Matriz A corresponde a matriz de dados inicial, ou seja, sem nenhum tratamento. Para formar a Matriz B, os dados foram submetidos à padronização estatística e à detecção de objetos atípicos. A Matriz C foi formada a partir da padronização estatística e ponderação dos fatores, mas não à detecção de objetos atípicos. A Matriz D foi composta por dados submetidos à detecção de objetos atípicos e ponderação dos fatores. A Matriz E foi formada a partir da aplicação de todos os três tratamentos. A Figura 15 apresenta um resumo quanto à formação das matrizes de dados a partir da combinação dos diferentes tratamentos. Cada um dos tratamentos está descrito nas seções a seguir.

Figura 15 - Formação das Matrizes de dados.



3.2.1. Padronização dos dados

Padronizar (ou normalizar) os dados significa deixá-los com amplitudes proporcionalmente semelhantes, independentemente da sua unidade de medida. Diferentes variáveis possuem diferentes unidades de medidas, sendo a dispersão dos seus dados

determinada por valores maiores ou menores. Por exemplo, a amplitude de valores encontrada para uma variável como "área total", que poderia variar de 20m^2 a 100m^2, é muito maior do que o que se encontraria para variáveis como "quantidade de pavimentos", que poderia variar de dois a seis. Essa diferença de medida, segundo a literatura (BUSSAB et al., 1990; HAIR et al., 2009), pode provocar resultados indesejáveis na análise de *cluster*, pois uma variável com maior valor de dispersão teria um impacto proporcionalmente maior no valor da medida de similaridade (um dos procedimentos da análise de *cluster*). Em função disso, um dos tratamentos de dados mais indicados na análise de *cluster* é a padronização dos dados.

Há várias formas de se obter dados padronizados. Nesta tese, optou-se pela padronização estatística (*z-scores*), por ser a forma mais comum de padronização (BUSSAB et al., 1990). Essa medida representa o quanto um objeto se afasta da média em termos de desvio padrão, assumindo normalmente valores entre -3 e 3. O escore z é positivo quando a medida de determinado dado está acima da média e negativa, quando abaixo. A Equação 4 apresenta a expressão numérica utilizada para transformar variáveis em uma escala escores Z. A padronização estatística foi aplicada a cada variável separadamente.

$$Z_{x_i} = \frac{(x_i - \bar{x})}{s} \tag{4}$$

Onde:
Z_{x_i} é o valor padronizado de x;
x_i é o valor da variável para cada objeto;
$\bar{x}$ é a média dos valores de determinada variável;
s é o desvio padrão dos valores de determinada variável.

3.2.2. Detecção de objetos atípicos

Objetos atípicos são objetos encontrados na amostra, mas que não seguem um padrão similar aos demais objetos. Esses objetos podem representar grupos muito pequenos dentro da população ou casos não convencionais, que não apresentam, portanto, importância para a representatividade do objeto de estudo. Devem, portanto, ser identificados e retirados da amostra sempre que possível.

Dado o caráter multivariado dos objetos em análise nesse estudo, os objetos atípicos foram detectados por meio da medida D^2 de Mahalanobis (Equação 5). Conforme descrito por Rosa (2014), o D^2 de Mahalanobis é uma medida obtida a partir de uma regressão multivariada,

cujas distâncias de cada objeto ao centroide do agrupamento (média multivariada) são calculadas, dada a covariância (variância multivariada) da distribuição amostral. Não há uma padronização que especifique um valor limite de D^2 para identificação do objeto atípico, sendo, entretanto, comumente usada a probabilidade associada ao D^2 como fator decisivo (HAIR et al., 2009). A probabilidade associada ao D^2 foi obtida a partir da probabilidade acumulada de um valor da distribuição normal χ^2, com k níveis de liberdade, ser menor que o valor amostral correspondente. Neste estudo, foram considerados objetos atípicos todos os objetos cuja a probabilidade associada ao D^2 for menor que 0,001. Esses objetos foram retirados da matriz de dados em que este tratamento estatístico foi especificado.

$$D^2{}_{nm} = \sqrt{(x_n - x_m)C^{-1}(x_n - x_m)}' \quad (5)$$

Onde:
$D^2{}_{nm}$ é a medida de Mahalanobis;
C^{-1} é a matriz de covariâncias;
x_n é o valor de n para cada variável;
x_m é o valor de m para cada variável.

3.2.3. Ponderação dos fatores

A ponderação dos fatores diz respeito à transformação dos valores da matriz de dados a partir da atribuição de pesos às variáveis, de forma a priorizar aquelas mais influentes no desempenho das habitações da amostra. Atribui-se, portanto, peso maior àquelas mais influentes e menor às menos influentes. Dessa forma, variáveis mais influentes no desempenho tornam-se também mais influentes no resultado da análise de *cluster*.

Uma forma de identificar a influência que determinada variável exerce sobre o desempenho das habitações é por meio da análise de variância (ANOVA). Segundo Hair et al. (2009), a ANOVA é uma combinação de processos estatísticos a partir do qual é possível obter modelos que podem estabelecer as relações existentes entre as variáveis explicativas de um determinado processo. Aplicando a análise a um conjunto de variáveis e um ou mais indicadores, é possível obter um peso para ponderar o valor de cada variável em relação às demais conforme a influência que exerce sobre a variável dependente. Dessa forma, variáveis mais importantes, ou seja, aquelas que exercem maior impacto sobre o desempenho da edificação, teriam também peso maior sobre o resultado da análise de agrupamento.

Como o próprio nome diz, a análise de variância usa a variância amostral para inferir sobre as relações entre um grupo de variáveis dependentes e um grupo de variáveis independentes. Nessa etapa do estudo, a ANOVA foi aplicada com o objetivo de determinar qual ou quais variáveis podem ser consideradas influentes ou sem relevância em um experimento numérico. As características geométricas das habitações foram classificadas como variáveis independentes e os indicadores de desempenho térmico do caso base como variáveis dependentes. O fator de ponderação atribuído a cada variável corresponde ao valor-F (medida de sensibilidade), obtido para cada parâmetro e calculado por meio das Equações 6 a 11. Para compor as matrizes, cada dado foi multiplicado pelo valor de F encontrado para a variável correspondente (por exemplo, os dados referentes à quantidade de dormitórios foram multiplicados pelo valor de F calculado para a quantidade de dormitórios, os dados de área de janela, pelo valor F correspondente à área de janela, e assim por diante).

$$SQ(A) = b \times \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{6}$$

$$SQ(B) = a \times \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 \tag{7}$$

$$SQ(Total) = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 \tag{8}$$

$$SQ(AB) = SQ(Total) - SQ(Erro) - SQ(A) - SQ(B) \tag{9}$$

$$MQ(A) = \frac{SQ(A)}{a - 1} \tag{10}$$

$$F(A) = \frac{MQ(A)}{MQ(Erro)} \tag{11}$$

Onde:
a é a quantidade de níveis no parâmetro A;
b é a quantidade de níveis no parâmetro B;
$\bar{y}_{i.}$ é a média do *i*ésimo nível do parâmetro A;
$\bar{y}_{..}$ é a média de todas as observações;
$\bar{y}_{.j}$ é a média o *j*ésimo nível do parâmetro B;
y_{ij} é cada observação individual do *i*ésimo nível do parâmetro A, e do *j*ésimo nível do parâmetro B;
$SQ(A)$é a soma dos quadrados do parâmetro A;

$SQ(B)$é a soma dos quadrados do parâmetro B;
$SQ(Total)$ é a soma dos quadrados total;
$SQ(AB)$ é a soma dos quadrados da interação entre parâmetros;
$SQ(Erro)$ é a soma dos quadrados do erro;
$MQ(A)$ é a média dos quadrados do parâmetro A (analogamente para o parâmetro B);
$F(A)$ é o valor-F do parâmetro A (analogamente para os demais parâmetros).

3.3. Aplicação da análise de *cluster*

A aplicação da análise de *cluster*, ou clusterização, corresponde aos procedimentos aplicados às matrizes de dados de forma a segregar a amostra em grupos com habitações semelhantes. Para isso, medidas de similaridade e algoritmos de partição foram combinados e aplicados a cada uma das matrizes formadas na seção 3.2, de forma a se obter diferentes métodos de análise de agrupamentos. O produto da aplicação de cada método é a separação da amostra em subgrupos. Em teoria, as diferentes combinações de medidas de similaridade, algoritmos de partição e tratamento de dados podem resultar em diferentes formas de separar a amostra. Definiu-se também, para cada método, a solução quanto à quantidade ideal de grupos formados. Ao final da seção, foram apresentadas todas as formações válidas (métodos e quantidade de grupos) para a amostra de dados utilizada.

3.3.1. Medidas de similaridade

A medida de similaridade é uma função usada para quantificar a semelhança entre dois objetos. A qualidade de uma boa análise de agrupamento está relacionada com a sua capacidade de dividir a amostra em grupos distintos, de forma a se obter alta homogeneidade interna e alta heterogeneidade entre os grupos (HAIR et al., 2009). Esse critério é obtido quando a análise consegue distinguir quais objetos são semelhantes e quais se diferem. Para isso, é preciso aplicar uma função que determine o seu grau de similaridade, ou seja, o quão similar são esses objetos. À medida matemática que exerce essa função na análise de agrupamento dá-se o nome de medida de similaridade.

Há diferentes medidas de similaridade e a opção por utilizar uma ou outra implica na formação de diferentes matrizes de similaridade (corresponde à matriz de distâncias entre os objetos). Pode, portanto, gerar diferentes resultados, mesmo quando aplicados à mesma matriz de dados e algoritmo de partição. A intenção de testar diferentes medidas é verificar se, de

forma geral, alguma delas proporciona resultados melhores que as demais ou se alguma delas é mais suscetível a algum tratamento de dados.

Nesse estudo, foram utilizadas cinco medidas de similaridade: (a) distância City-Block ou Manhattan, (b) distância Euclidiana, (c) distância Euclidiana Quadrada, (d) distância Chebyshev e (e) coeficiente de Pearson. Essas medidas foram selecionadas por serem as medidas mais comumente aplicadas a dados quantitativos.

A Equação 12 apresenta a expressão numérica utilizada para obter a distância entre dois objetos com a medida City-Block. Ela mede a semelhança entre dois objetos a partir da soma das diferenças entre todas as variáveis. Essa medida é baseada na distância de Minkowsky, que é uma generalização para várias medidas de distância. Segundo ela, a distância entre dois objetos é definida pela raiz de ordem "p", tirada a partir da soma das diferenças entre todas as variáveis de dois objetos, elevadas a "p". A medida City-Block é, portanto, a medida de Minkowsky quando se adota p igual a 1.

$$d_{ij} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{im} - x_{jm}| \quad (12)$$

Onde:
d_{ij} é a distância de i até j;
x_{im} são os valores do objeto x_i para cada variável;
x_{jm} são os valores do objeto x_j para cada variável.

A distância Euclidiana foi obtida por meio da Equação 13. Essa é a medida de similaridade mais comumente utilizada para valores numéricos (JAIN et al., 1999; WITTEN et al., 2005; HAN et al., 2011). O valor da distância é calculado a partir da raiz quadrada da soma entre as diferenças de todas as variáveis elevadas ao quadrado. Também é baseada na distância de Minkowsky, adotando-se p igual a dois.

$$d_{ij} = \sqrt[2]{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{im} - x_{jm})^2} \quad (13)$$

Onde:
d_{ij} é a distância de i até j;
x_{im} são os valores do objeto x_i para cada variável;
x_{jm} são os valores do objeto x_j para cada variável.

Uma variação muito utilizada da distância Euclidiana é a distância Euclidiana Quadrada. É definida, portanto, pela soma dos quadrados das diferenças entre cada variável, entre dois pares de objetos. Ela diferencia-se da distância Euclidiana por ressaltar as diferenças entre os objetos que estão mais distantes. A distância Euclidiana Quadrada foi calculada por meio da Equação 14.

$$d_{ij} = (x_{i1} - x_{j2})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{im} - x_{jm})^2 \quad (14)$$

Onde:
d_{ij} é a distância de i até j;
x_{im} são os valores do objeto x_i para cada variável;
x_{jm} são os valores do objeto x_j para cada variável.

A distância Chebyshev é outro caso especial da distância de Minkowsky. Obtém-se essa medida quando o valor de p tende ao infinito. Nessa situação, o valor da distância é, na verdade, o máximo valor absoluto da diferença entre qualquer dos atributos dos vetores (atributos correspondem aos valores assumidos por cada objeto para cada variável). A distância Chebyshev foi obtida por meio da Equação 15.

$$d_{ij} = \max_{f=1} |x_{if} - x_{jf}| \quad (15)$$

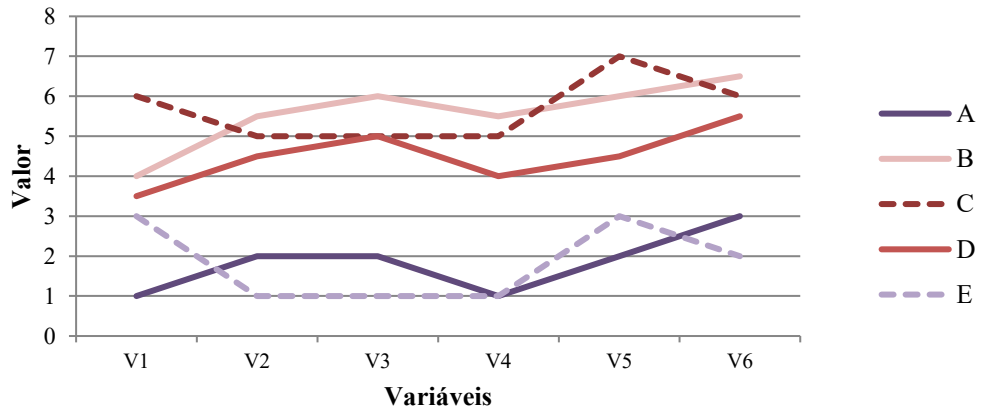
Onde:
d_{ij} é a distância de i até j;
x_{im} são os valores do objeto x_i para cada variável;
x_{jm} são os valores do objeto x_j para cada variável.

As medidas de similaridade apresentadas anteriormente são aplicadas preferencialmente a situações nas quais se deseja conhecer o grau de proximidade ou distância entre os objetos, dada uma certa amostra de dados. Entretanto, há situações em que um dado grupo de objetos é melhor subdividido a partir da existência de associação ou padrão entre os dados. Nesses casos, a medida que possui melhor habilidade para determinar a semelhança entre os objetos é um coeficiente de correlação.

A Figura 16 exemplifica a forma como medidas de distância e medidas de correlação consideram a semelhança entre objetos de uma amostra. Uma medida de proximidade segregaria a amostra pelos valores mais próximos dos seus elementos, mantendo no mesmo agrupamento os elementos de mesma cor (B, C e D em um grupo e A e E, em outro). Se a

medida fosse de correlação, seriam agrupados os elementos cujo comportamento assume um mesmo padrão (elementos C e E em um grupo e A, B e D em outro).

Figura 16 – Comparação entre medidas de proximidade e de correlação.



Em função dessa diferença, nessa tese foi adotada também uma medida de correlação. A Equação 16 apresenta a expressão numérica utilizada para obter o coeficiente de correlação de Pearson, um dos coeficientes de correlação utilizados para o propósito da análise de *cluster*.

$$\rho = \frac{cov\,(X_1, X_2)}{\sqrt{\sigma_1 + \sigma_2}} \tag{16}$$

Onde:
ρ é o coeficiente de correlação obtido;
σ_i é a variância do objeto i;
σ_j é a variância do objeto j.

3.3.2. Algoritmos de partição

Os algoritmos de partição representam um conjunto de regras de como os grupos devem ser formados a partir de uma matriz de dados, ou seja, como separar os objetos da amostra em grupos. No processo de clusterização, os grupos vão sendo unidos a partir da similaridade entre eles. A decisão quanto a quais grupos devem ser unidos corresponde sempre àqueles que possuem menor distância. A função do algoritmo de partição é determinar de que forma a medida de similaridade será aplicada a dois grupos, ou seja, a partir de quais objetos de cada grupo ela será medida.

Após a seleção das medidas de similaridade, o próximo passo foi o processo de partição da amostra em grupos. Foram adotados nesta tese algoritmos de técnicas hierárquicas e não hierárquicas e suas características estão apresentadas nas seções a seguir.

3.3.2.1. Técnica hierárquica de agrupamento

As técnicas hierárquicas caracterizam-se pela formação de grupos através de uma sucessão de etapas, formando uma hierarquia de partições. Cada objeto inicia o processo como um agrupamento unitário. A cada nova etapa, o algoritmo de partição escolhido especifica de que forma a medida de similaridade será aplicada, resultando em um novo par de objetos ou grupos que deverão ser unidos. Esse processo se repete até que todos os objetos da amostra em análise sejam unidos em um único agrupamento. Essa forma de construir relações é chamada de formação em árvore e permite a construção de um gráfico chamado dendograma. Nesse tipo de gráfico, é possível fazer uma leitura visual da forma como os objetos foram se agrupando. O gráfico também indica o nível de similaridade obtido a cada nova união. O nível de similaridade indica a distância que possuíam os grupos no momento em que foram unidos.

Quando os agrupamentos são unitários, a medida de similaridade é calculada para cada par de objetos. Mas como fazer para obter a medida de similaridade entre agrupamentos formados por mais de um objeto? Essa resposta é dada pelo algoritmo de partição. Cada algoritmo apresenta uma forma diferente de aplicar essa medida.

Os algoritmos de partição baseados em técnicas hierárquicas que foram objeto de investigação nesta tese são o algoritmo da Ligação Simples, algoritmo da Ligação Completa, algoritmo do Centroide e algoritmo de Ward. Suas características e as regras de partição específicas de cada um estão apresentadas a seguir.

a) Ligação Simples (ou vizinho mais próximo)

Esse é um algoritmo simples e detecta grupos de forma muito variada. Neste método, a similaridade ou distância entre dois *clusters* é definida a partir dos objetos mais similares ou próximos entre esses *clusters*, como mostra a Figura 17. A medida entre apenas dois objetos basta para determinar a distância entre os grupos. Essa medida dá ênfase a regiões mais próximas, desconsiderando a estrutura geral do *cluster*. Em função disso, é capaz de detectar formas irregulares ou não-elípticas, mas é sensível a ruídos e objetos atípicos (JAIN et al., 1999;

WITTEN et al., 2005; HAN et al., 2011). A Figura 18 apresenta os procedimentos que devem ser adotados quando se realiza a partição dos objetos pelo algoritmo de Ligação Simples.

Figura 17 - Representação do algoritmo de partição: Ligação Simples.

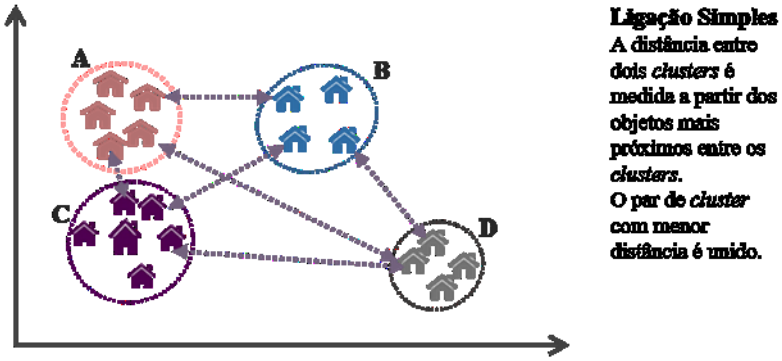
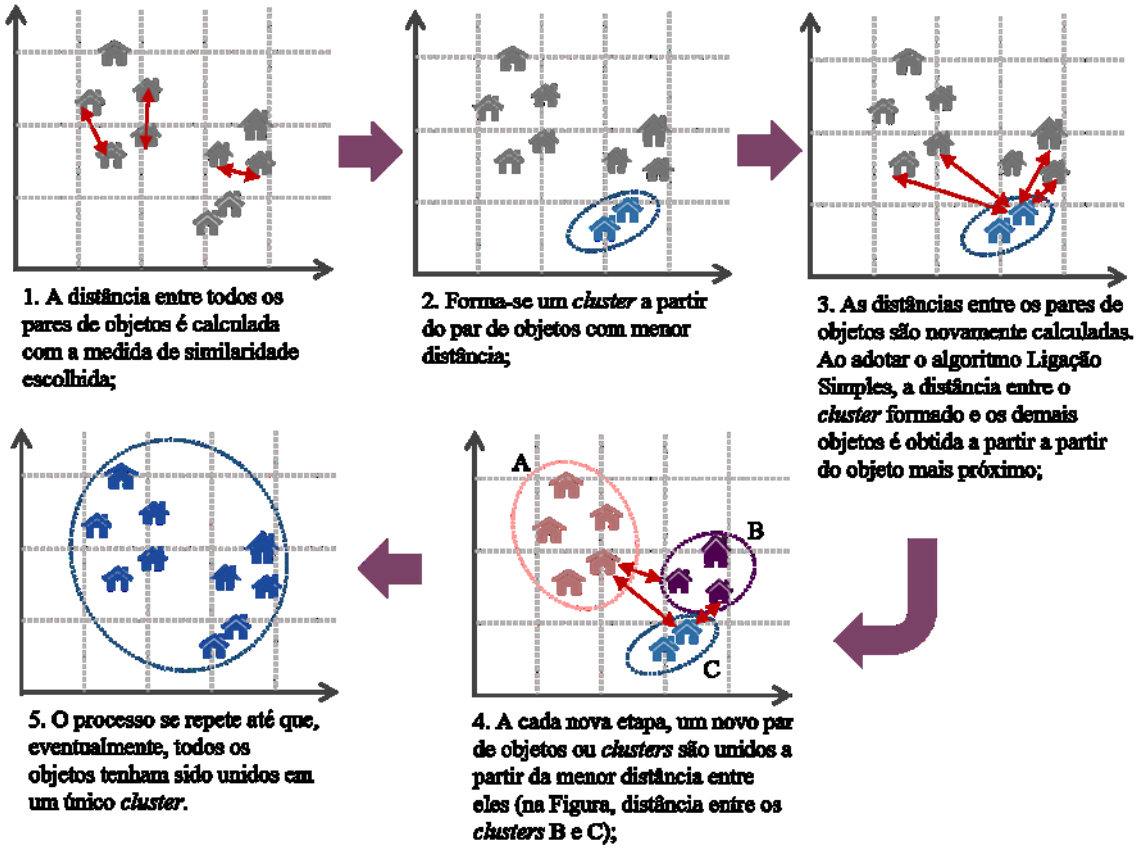


Figura 18 - Procedimentos de partição: Ligação Simples.

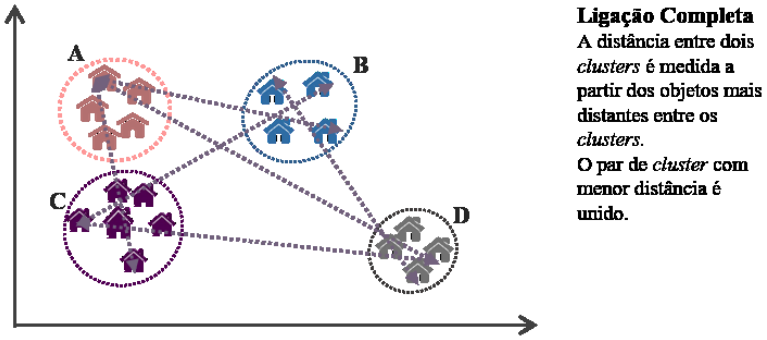


b) Ligação Completa (ou vizinho mais distante)

O procedimento adotado com esse algoritmo é muito similar ao da Ligação Simples, com a diferença de que a similaridade ou distância entre dois *clusters* é definida a partir dos objetos mais distantes entre esses *clusters*, representado na Figura 19. Esse algoritmo busca

formar *clusters* com formas bem compactas, com menor diâmetro possível, por isso também é conhecido como Método do Diâmetro. É um pouco menos sensível a objetos atípicos e tende a formar grupos com formato mais oval. Uma desvantagem é que tende a quebrar grupos muito grandes (JAIN et al., 1999; WITTEN et al., 2005; HAN et al., 2011).

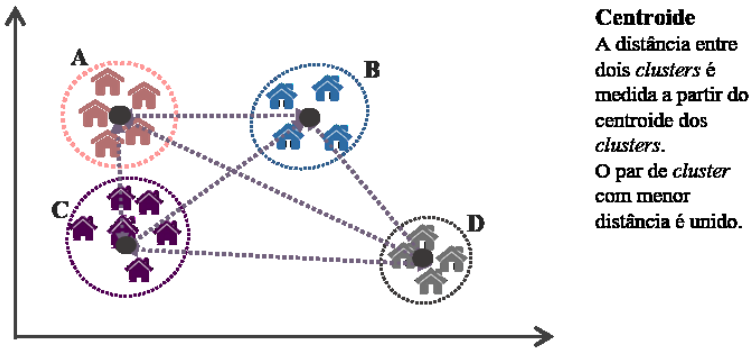
Figura 19 - Representação do algoritmo de partição: Ligação Completa.



c) Centroide

O algoritmo do Centroide computa a distância entre o centro de cada *cluster* (chamado centroide), definido pela média entre seus objetos considerando todas as variáveis (Figura 20). O procedimento para o cálculo da medida de similaridade é similar aos métodos anteriores, tomando por base a distância entre os centroides (JAIN et al., 1999; WITTEN et al., 2005; HAN et al., 2011).

Figura 20 - Representação do algoritmo de partição: Centroide

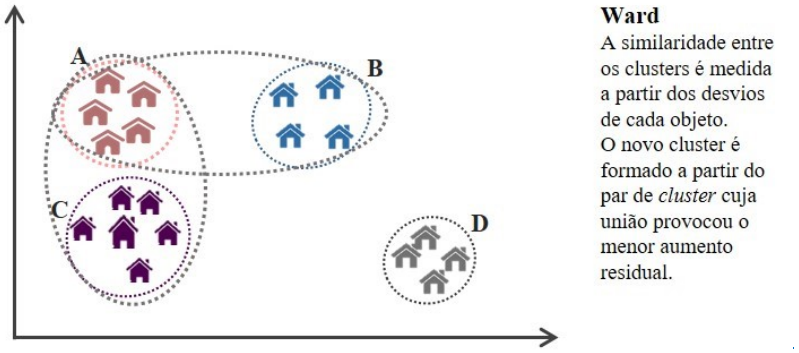


d) Ward

No algoritmo de Ward a medida de similaridade usada para juntar os agrupamentos é calculada a partir da união de dois agrupamentos cuja combinação minimiza o aumento residual

dos quadrados ao longo de todas as variáveis. Uma característica desse algoritmo é formar agrupamentos com tamanho aproximado, devido à minimização da variação interna. A Figura 21 representa o método, onde os desvios aos centroides de cada grupo são computados (JAIN et al., 1999; WITTEN et al., 2005; HAN et al., 2011).

Figura 21 - Representação do algoritmo de partição: Ward.



3.3.2.2. Técnica não hierárquica de agrupamento

Nas técnicas não hierárquicas, cada objeto da amostra é designado a um agrupamento entre os definidos previamente, em uma única etapa. Diferentemente das técnicas hierárquicas, que definem os agrupamentos através de um processo em árvore (ou seja, uma vez designado um objeto a um agrupamento, ele não poderá ser redesignado a outro agrupamento), as técnicas não hierárquicas são interativas. Isso significa que os objetos podem ser redesignados, ou seja, um objeto poderá mudar de agrupamento caso no final do processo ele venha a se assemelhar mais com os objetos de outro agrupamento. Esse processo se repete até que a convergência seja obtida, ou seja, nenhum objeto mude mais de agrupamento. O algoritmo de partição da técnica não hierárquica que foi objeto de investigação nesta tese é o algoritmo K-means.

e) Algoritmo K-means

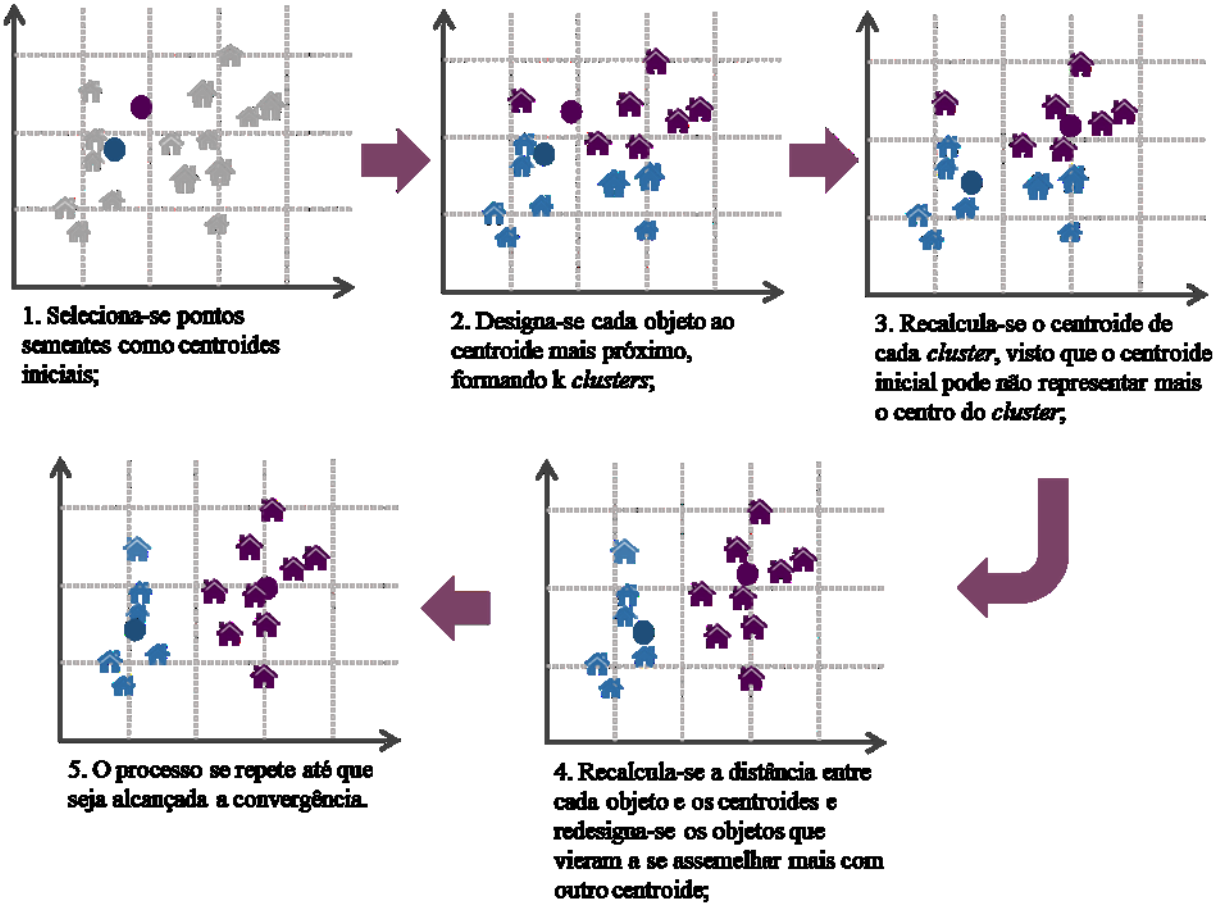
Este é o mais popular dos algoritmos de partição não hierárquicos. Nesse método, cada *cluster* é representado pelo seu centro (centroide), que é definido pela média multivariada de todos os objetos do grupo (JAIN et al., 1999; HAIR et al., 2009; HAN et al., 2011). O algoritmo K-means separa os grupos tentando minimizar a função objetivo (Equação 17).

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} d(x_j, \bar{x}_i)^2 \qquad (17)$$

Onde:
J é a função objetivo;
xj é o valor da variável para cada objeto;
$\bar{x}$ é a média dos valores de determinada variável;
d é a distância do objeto ao centroide.

A Figura 22 apresenta os procedimentos que devem ser adotados quando se realiza a partição dos objetos pelo algoritmo K-means.

Figura 22 - Procedimentos de partição: K-means.



É um método muito eficiente em relação aos demais, visto que a quantidade de interações é normalmente muito menor que a de objetos. Um ponto negativo é a necessidade de informar previamente a quantidade de agrupamentos a serem formados. Para isso, o pesquisador já deve ter uma expectativa em relação a grupos existentes, ou deve então realizar algumas

análises e comparar os resultados para verificar qual deles apresenta a solução mais coerente. Também é sensível a sua inicialização, ou seja, à escolha dos pontos sementes. Diferentes inicializações podem resultar em diferentes formações de agrupamento, por isso a correta escolha dos pontos sementes é muito importante (BUSSAB et al., 1999, HAIR et al., 2009).

Por trabalhar com valores de médias, é também suscetível a ruídos e objetos atípicos (opções para contornar esses problemas seriam os algoritmos k-medians e k-medoids, que não foram trabalhados nesse estudo). Pelo mesmo motivo, não funciona para variáveis categóricas, sendo necessário transformá-la em dados contínuos.

3.3.3. Conjunto preliminar de soluções

O conjunto preliminar corresponde a todas as formações possíveis a partir dos parâmetros apresentados anteriormente, apresentadas no Quadro 3. Os métodos apresentados foram formados a partir da combinação entre as matrizes de dados, medidas de similaridade e algoritmos de partição selecionados nas seções anteriores. Foram formados 105 métodos de agrupamento distintos. No Quadro 3, a combinação entre os algoritmos de partição e medidas de similaridade estão apresentados nas linhas. Nas colunas à direita, estão as matrizes de dados. Na intersecção entre eles, estão apresentados os métodos a partir da nomenclatura utilizada a cada um na sequência desse estudo.

Quadro 3 – Conjunto preliminar de métodos de clusterização.

Métodos		**Matriz de dados**				
Algoritmo de partição	Medida de similaridade	A	B	C	D	E
Ligação simples	City-Block	M01_A	M01_B	M01_C	M01_D	M01_E
	Euclidiana	M02_A	M02_B	M02_C	M02_D	M02_E
	Euclidiana quadrada	M03_A	M03_B	M03_C	M03_D	M03_E
	Chebyshev	M04_A	M04_B	M04_C	M04_D	M04_E
	Pearson	M05_A	M05_B	M05_C	M05_D	M05_E
Ligação completa	City-Block	M06_A	M06_B	M06_C	M06_D	M06_E
	Euclidiana	M07_A	M07_B	M07_C	M07_D	M07_E
	Euclidiana quadrada	M08_A	M08_B	M08_C	M08_D	M08_E
	Chebyshev	M09_A	M09_B	M09_C	M09_D	M09_E
	Pearson	M10_A	M10_B	M10_C	M10_D	M10_E

Quadro 3 – Conjunto preliminar de métodos de clusterização (continuação).

Métodos		**Matriz de dados**				
Algoritmo de partição	Medida de similaridade	A	B	C	D	E
Centroide	City-Block	M11_A	M11_B	M11_C	M11_D	M11_E
	Euclidiana	M12_A	M12_B	M12_C	M12_D	M12_E
	Euclidiana quadrada	M13_A	M13_B	M13_C	M13_D	M13_E
	Chebyshev	M14_A	M14_B	M14_C	M14_D	M14_E
	Pearson	M15_A	M15_B	M15_C	M15_D	M15_E
Ward	City-Block	M16_A	M16_B	M16_C	M16_D	M16_E
	Euclidiana	M17_A	M17_B	M17_C	M17_D	M17_E
	Euclidiana quadrada	M18_A	M18_B	M18_C	M18_D	M18_E
	Chebyshev	M19_A	M19_B	M19_C	M19_D	M19_E
	Pearson	M20_A	M20_B	M20_C	M20_D	M20_E
K-means	Otimização	M21_A	M21_B	M21_C	M21_D	M21_E

3.3.4. Conjunto de soluções final

O conjunto de soluções final corresponde à última etapa do processo de análise de agrupamento, quando foi determinada a quantidade “k” final de grupos a serem formados a partir de cada método. Devido ao fato de que a formação dos agrupamentos pelos métodos hierárquicos e não hierárquicos são diferentes, os procedimentos adotados para a determinação da quantidade de agrupamentos para cada um dos métodos foram também diferentes.

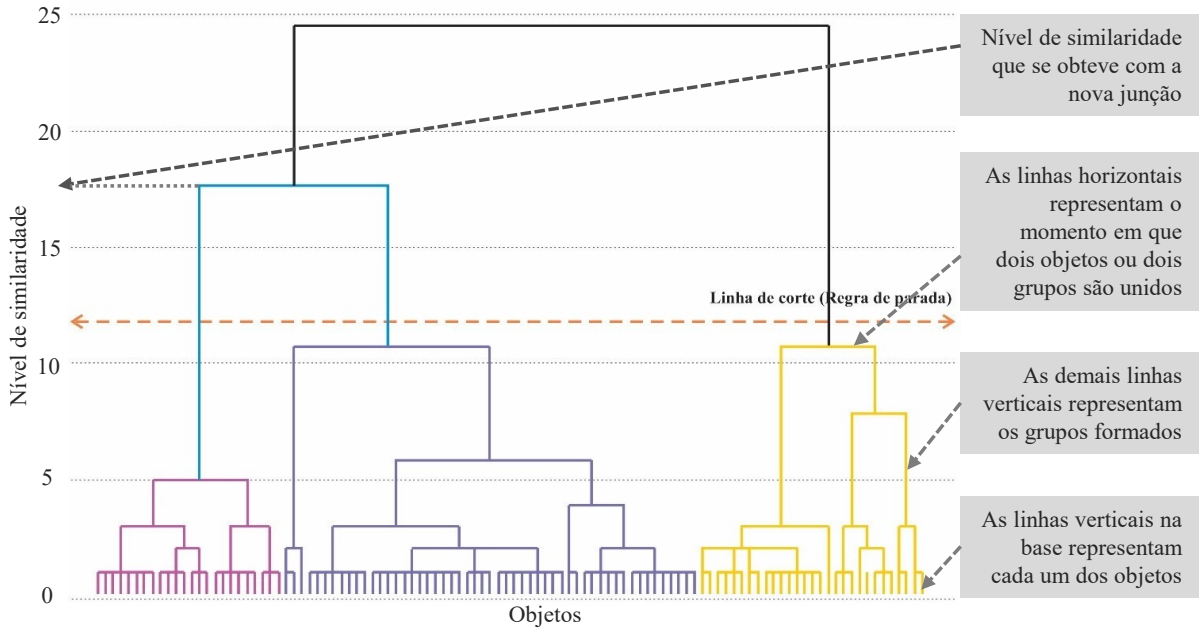
3.3.4.1. Técnica hierárquica de agrupamento

Para os métodos definidos a partir de técnicas hierárquicas, a definição da quantidade de agrupamentos se deu através do dendograma e dos coeficientes de aglomeração obtidos no processo de partição. No dendograma, é representado graficamente como se deu a formação dos agrupamentos ao longo de todo processo, a cada etapa. Como mencionado anteriormente, são registrados ali o nível de similaridade entre os objetos unidos em cada etapa. Uma análise visual permite verificar em que etapa a união de dois agrupamentos provocou um aumento de similaridade relativamente maior que as etapas anteriores. Esse aumento relativamente maior indica a etapa ideal onde deve ser feito o corte. A quantidade agrupamentos formados

corresponde à quantidade de agrupamentos existente na etapa anterior em que o corte foi realizado.

A Figura 23 apresenta um exemplo de dendograma. Na base, no eixo das abscissas, estão representados por pequenas linhas verticais todos os objetos envolvidos na análise. A linha horizontal que liga os objetos representa a etapa em que foram unidos em um mesmo agrupamento. Essa linha corresponde também ao nível de similaridade obtido com a nova união. O nível de similaridade está representado no eixo das ordenadas e representa a distância que possuíam os agrupamentos no momento em que foram unidos. Com o dendograma, é possível verificar a forma como foram sendo construídos os agrupamentos, etapa por etapa (cada nova união é considerada uma etapa), até que todos os objetos sejam unidos em um único grupo. Observa-se na Figura 23 que após a união de um grupo correspondente ao nível de similaridade próximo a 10, a próxima união ocorreu a um nível de similaridade próximo a 18, uma diferença relativamente maior que nas etapas anteriores (quase o dobro). Nesse momento, aplica-se a regra de parada e a solução final corresponde à quantidade de grupos formada imediatamente abaixo dessa linha (no caso da Figura 23, corresponde a três grupos). Neste estudo, por tratar-se de uma amostra pequena, com apenas 102 casos e características bem específicas, foi adotado um intervalo de soluções com limite até 5 grupos, de forma a não obter como resultado agrupamentos muito pequenos.

Figura 23 - Exemplo de dendograma obtido a partir da técnica hierárquica.



Fonte: Adaptado de Rosa (2014)

No exemplo apresentado com a Figura 23, pode-se ver claramente a separação da amostra em três grupos. Entretanto, nem sempre é possível, através do dendograma, definir uma única solução quanto à quantidade de agrupamentos. Há situações em que o dendograma sugere mais de uma solução. Quando for o caso, o coeficiente de aglomeração pode ser utilizado.

O coeficiente de aglomeração indica o grau de heterogeneidade obtido a cada união. Esse valor aumenta a cada nova etapa, visto que à medida que mais objetos vão sendo agregados ao agrupamento existente, maior fica a heterogeneidade interna desse agrupamento, ou seja, as características encontradas nos objetos do grupo são cada vez mais divergentes. Conforme indicado por Rosa (2014), o aumento da heterogeneidade é melhor quantificado pelo cálculo das variações percentuais de heterogeneidade, valor obtido pela razão da diferença entre os coeficientes de aglomeração da etapa atual e anterior pelo coeficiente de aglomeração da etapa anterior (Equação 18). O melhor ponto para fazer o corte (regra de parada) é na etapa onde o maior percentual de heterogeneidade foi obtido. A quantidade ideal de agrupamentos é aquela da etapa anterior a este aumento. O coeficiente de aglomeração foi obtido através do programa SPSS.

$$P_{het} = \frac{(CA_{(n)} - CA_{(n-1)})}{CA_{(n-1)}} \tag{18}$$

Onde:
P_{het} é a variação percentual de heterogeneidade obtida para cada etapa;
$CA_{(n)}$ é o coeficiente de aglomeração da etapa atual;
$CA_{(n-1)}$ é o coeficiente de aglomeração da etapa imediatamente anterior.

A partir desses dois procedimentos (análise do dendograma e dos coeficientes de aglomeração), definiu-se a solução final quanto à quantidade de grupos a serem formados por cada um dos métodos hierárquicos de agrupamento.

3.3.4.2. Técnica não hierárquica de agrupamento

Ao contrário da técnica hierárquica, na técnica não hierárquica é necessário definir inicialmente a quantidade de grupos a serem formados. Dessa forma, foi adotada como limite a quantidade de grupos correspondente à quantidade máxima encontrada nas soluções das técnicas hierárquicas. O método de clusterização foi aplicado paralelamente a todas as possíveis soluções (por exemplo, se com as técnicas hierárquicas obteve-se soluções formando até 5

grupos, com a técnica não hierárquica obteve-se soluções para k igual 2, k igual a 3, k igual a 4 e k igual a 5).

Devido ao fato de sua análise ser interativa, no método não hierárquico não é possível obter um gráfico como o dendograma ou o programa de aglomeração para definição da melhor solução quanto à quantidade ideal de agrupamentos. Uma forma de obter essa informação é através do histórico de interação.

O histórico de interação representa as mudanças sofridas pelos centroides a cada nova interação. Na primeira etapa, o valor do centroide de cada grupo é definido pelos pontos sementes. Após os objetos serem designados ao grupo com centroide mais próximo, a média multivariada do grupo é recalculada e se obtém um novo valor para o centroide. O processo é repetido até que não haja mais alteração no centroide. Diz-se então que a convergência é alcançada.

O valor da diferença dos centroides no histórico de interação vai diminuindo à medida que os objetos vão sendo redesignados e os novos centros vão sendo recomputados. Quanto antes a convergência for alcançada, melhor é considerada a divisão dos grupos. A forma como o centro dos *clusters* muda está relacionada à estabilidade do agrupamento. Em um processo de clusterização que demora a convergir há uma chance grande de que um ou mais grupos apresentem comportamento instável e que, portanto, provavelmente não representem boa divisão dos objetos.

Foram consideradas válidas a quantidade de agrupamentos em que todos os grupos formados convergiram antes da décima interação. Os métodos que apresentaram soluções que não convergiram, foram descartados.

Para complementar a análise, foram realizadas análises de variância para todas as variáveis, obtidas a cada aumento da quantidade de agrupamentos, para cada método. A análise de variância é uma técnica estatística que tem por objetivo avaliar se há diferença significativa entre médias. Além disso, verifica se há fatores que exercem maior impacto na separação dos grupos e o tamanho desse impacto. Detalhes sobre o procedimento para obtenção do valor-F foram apresentados na seção 3.2.3.

Ao aplicar a análise estatística ANOVA, obteve-se o valor de F para cada solução de agrupamento. Para cada método, definiu-se como solução quanto à quantidade de agrupamentos aquela cujos valores de F foram os maiores para a maioria das variáveis com significância estatística. Na impossibilidade de definir apenas uma solução, adotou-se todas as que se apresentaram equivalentes.

A partir desses dois procedimentos (análise do histórico de interação e análise de variância), definiu-se a solução final quanto à quantidade de grupos a serem formados para cada um dos métodos não hierárquicos de agrupamento.

3.4. Determinação dos modelos de referência

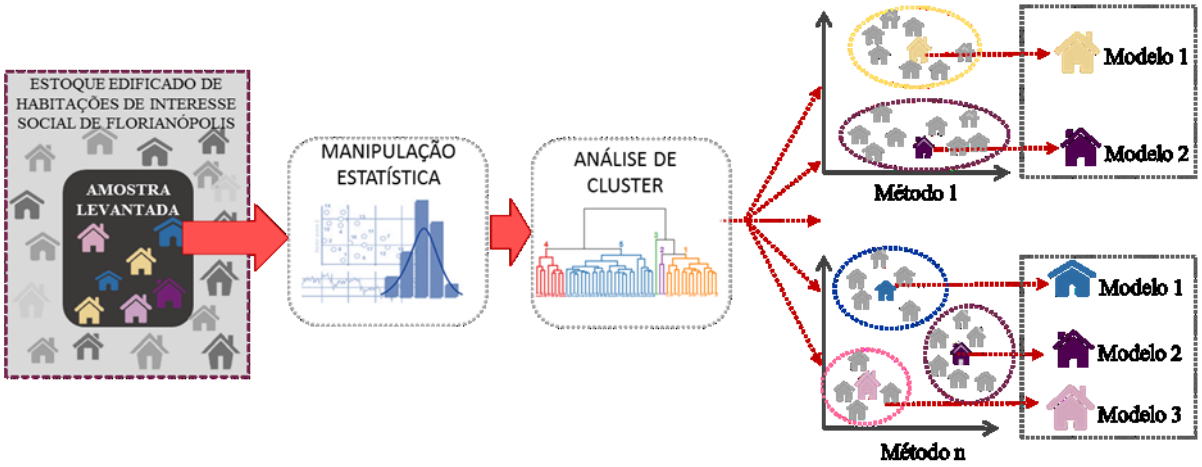
Foram determinados os modelos de referência para cada grupo dos métodos resultantes de técnicas hierárquicas e não hierárquicas.

Os modelos de referência, como apresentado na revisão de literatura, possuem diversas aplicações, como a classificação quanto ao desempenho energético ou para acessar o balanço térmico de edificações. Nesta tese, o modelo de referência foi utilizado com o objetivo de representar o estoque de edificações existente, ou seja, espera-se que ele possua as características de maior representatividade do grupo de habitações reais. Não se pretende, nesse momento, fazer uma análise qualitativa, classificando os modelos como bons exemplos de arquitetura ou inadequados, com desempenho bom ou ruim, nem tão pouco que ele seja um modelo a ser replicado. Apenas espera-se que ele apresente comportamento e características similares às demais habitações do grupo, de forma a permitir que se identifiquem as estratégias mais adequadas a estas habitações. Por tratar-se de uma amostra de habitações de interesse social, resultantes de auto construção e, portanto, sem orientação de um profissional qualificado, é esperado que o modelo apresente alguma inadequação funcional. Entretanto, se essa inadequação é uma característica comum da amostra em análise, ela não deve ser ignorada. Pelo contrário, deve ser mantida para que as análises envolvendo o modelo indiquem as melhores soluções para as problemáticas existentes.

A definição do modelo de referência baseou-se no conceito de edifício real (LOGA et al., 2008), definido como uma edificação existente, cujas características mais se assemelham ao centroide do grupo (média multivariada do agrupamento). Optou-se pela adoção do conceito de edifício real em vez do edifício teórico pois este, por tratar-se de uma edificação fictícia, implicaria na necessidade de adotar diversas suposições para criar o modelo (como a determinação das características que não compuseram o banco de dados). Por sua vez, o modelo real apresenta maior praticidade de aplicação, visto que o modelo corresponde a uma geometria existente, com todas as características já definidas. Dessa forma, para cada grupo formado, o objeto mais próximo ao centroide foi selecionado como modelo de referência daquele grupo. A proximidade de cada objeto ao centroide foi determinada tomando por base a medida de

distância encontrada na etapa de partição. A Figura 24 exemplifica o processo de determinação do modelo de referência.

Figura 24 – Fluxograma do processo de determinação do modelo de referência.



3.5. Validação dos métodos

A validação dos métodos consiste em avaliar a partir de outras técnicas estatísticas se os resultados obtidos com a análise de *cluster* são consistentes e coerentes. Por tratar-se de uma técnica exploratória e não inferencial, aplicar medidas de validação representa uma etapa importante e auxilia no processo de decisão quanto ao melhor método.

A validação dos métodos foi aplicada para verificar se os modelos de referência obtidos para os agrupamentos encontrados podem representar os demais objetos de seu agrupamento em estudos de desempenho térmico. Isso quer dizer que ao se variar alguns parâmetros no modelo, comportamento similar quanto ao seu desempenho poderá ser encontrado nos demais objetos do grupo. Se isso acontecer, estudos futuros poderão ser simplificados e realizados apenas com os modelos, e os resultados obtidos estendidos para toda a amostra.

Foram utilizadas medidas internas e relativas para validação dos resultados da análise de *cluster*. Como medida interna, foi utilizado o cálculo da inércia, que estabelece uma relação entre a inércia *intra-cluster* e a inércia *inter-cluster*. Essa medida foi calculada para todos os métodos selecionados como válidos na seção 3.3.4 e obtida a partir da matriz de dados inicial. Como medida relativa, foram aplicados seis índices estatísticos que quantificam o erro amostral ao comparar cada objeto com a média e modelo de referência do seu grupo. Para o cálculo dessa

medida foram utilizados os dados de desempenho térmico obtidos a partir das simulações computacionais (matriz de desempenho).

Nenhuma medida externa foi aplicada visto que não há classes pré-definidas para classificação das edificações utilizadas nesse estudo.

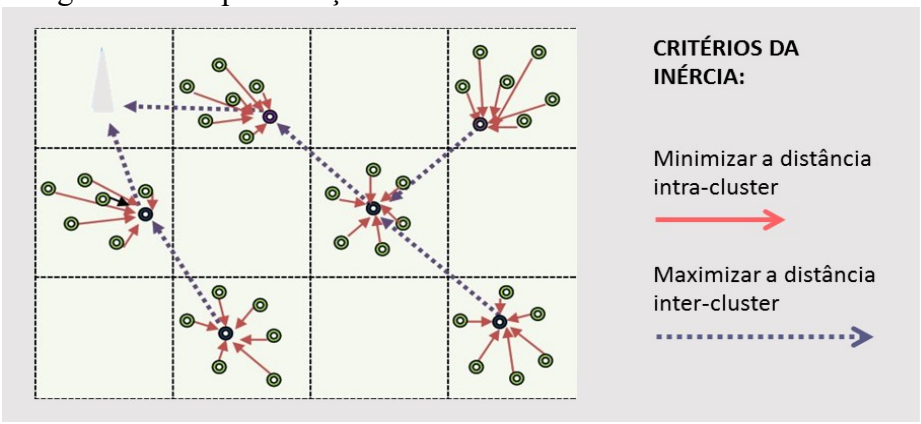
3.5.1. Validação: medida interna

A medida interna utilizada para validação da qualidade dos agrupamentos formados por cada método foi a Inércia. Trata-se de uma função de semelhança de cada elemento obtida a partir dos desvios ao centroide dos agrupamentos e da nuvem de dados. Também pode ser vista como uma avaliação da dispersão dos elementos dentro de um *cluster*.

Como já mencionado, as avaliações internas não necessitam o conhecimento de um particionamento a priori, e utilizam, portanto, os dados envolvidos na própria análise de *cluster*. Para esse tipo de avaliação, utilizam-se medidas de semelhança entre *clusters* (*intra-cluster*) e dentro de *clusters (inter-cluster)*.

Há diferentes formas de dividir uma amostra em k *clusters*, o que é obtido a partir da aplicação de vários métodos. Sabe-se que uma boa separação se baseia na capacidade de separar a amostra em grupos compactos e distantes entre si. Assim, a melhor solução é aquela que minimiza a medida *intra-cluster* e maximiza a medida *inter-cluster* (Figura 25). A *medida intra-cluster* representa os desvios dos objetos ao centro do grupo e a medida *inter-cluster,* os desvios dos centroides de cada grupo ao centro da nuvem. Quanto menor a inércia *intra-cluster*, mais homogêneos são os grupos formados. Quanto maior a inércia *inter-cluster*, maior é a separação dos grupos. A inércia total permanece constante e independe da forma como os objetos são agrupados.

Figura 25 – Representação das medidas *intra-cluster* e *inter-cluster*.



O cálculo da Inércia é realizado a partir de uma sequência de procedimentos. O objetivo final é encontrar a relação entre inércia *inter-cluster* e inércia *intra-cluster* para cada método.

O primeiro passo é encontrar o centro de gravidade da nuvem de pontos (G). As coordenadas do centro de gravidade representam a média dos seus objetos, descritos através da média de seus atributos (Equação 19).

$$G = \frac{\sum xij}{n} \tag{19}$$

Onde:
G é o centro de gravidade da nuvem;
x_{ij} é cada objeto da amostra, representado por seus atributos;
n é a quantidade de objetos da amostra.

Após a obtenção do centro de gravidade da nuvem, obtém-se a Inércia Global (I_G). A Inércia Global representa a medida de dispersão da nuvem de pontos ao redor do centro de gravidade da nuvem e é obtida a partir da soma dos quadrados das distâncias entre cada objeto e o centro de gravidade G (Equação 20).

$$I_G = \sum_{i=1}^{n} ||x_{ij} - G||^2 \tag{20}$$

Onde:
I_G é Inércia Global;
x_{ij} é cada objeto da amostra, representado por seus atributos;
G é o centro de gravidade da nuvem.

Paralelamente à inércia global (I_G), calcula-se a inércia total (I_{tot}). A inércia total representa a soma da inércia *intra-cluster* (I_{intra}) e inércia *inter-cluster* (I_{inter}) e deve resultar no mesmo valor obtido pela inércia global. O valor da inércia global diferir da inércia total é um indicativo de que algo não foi calculado adequadamente.

Para obter a inércia *intra-cluster* e inércia *inter-cluster*, é preciso conhecer o centro de gravidade de cada agrupamento (C_i). O centro de gravidade de um cluster é a média multivariada de seus objetos (em outras palavras, corresponde ao centroide do grupo). O Centro de gravidade foi calculado por meio da Equação 21.

$$C_i = \frac{\sum xij}{n_i} \quad (21)$$

Onde:
C_i é o centro de gravidade do *cluster*;
x_{ij} é cada objeto do *cluster*, representado por seus atributos;
n_i é a quantidade de objetos do grupo.

A inércia *inter-cluster* (I_{inter}) corresponde à medida de separação dos *clusters* e é obtida a partir da soma das contribuições individuais de cada *cluster* ($I_{inter_{Ci}}$), conforme apresentado na Equação 22. A contribuição de cada *cluster* é obtida a partir da soma dos quadrados das distâncias dos centroides dos *clusters* (C_i) ao centro de gravidade da nuvem de pontos G, ponderadas pelos respectivos tamanhos dos *clusters* (Equação 23).

$$I_{inter} = \sum_{i-1}^{n} I_{inter_{Ci}} \quad (22)$$

$$I_{inter_{Ci}} = \sum_{i=1}^{n} n_i * ||C_i - G||^2 \quad (23)$$

Onde:
I_{inter} é a inércia *inter-cluster*;
$I_{inter_{Ci}}$ é a contribuição de cada *cluster* à inércia *inter-cluster*;
n_i é a quantidade de objetos do grupo;
C_i é o centro de gravidade do *cluster*;
G é o centro de gravidade da nuvem.

Por sua vez, a inércia *intra-cluster* (I_{intra}) representa a medida de heterogeneidade interna dos *clusters* e é obtida a partir da soma das contribuições individuais de cada *cluster* ($I_{intra_{Ci}}$), calculado por meio da Equação 24. A contribuição de cada *cluster* é obtida com a soma dos quadrados das distâncias dos objetos aos centroides dos *clusters* (C_i) onde estão alocados (Equação 25).

$$I_{intra} = \sum_{i-1}^{n} I_{intra_{Ci}} \quad (24)$$

$$I_{intra_{Ci}} = \sum_{i=1}^{n} ||x_{ij} - C_i||^2 \quad (25)$$

Onde:
I_{intra} é a inércia *inter-cluster*;
$I_{intra_{Ci}}$ é a contribuição de cada *cluster* à inércia *intra-cluster*;
x_{ij} é cada objeto do *cluster*, representado por seus atributos;
C_i é o centro de gravidade do *cluster*.

A inércia total é obtida por meio da Equação 26.

$$I_{tot} = I_{inter} + I_{intra} \quad (26)$$

Onde:
I_{itot} é a inércia total;
I_{inter} é a inércia *inter-cluster*;
I_{intra} é a inércia *intra-cluster*.

O valor da inércia *inter-cluster* e da inércia *intra-cluster* varia com a quantidade de *clusters* (k), porém o valor da inércia total é constante para uma mesma matriz de dados e independe do valor de k. A inércia *intra-cluster* tende a diminuir ao aumentar a quantidade de *clusters*. Inversamente, a inércia *inter-cluster* tende a aumentar à medida que se aumenta a quantidade de *clusters*. A medida de qualidade de uma partição é dada pela maximização da relação entre inércia *inter-cluster* e inércia *intra-cluster*, calculada por meio da Equação 27. Entretanto, em um processo de clusterização, há interesse em reduzir a quantidade de grupos ao menor número possível. Dessa forma, foram consideradas soluções válidas a partir dos métodos apresentados aquelas definidas pela partição da amostra na menor quantidade de *clusters* que satisfaça a desigualdade determinada por meio da Equação 28.

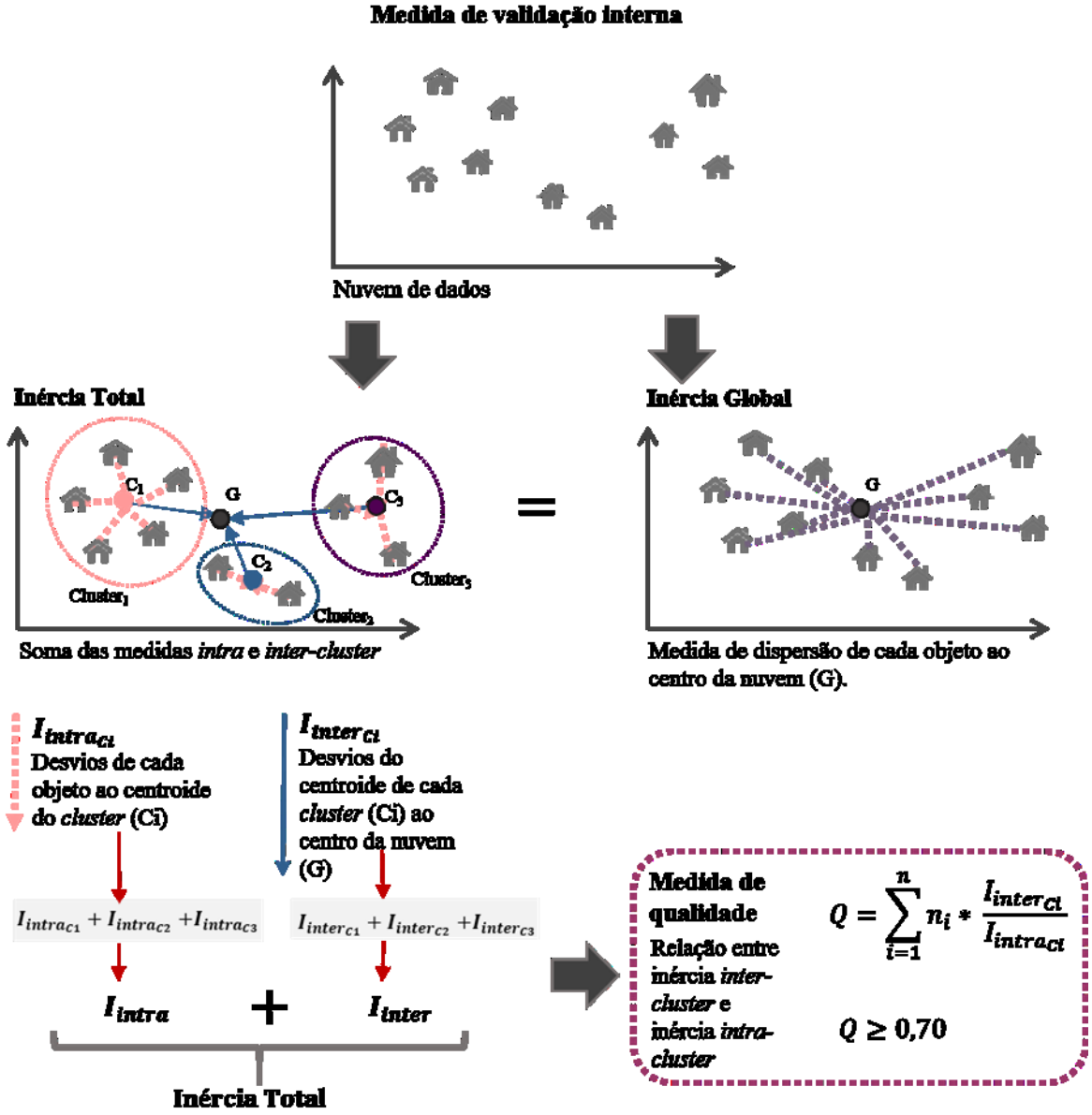
$$Q = \sum_{i=1}^{n} n_i * \frac{I_{inter_{Ci}}}{I_{intra_{Ci}}} \quad (27)$$

$$Q \geq 0{,}70 \quad (28)$$

Onde:
Q é a medida de qualidade da clusterização;
n_i é a quantidade de objetos do grupo;
$I_{inter_{Ci}}$ é a contribuição de cada *cluster* à inércia *inter-cluster*;
$I_{intra_{Ci}}$ é a contribuição de cada *cluster* à inércia *intra-cluster*.

A Figura 26 apresenta um resumo dos procedimentos para o cálculo da validação interna.

Figura 26 – Procedimentos para validação interna.



3.5.2. Validação: medida relativa

A medida de validação relativa tem por objetivo analisar a qualidade do agrupamento a partir de variáveis que descrevam os objetos, mas que não tenham sido envolvidas na análise propriamente. Essas variáveis devem estar relacionadas com o objetivo de pesquisa e permitir a criação um rótulo para os grupos formados. Em outras palavras, as variáveis escolhidas devem ser capazes de descrever consistentemente os grupos formados através de critérios que deem respaldo ao problema de pesquisa.

O objetivo da aplicação da análise de *cluster* nessa tese é a busca por modelos de referência capazes de representar o desempenho térmico do seu grupo em simulações computacionais. As variáveis selecionadas para validação através de medida relativa devem então, naturalmente, representar o desempenho térmico das habitações. Com esse objetivo, a medida de validação foi aplicada à matriz de desempenho obtida na seção 3.2.

Uma das principais proposições do modelo de referência é investigar a adequação de medidas de eficiência em edificações do mesmo tipo. Nesse sentido, o que se espera do modelo é que as variações de desempenho encontradas ao alterar as suas configurações sejam próximas às variações encontradas para os demais objetos do grupo, quando submetidos às mesmas configurações. Dessa forma, utilizou-se as variações percentuais obtidas entre o caso base e as demais configurações de simulação, calculadas por meio da Equação 29.

$$\Delta_{Id} = \frac{|Id_A - Id_i|}{Id_A} \tag{29}$$

Onde:
Δ_{Id} é a variação percentual do indicador de desempenho (aquecimento ou resfriamento) entre duas configurações;
Id_A é o indicador de desempenho da configuração base;
Id_i é o indicador de desempenho das demais configurações.

Desse processo, foram criadas oito variáveis para representar o desempenho térmico das habitações, apresentadas no Quadro 4.

Quadro 4 - Variáveis submetidas à medida de validação relativa.

Nome da variável	**Indicador de desempenho**	**Ambiente**	**Variação na configuração**
Δ_{resf}**_dorm_B**	Resfriamento	Dormitórios	Sistemas construtivos (B)
Δ_{aquec}**_dorm_B**	Aquecimento	Dormitórios	Sistemas construtivos (B)
Δ_{resf} **_dorm_C**	Resfriamento	Dormitórios	Ocupação (C)
Δ_{aquec} **_dorm_C**	Aquecimento	Dormitórios	Ocupação (C)
Δ_{resf} **_sala_B**	Resfriamento	Salas	Sistemas construtivos (B)
Δ_{aquec} **_sala_B**	Aquecimento	Salas	Sistemas construtivos (B)
Δ_{resf} **_sala_C**	Resfriamento	Salas	Ocupação (C)
Δ_{aquec} **_sala_C**	Aquecimento	Salas	Ocupação (C)

Para a análise da validação relativa, foram aplicados seis índices estatísticos que quantificam o erro amostral de cada objeto em relação ao modelo de referência e média do grupo. Quanto menor o desvio encontrado, maior a qualidade da formação obtida pelo método de agrupamento em análise. Os índices utilizados foram: o erro máximo (ME), o erro absoluto médio (EAM), a raiz do erro médio quadrático normalizado (RMSE), o coeficiente de massa residual (CRM), o índice de concordância (d) e a eficiência (EF). Cada um deles foi selecionado considerando as habilidades em detecção de particularidades da distribuição de erros amostral.

O erro máximo é o índice mais simples de obtenção do erro. Ele considera o maior desvio encontrado na distribuição amostral. Quanto mais próximo de zero, menos dispersos são os objetos e melhor é a formação de agrupamento. O erro máximo foi calculado por meio da Equação 30.

$$ME = max(|O_i - P_i|)_{i=1}^{n} \quad (30)$$

Onde:
ME é o erro máximo;
O_i é o valor atribuído a cada objeto da amostra;
P_i é o valor atribuído ao modelo de referência.

O erro médio absoluto é uma alternativa ao erro máximo e representa a média dos erros amostrais. É uma medida menos afetada por objetos atípicos ou extremos e mede a tendência de superestimação ou subestimação do modelo. Quanto mais próximo a zero, melhor é a representatividade do modelo. O erro médio absoluto foi calculado por meio da Equação 31.

$$EAM = \left[\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)\right] \quad (31)$$

Onde:
EAM é o erro absoluto médio;
n_i é a quantidade de objetos do grupo;
O_i é o valor atribuído a cada objeto da amostra;
P_i é o valor atribuído ao modelo de referência.

A raiz do erro médio quadrático é similar ao erro médio absoluto. Entretanto, por elevar as diferenças individuais ao quadrado, torna-se mais sensível a desvios grandes. Adicionalmente, aplicando a raiz quadrada, obtém-se a medida de erro com a mesma amplitude

da variável analisada. A raiz do erro médio quadrático foi obtida por meio da Equação 32. A qualidade do método é maximizada à medida que a RMSE se aproxima de zero.

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2\right]^{0,5} \quad (32)$$

Onde:
RMSE é a raiz do erro médio quadrático normalizado;
n_i é a quantidade de objetos do grupo;
O_i é o valor atribuído a cada objeto da amostra;
P_i é o valor atribuído ao modelo de referência.

O coeficiente de massa residual representa uma medida de tendência de superestimar ou subestimar o modelo, variando de 1 a -1, respectivamente. A condição ótima é igual a zero. Essa medida é interessante pois acumula as diferenças de cada objeto em relação ao modelo e foi calculada por meio da Equação 33.

$$CRM = \frac{(\sum_{i=1}^{n} O_i) - (P_i * n_i)}{\sum_{i=1}^{n} O_i} \quad (33)$$

Onde:
CRM é o coeficiente de massa residual;
n_i é a quantidade de objetos do grupo;
O_i é o valor atribuído a cada objeto da amostra;
P_i é o valor atribuído ao modelo de referência.

O índice de concordância mede o grau de exatidão de uma observação, atingindo a condição ótima quando é igual a zero. Essa medida é interessante pois mede a relação de proximidade não apenas entre o objeto e o modelo, mas também entre esses e a média amostral. O índice de concordância foi obtido por meio da Equação 34.

$$d = \left[\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2}\right] \quad (34)$$

Onde:
d é o índice de concordância;
O_i é o valor atribuído a cada objeto da amostra;
P_i é o valor atribuído ao modelo de referência;
$\bar{O}$ é a média dos valores atribuídos a cada objeto da amostra.

Por fim, a medida de eficiência também representa uma relação entre a proximidade do objeto, da média amostral e do modelo. A condição ótima dessa medida é igual a um. A medida de eficiência foi obtida por meio da Equação 35.

$$EF = \frac{[\sum_{i=1}^{n}(O_i - \bar{O})^2 - \sum_{i=1}^{n}(O_i - P_i)^2]}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (35)$$

Onde:
EF é o coeficiente de eficiência;
O_i é o valor atribuído a cada objeto da amostra;
P_i é o valor atribuído ao modelo de referência;
$\bar{O}$ é a média dos valores atribuídos a cada objeto da amostra.

Os índices estatísticos apresentados acima foram calculados para todos os métodos resultantes da etapa de validação interna.

Por apresentar diferentes escalas numéricas, os índices foram recalculados, aplicando-se uma escala min-max (Equação 36). Essa ponderação foi feita de forma a possibilitar a agregação dos índices estatísticos em um único indicador por método, chamado aqui de índice global e calculado por meio da Equação 37.

$$E_j = \sum_{i=1}^{n} \frac{I_{ij} - \min(I_i)}{\max(I_i) - \min(I_i)} \quad (36)$$

$$E_{global} = \sum_{i=1}^{n} E_j \quad (37)$$

Onde:
E_j é o erro normalizado encontrado para índice estatístico, para cada método;
I_{ij} é a matriz de erros obtida a partir de cada índice estatístico;
$\min(I_i)$ é o menor erro obtido na matriz;
$\max(I_i)$ é o maior erro obtido na matriz;
E_{global} é o índice global encontrado para cada método.

A qualidade de cada método foi avaliada a partir do índice global de erros. O método de clusterização selecionado como mais adequado quanto à partição dos dados nessa pesquisa foi o que obteve o menor índice global.

3.6. Interpretação e caracterização dos resultados

3.6.1. Perfis a partir das características geométricas

Como resultado final, foram apresentados os perfis dos agrupamentos formados com o método resultante da seção 3.5.2. Os agrupamentos foram descritos a partir de variáveis envolvidas na análise de agrupamentos, listadas no Quadro 1 da seção 3.1.1. As variáveis qualitativas foram descritas com a frequência de casos por categoria. Variáveis quantitativas foram descritas a partir da média e desvio padrão dos objetos do grupo.

Informações a respeito dos modelos de referência foram apresentadas para cada uma das variáveis, juntamente com a descrição dos grupos, a fim de comparação. Verificou-se se o modelo apresentava valor similar à média do grupo e dentro do desvio padrão, ou correspondia à categoria de maior frequência.

Os modelos de referência também foram apresentados a partir da planta baixa, detalhando as principais medidas, configuração espacial dos ambientes (*layout*), orientação solar e relação com a via pública. Adicionalmente, foram apresentadas imagens da maquete eletrônica utilizada para a realização das simulações computacionais.

3.6.2. Perfis a partir do desempenho térmico

Adicionalmente à geometria, os grupos e modelos também foram caracterizados quanto ao seu desempenho térmico. Foram apresentados os valores de cada indicador de desempenho dos modelos, apresentados no Quadro 4 da seção 3.5.2. Os grupos foram descritos a partir da média e erro padrão.

A posição central do modelo foi analisada por meio de gráficos de caixa, verificando-se o desvio do modelo em relação ao centroide do grupo, considerando as medidas de média, mediana e quadrantes. Para assegurar a qualidade do modelo em relação à distribuição amostral do agrupamento, este deve estar locado entre o primeiro e terceiro quartis, que representam o intervalo que compreende 50% dos objetos da amostra. Quanto mais central estiver a posição do modelo, melhor a representatividade deste.

Para complementar o estudo, foram aplicados testes de hipóteses, tanto para a distribuição amostral quanto para as medianas. Por meio desses testes, pode-se comprovar estatisticamente se as amostras (ou seja, os grupos) se diferem ou se igualam. Adotou-se

significância estatística igual a 0,05. Dessa forma, foram consideradas amostras independentes sempre que o p_{valor} obtido foi menor ou igual a 0,05. Foram aplicados testes de hipóteses a todos os indicadores de desempenho listados no Quadro 4 da seção 3.5.2. Há diferentes testes de hipótese e aplica-se cada um dependendo da verificação da normalidade dos dados amostrais e também da quantidade de grupos a serem comparados. A decisão quanto à qual teste de hipótese aplicar baseou-se nas informações indicadas no Quadro 5.

Quadro 5 – Teste de hipóteses utilizados nas análises.

Análise	**Quantidade de agrupamentos formados**	**Verificação quanto à normalidade**	**Teste de hipótese aplicado**
Comparação entre agrupamentos	2 agrupamentos	Paramétrico	t de Student
		Não paramétrico	U de Whitney
	k agrupamentos	Paramétrico	t de Student
		Não paramétrico	ANOVA Kruska Wallis

4. Resultados

4.1 Composição do banco de dados inicial

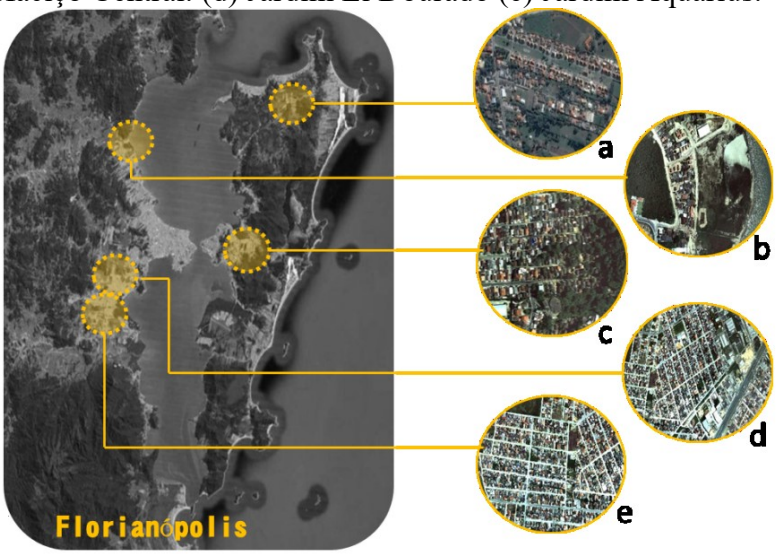
Nesta seção, foram apresentadas, de forma geral, as características que definem as habitações da amostra quanto à sua forma e geometria e também quanto ao desempenho térmico das habitações.

4.1.1. Dados referentes à geometria das edificações

A partir dos dados obtidos com o Projeto FINEP, 120 habitações foram levantadas, das quais dezoito foram excluídas por apresentarem dados inconsistentes ou incompletos, resultando em uma matriz de dados composta por 102 objetos e 35 variáveis. As variáveis que descrevem cada objeto foram transformadas em variáveis quantitativas, a fim de simplificar o processo de tratamento de dados, visto que as medidas de similaridade e algoritmos de partição aplicam-se geralmente para apenas um determinado tipo de variável (categórica, binária, ordinal, etc.). Dessa forma, ao transformar as variáveis todas para um mesmo tipo, é possível aplicar a análise de *cluster* a todos os dados, de forma conjunta. A preferência por variáveis quantitativas deu-se porque essas representam a maior parte dos dados obtidos.

A Figura 27 apresenta as regiões da Grande Florianópolis onde foram realizados os levantamentos.

Figura 27 – Localização dos pontos de levantamento: (a) Vargem Grande. (b) Foz do Rio. (c) Maciço Central. (d) Jardim El Dourado (e) Jardim Aquárius.



As Figuras 28 e 29 apresentam o resumo das características geométricas da amostra. Mostram equilíbrio entre a quantidade de habitações com sala e cozinha conjugadas ou independentes, prevalecendo, ainda assim, habitações com sala e cozinha conjugadas (56% dos casos). De forma geral, são habitações térreas (92% dos casos), com poucos casos com dois pavimentos. Prevalece a quantidade de habitações com dois dormitórios (44%), seguida de três dormitórios (31%). Habitações com três a quatro ambientes de permanência prolongada correspondem a quase 75% da amostra. Também predominam na amostra habitações com dois ambientes de permanência transitória (42%), correspondendo, geralmente, à cozinha e banheiro, embora haja casos com mais de um banheiro, área de serviço, garagem e ambiente específico para trabalho.

As áreas totais se concentram nas faixas entre $38m^2$ e $68m^2$. A distribuição de frequências da área de cobertura não apresenta distribuição normal, com valores um pouco abaixo das áreas totais devido à existência de habitações com dois pavimentos. O volume das habitações concentrou-se entre 84 m^3 e 164 m^3. O pé-direito, na maior parte da amostra, concentrou-se entre 2,30m e 2,65m. A soma das áreas de permanência prolongada variou de $15m^2$ a $75m^2$, aproximadamente. A soma das áreas de ambiente íntimo variou mais que a soma das áreas de ambiente social, o que pode ser explicado pelo fato de que o aumento da quantidade de dormitórios não é proporcional ao aumento das áreas sociais da habitação.

Quanto às áreas de parede e aberturas, percebe-se distribuição similar entre as orientações norte, leste e oeste, com área de parede voltada ao sul um pouco abaixo das demais. O aumento das áreas de janela não acompanhou linearmente o aumento das áreas de parede exposta, independente da orientação. A relação de área de janela por área de parede mostrou grande variação entre as habitações, concentrando-se entre 5% e 20%, com medianas próximas a 7% para todas as orientações, nos ambientes de permanência prolongada.

As tipologias habitacionais unifamiliares apresentaram *layouts* bem variados. Além da quantidade de ambientes e independência ou não das salas e cozinhas, a disposição interna dos ambientes mostrou-se bem heterogênea. Essa forma como configuram-se as geometrias representa uma das maiores dificuldades em lidar com a padronização desses objetos. Em outras palavras, as possibilidades de desenho da forma são tão grandes que dificultam o estabelecimento de critérios subjetivos de comparação entre um e outro.

A Figura 30 mostra algumas dessas habitações em planta baixa, como exemplo. A Hab_010 mostra a configuração de uma habitação com sala e cozinha conjugadas cujos dormitórios possuem acesso direto ao ambiente social (não há delimitação entre área íntima e

Figura 28 – Resumo das características geométricas da amostra: áreas e ambientes.

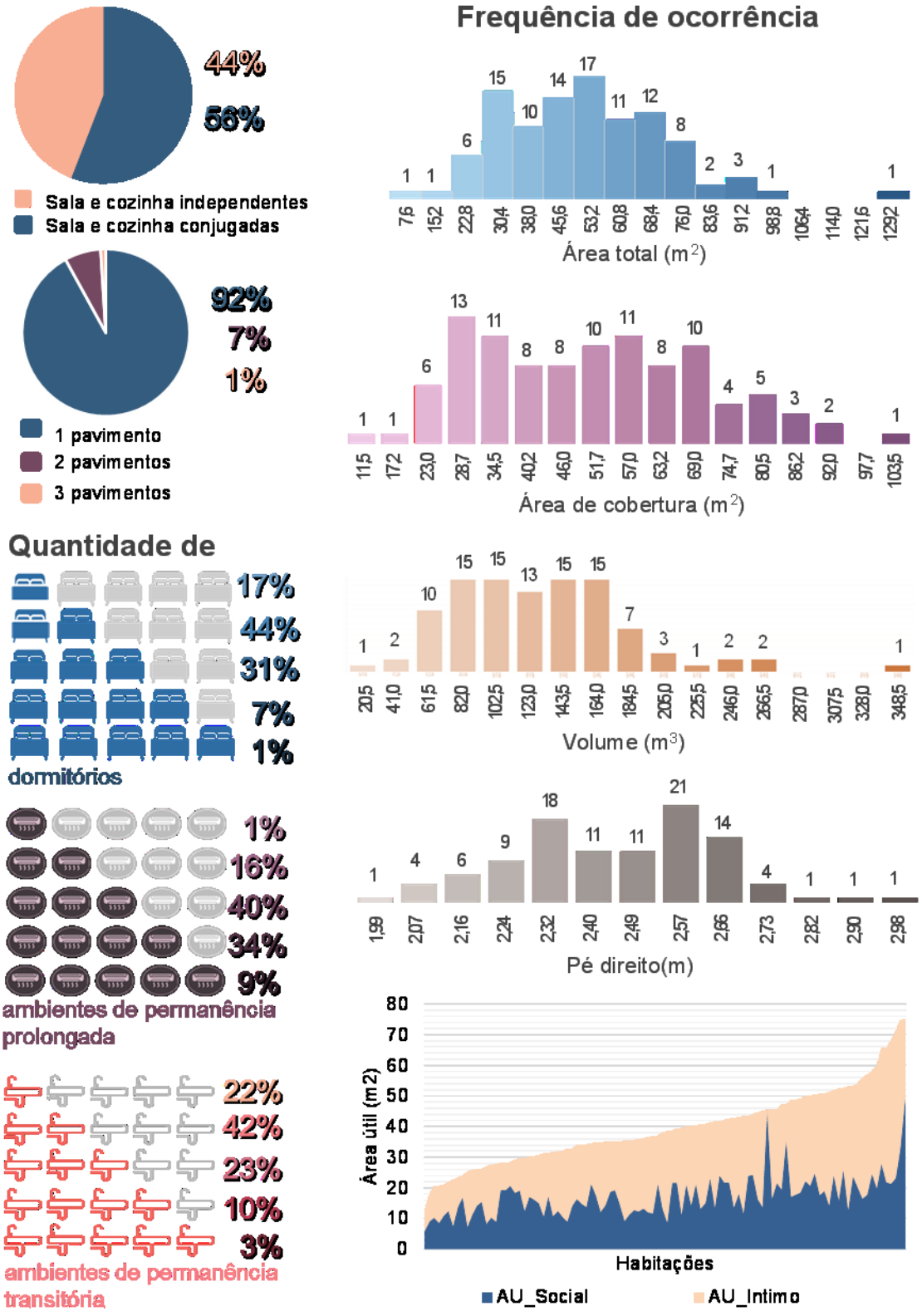
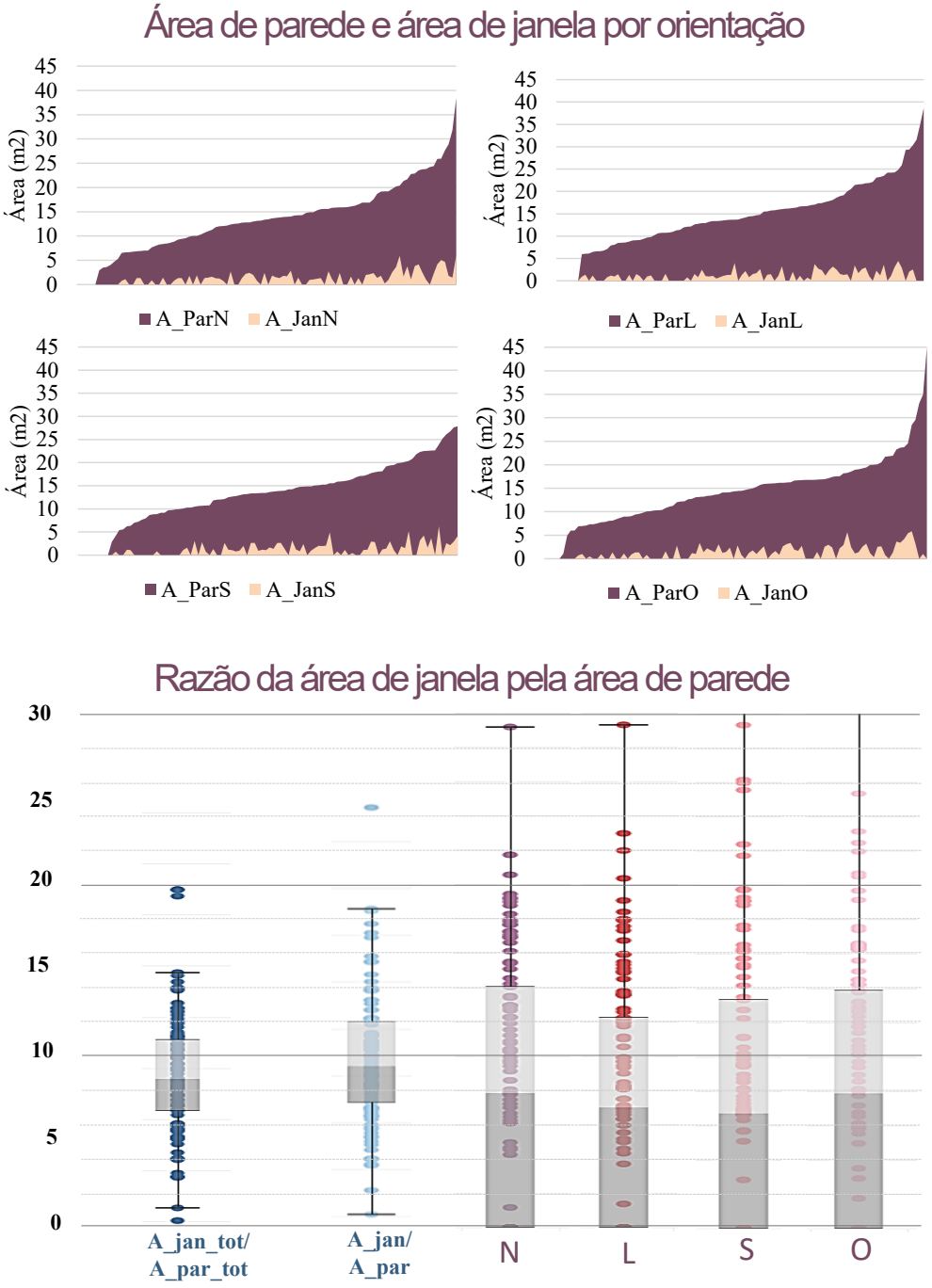


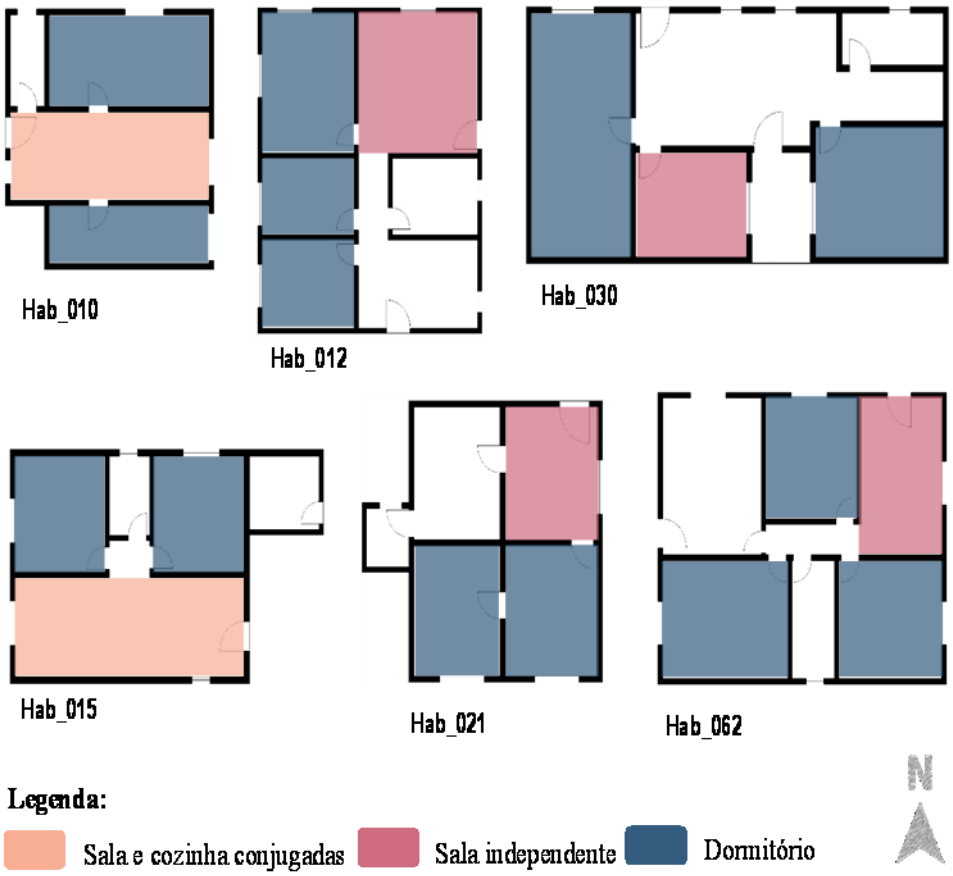
Figura 29 – Resumo das características geométricas da amostra: relação entre paredes e aberturas.



social, como geralmente ocorre em edificações multifamiliares ou habitações de maior classe social). A Hab_012 possui todos os dormitórios alinhados em uma mesma face. A Hab_015 apresenta a distribuição dos dormitórios separados pelo banheiro e organizados por um pequeno hall íntimo. Na Hab_021, o acesso a um dos dormitórios se dá pelo outro dormitório. Na Hab_030, os dormitórios estão em pontos extremos da casa e a circulação acontece pela

cozinha. Na Hab_062, há um corredor de acesso aos dormitórios, que estão dispostos em fachadas diferentes.

Figura 30 – Representação gráfica em planta baixa de algumas habitações com *layouts* variados.



Do ponto de vista arquitetônico, pode-se observar que os projetos representados na Figura 30, assim como boa parte das demais habitações da amostra, apresentam várias inadequações funcionais, como banheiros com porta em frente ao acesso de dormitórios (Hab_012) ou acesso aos dormitórios pela cozinha, sem uma separação das áreas sociais e áreas íntimas (Hab_030). Vale ressaltar que o objetivo deste estudo não é avaliar a qualidade das habitações e sim, achar uma forma de representar o estoque edificado, não cabendo aqui poderar sobre essas questões. As habitações levantadas compõem uma amostra de habitações de interesse social, das quais grande parte caracteriza-se por autoconstrução, ou seja, sem aprovação em órgão legal ou sequer acompanhamento por profissional da área. Dessa forma, é importante preservar a característica do estoque a ser investigado para que as respostas que se obtenham para sua melhoria sejam também adequadas.

A matriz completa com os dados está disponibilizada no Apêndice A. A representação gráfica de todas as edificações envolvidas nesse estudo está disponibilizada no Anexo A.

Muitos países classificam seu estoque edificado a partir de características como área construída e ano de contrução, o que pode ser feito pois as características arquitetônicas desse estoque são muito similares para uma mesma época de construção. Mesmo no Brasil, edificações como as comerciais e até residenciais multifamiliares, por seu caráter coletivo, acabam apresentando desenho arquitetônico com formato mais padronizado. A edificação residencial unifamiliar, por seu caráter individual, reflete na sua geometria características muito singulares e específicas do local e dos seus usuários, como mostrado por meio dos dados apresentados nessa seção. Essas particularidades dificultam a criação de padrões para esse tipo de edificação. Por esse motivo, se faz necessária a aplicação de um método que auxilie no processo de decisão de quais critérios usar para comparar e padronizar as unidades.

4.1.2. Dados referentes ao desempenho térmico das edificações

A Figura 31 apresenta o resumo dos indicadores de desempenho obtidos por meio das simulações computacionais. O eixo das ordenadas indica os valores do indicador de desempenho ponderado. Nas colunas, são apresentadas as distribuições amostrais para cada configuração (A, B e C), separadas por aquecimento ou resfriamento. As linhas horizontais representam o indicador de desempenho de cada habitação, para cada variável. Os valores destacados nas extremidades de cada distribuição representam o valor máximo e mínimo obtidos para a variável. Distribuições em azul representam os dormitórios e em rosa, as salas.

A rápida visualização dos dados antes da aplicação da análise de *cluster* auxilia no sentido de compreender ou criar expectativas sobre quais resultados podem ser obtidos ao final do processo para os modelos e grupos formados. Nesse sentido, é possível, ao final da análise, comparar os resultados encontrados (obtidos a partir dos modelos e grupos formados) com os esperados (que descrevem a amostra). No caso de haver divergência, há indicativo de que algo não foi bem executado e o processo deve ser examinado com mais cuidado.

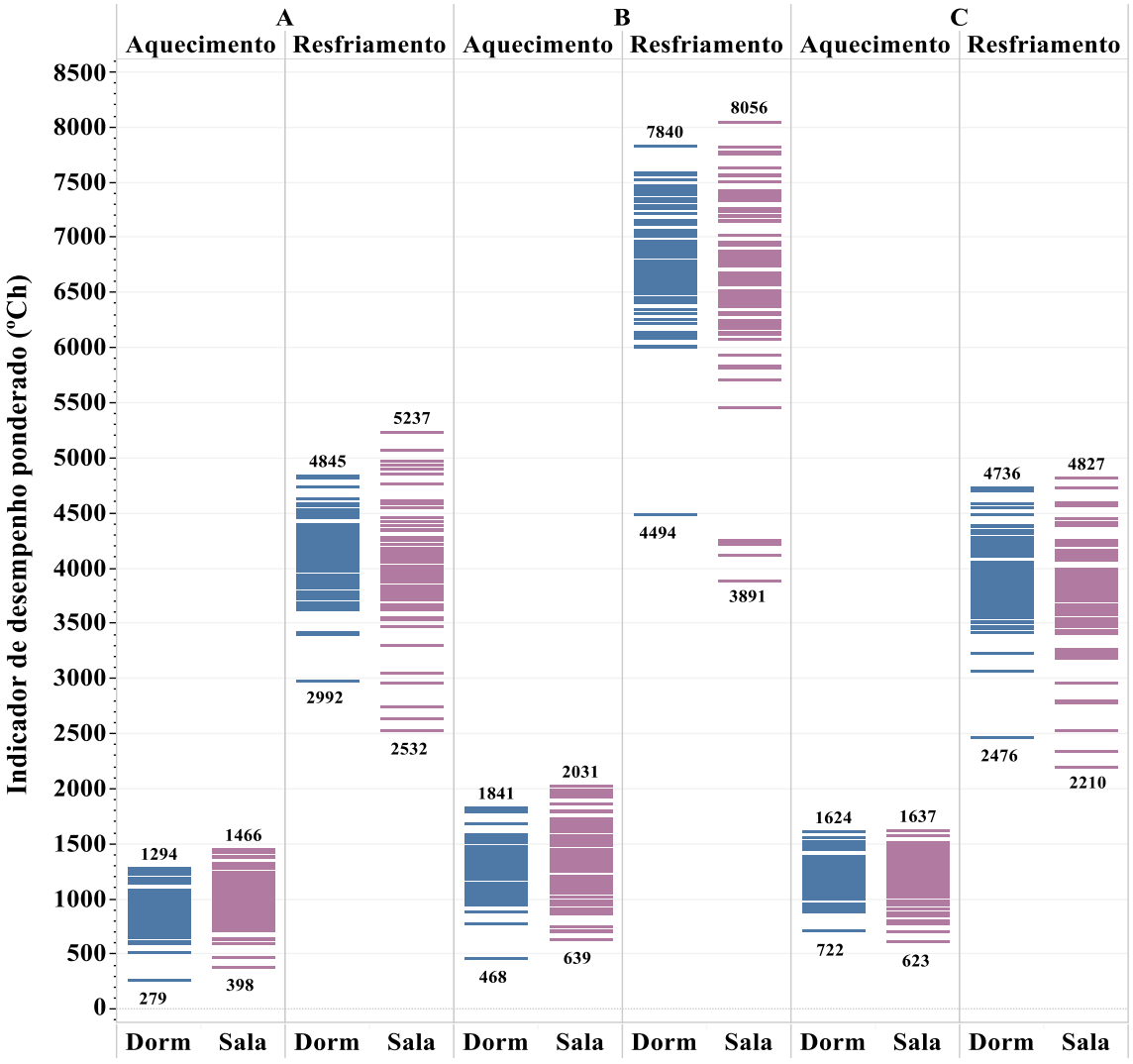
Na Figura 31, pode-se observar que a amostra apresenta maior desconforto por calor, com os indicadores de desempenho de resfriamento bem acima do aquecimento para todas as variações de configuração de simulação e ambientes. Os maiores valores foram encontrados para a configuração B. Essa variação foi configurada alterando-se os sistemas construtivos da edificação com maior transmitância e menor capacidade térmica, o que faz com que a geometria da edificação tenha maior influência no seu desempenho. Dessa forma, é importante que os

grupos formados a partir da análise de agrupamento apresentem independência estatística para essa variável, tanto para as salas quanto para os dormitórios. Outra expectativa é quanto as médias dos grupos. Para a configuração B, por exemplo, espera-se que fiquem entre 6.500 e 7.000ºCh, pois é onde está a maior concentração de casos.

Quanto ao indicador de desempenho para aquecimento, espera-se encontrar grupos com amostras similares, principalmente para as configurações A e C do dormitório, devido a sua distribuição apresentar pouca variância. Não devem ser encontrados valores acima de 2000ºCh.

Objetos com valores atípicos para o indicador de desempenho também devem ser encontrados. Esses valores estão representados pelas linhas horizontais mais afastadas da distribuição. Na amostra em análise, observa-se a presença de valores atípicos principalmente para o indicador de desempenho para resfriamento na sala.

Figura 31 – Resumo dos indicadores de desempenho das habitações obtidos por meio de simulações computacionais.



4.2. Formação das matrizes de dados

Com a Matriz de dados inicial (Matriz A) pronta, deu-se início ao processo de tratamento de dados. As demais matrizes (Matriz B, C, D e E) foram formadas a partir dos resultados obtidos com a submissão da Matriz A às diferentes combinações de tratamento. A seguir, são apresentados os resultados obtidos a partir da aplicação da padronização dos dados, da detecção de objetos atípicos e da ponderação dos fatores.

4.2.1. Padronização dos dados

Com esse tratamento, todos os dados foram transformados de forma a definir cada variável com um conjunto de dados cujos valores (*scores*) possuem média igual a zero e variam em termos de desvio padrão.

A Figura 32 mostra os valores padronizados obtidos para cada variável. As variáveis estão relacionadas no eixo horizontal e a escala de valores padronizados (*scores*), no eixo vertical. Os marcadores circulares representam o valor padronizado assumido por cada objeto ao longo das variáveis.

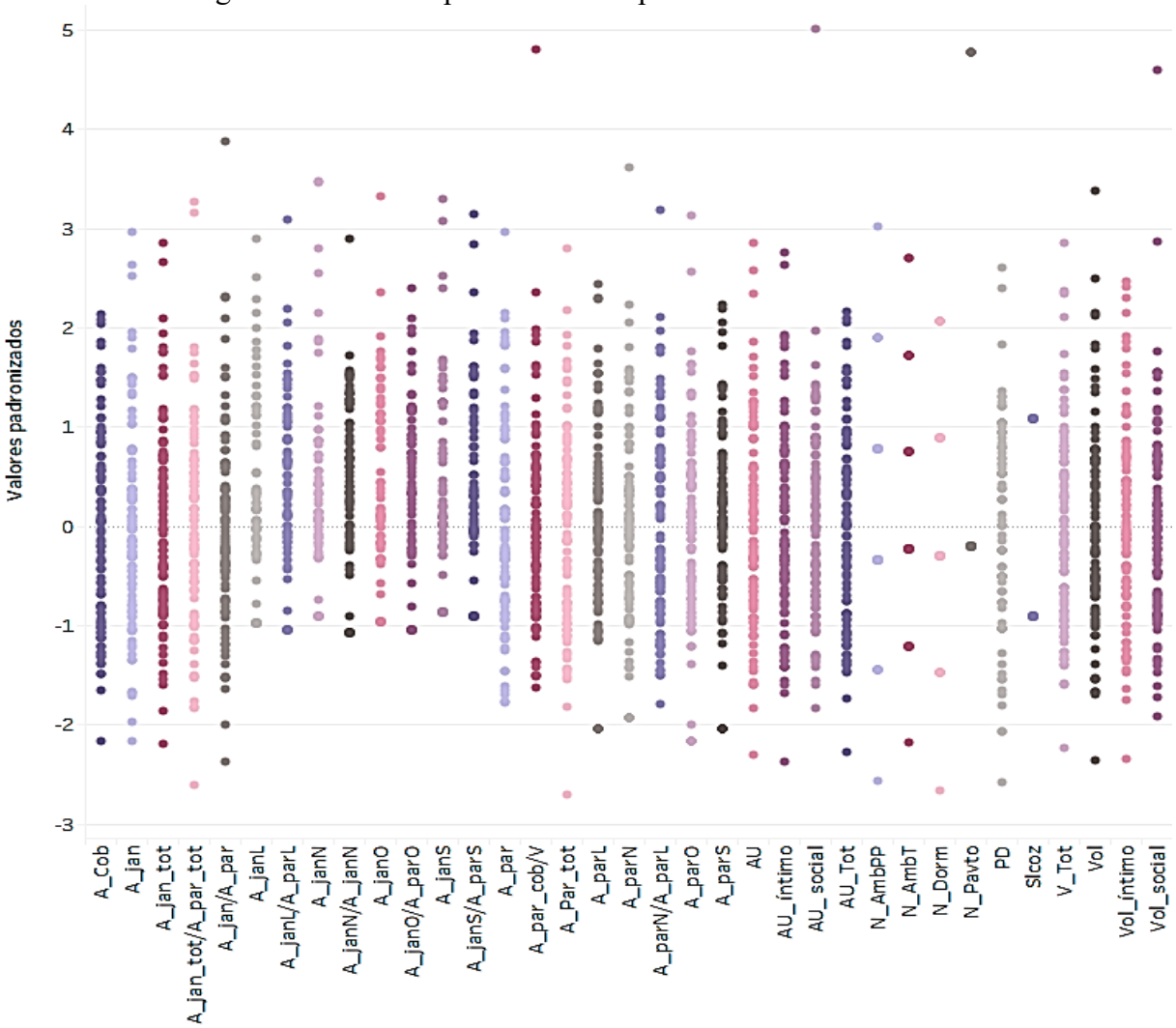
Os valores padronizados da amostra concentraram-se de forma geral em um intervalo entre -3 e 3, com valores concentrando-se em torno de zero, o que se espera normalmente de uma amostra bem representada. Valores padronizados acima ou abaixo desse intervalo normalmente representam objetos atípicos e pouco representativos. NaFigura 32, é possível verificar a presença de alguns casos com valores acima de 3, chegando próximo a 5. Este é um primeiro indicador de que há um ou alguns, porém poucos, objetos atípicos na amostra.

É possível distinguir as variáveis contínuas e discretas a partir da dispersão dos dados. Nas variáveis contínuas, os objetos distribuem-se de forma mais dispersa ao longo do intervalo, enquanto nas variáveis discretas, assumem valores bem pontuais. É possível observar, por exemplo, que a variável quantidade de dormitórios (N_dorm) resume-se a cinco possíveis valores, mesmo estando padronizados. Nessas variáveis, não fica tão clara a distribuição dos valores em relação a posição central, embora ainda seja possível perceber a posição de um determinado ponto mais próximo a zero (que aqui, representa a média do grupo) do que os demais. Essa observação permite inferir sobre a quantidade de dados que estão acima ou abaixo da média. Por exemplo, considerando a variável quantidade de pavimentos (N_pavto), nota-se

um marcador bem próximo a zero, enquanto outro muito acima. Isso significa que há uma quantidade grande de casos com valores próximos à média, enquanto poucos estão distantes.

Nas variáveis contínuas, é mais fácil de perceber a tendência dos dados a partir de sua dispersão. Na Figura 32, observa-se que a maioria das variáveis apresenta tendência de posição central próxima a zero, ou seja, existe uma tendência nessas variáveis de os objetos assumirem valores próximos à média. Em algumas variáveis, entretanto, verifica-se a tendência de os objetos assumirem valores acima da média, como é o caso das áreas de janela (como as variáveis A_janL, A_janN, A_janS, etc.).

Figura 32 – Valores padronizados a partir da medida *z-scores*.



A padronização dos dados foi aplicada na formação das Matrizes B, C e E.

4.2.2. Detecção de objetos atípicos

O próximo passo foi a detecção de objetos atípicos. A Tabela 7 mostra os resultados obtidos a partir da aplicação da medida D^2 de Mahalanobis, medida utilizada aqui para detectar objetos atípicos, dada a característica multivariável dos objetos. Na primeira coluna, são apresentados os objetos envolvidos na análise. Na segunda coluna, estão apresentados os valores de D^2, que representa a distância teórica de cada objeto até o centroide (média multivariada) do grupo. Quanto maior a distância ao centroide, maior também é a probabilidade desse objeto ser designado como atípico. Os valores da probabilidade associada ao D^2 estão apresentados na quarta coluna. Foram definidos como atípicos os objetos cuja probabilidade foi menor ou igual a 0,001. Como pode ser observado na Tabela 7, foram, ao todo, seis objetos designados como atípicos. Os objetos designados como atípicos estão destacados nas últimas linhas da Tabela 7.

Tabela 7 - Identificação de potenciais objetos atípicos com a medida D^2 de Mahalanobis.

Identificação dos objetos	**D^2 de Mahalanobis**	**Diferença do valor de D^2 em relação ao objeto anterior**	**Probabilidade associada ao D^2**
Hab_010	8,41	0,00	1,000
Hab_108	8,60	0,19	1,000
Hab_061	9,14	0,54	1,000
Hab_057	9,71	0,57	1,000
Hab_088	10,60	0,89	1,000
Hab_052	10,76	0,16	1,000
Hab_005	11,01	0,25	1,000
Hab_012	11,82	0,81	1,000
(valores intermediários omitidos)			
Hab_109	61,84	2,69	0,002
Hab_027	61,94	0,10	0,002
Hab_110*	**63,80**	**1,87**	**0,001**
Hab_091*	**73,12**	**9,31**	**0,000**
Hab_092*	**73,88**	**0,77**	**0,000**
Hab_082*	**74,14**	**0,25**	**0,000**
Hab_107*	**85,81**	**11,67**	**0,000**
Hab_016*	**95,51**	**9,70**	**0,000**

* Objetos identificados como atípicos dada a probabilidade associada ao D^2.

A Tabela 8 apresenta as principais características de cada um dos objetos designados como atípicos, de forma a permitir a compreensão de quais valores diferem do padrão da amostra ao longo das variáveis analisadas. Como pode ser observado na Tabela 8, de forma geral, os objetos designados como atípicos apresentaram dimensões maiores que o padrão da amostra. Dentre as apresentadas, destacam-se as variáveis quantidade de dormitórios, quantidade de pavimentos, área total, volume total, área de fachada e área e volume total dos ambientes condicionados. Os objetos Hab_016 e Hab_092 foram os que apresentaram as maiores diferenças. O objeto Hab_110 não se diferenciou tanto da amostra como os demais objetos atípicos. Entretanto, apresentou valores para algumas variáveis bem diferente das demais, como a quantidade de dormitórios, quantidade de pavimentos e pé direito.

A medida de Mahalanobis possibilita analisar os objetos de forma multivariada, ou seja, não considera apenas a discrepância nos dados em cada variável exclusivamente, mas sim o padrão que aquele conjunto de variáveis (que define cada objeto) representa em relação ao conjunto de objetos da amostra. Por isso, objetos como a Hab_091 que, de forma geral, não apresenta variáveis com valores que diferem muito da média, é designada como atípica. Esse objeto apresenta um padrão diferente das demais habitações, pois não apresenta uma proporção direta entre os dados. Por exemplo, embora tenha uma maior área de parede (A_par_tot) em relação à média, apresenta também a área de janela (A_jan_tot) e o volume dos ambientes de convívio íntimo (Vol_íntimo) bem menor que a média. Entretanto, a tendência da amostra indica que esses valores devem ser diretamente proporcionais.

A descrição completa dos objetos designados como atípicos está apresentada no Apêndice B.

Tabela 8 - Características dos objetos identificados como atípicos.

Variáveis envolvidas na análise	Identificação dos objetos						**Estatísticas da amostra**	
	Hab 016	Hab 107	Hab 082	Hab 092	Hab 091	Hab 110	Média	Desvio padrão
N_dorm	4	1	3	2	2	5	2	1
N_ambT	7	3	2	5	3	2	2	1
N_ambPP	6	4	4	4	3	6	3	1
N_pavto	2	1	3	1	2	2	1	0
A_tot	135,64	66,48	74,92	105,55	46,60	54,02	55,86	20,31

* Valores atípicos em relação à amostra.

Tabela 8 - Características dos objetos identificados como atípicos. (continuação).

Variáveis envolvidas na análise	Identificação dos objetos						**Estatísticas da amostra**	
	Hab 016	Hab 107	Hab 082	Hab 092	Hab 091	Hab 110	Média	Desvio padrão
A_cob	76,54	66,48	25,58	105,55	31,71	36,26	53,68	19,62
PD	2,60	2,45	2,35	2,56	2,45	2,15	2,49	0,19
V_tot	362,25	162,87	180,34	270,21	139,22	118,03	140,36	54,24
A_par_tot	209,56	163,67	150,02	114,63	107,46	78,91	81,99	24,12
A_jan	12,20	4,14	7,93	17,16	4,02	8,33	6,28	2,50
A_jan_tot/ A_par_tot	5,82	2,53	5,29	14,97	3,74	10,55	7,74	2,39
Au_íntimo	42,02	11,73	43,79	25,41	13,14	26,16	22,57	9,64
Au_social	32,85	35,55	21,84	49,53	21,31	9,21	17,21	6,99
A_ambPP	74,87	47,29	65,63	74,95	34,45	35,38	39,78	12,45
Vol_íntimo	109,26	28,75	105,04	65,06	32,18	56,27	56,31	24,40
Vol_social	85,42	87,11	51,33	126,81	52,21	21,69	42,97	17,92
Vol	194,67	115,85	156,37	191,87	84,39	77,95	99,28	32,31
A_par_ambPP	94,67	87,69	132,16	61,15	58,68	52,76	55,71	15,35

* Valores atípicos em relação à amostra.

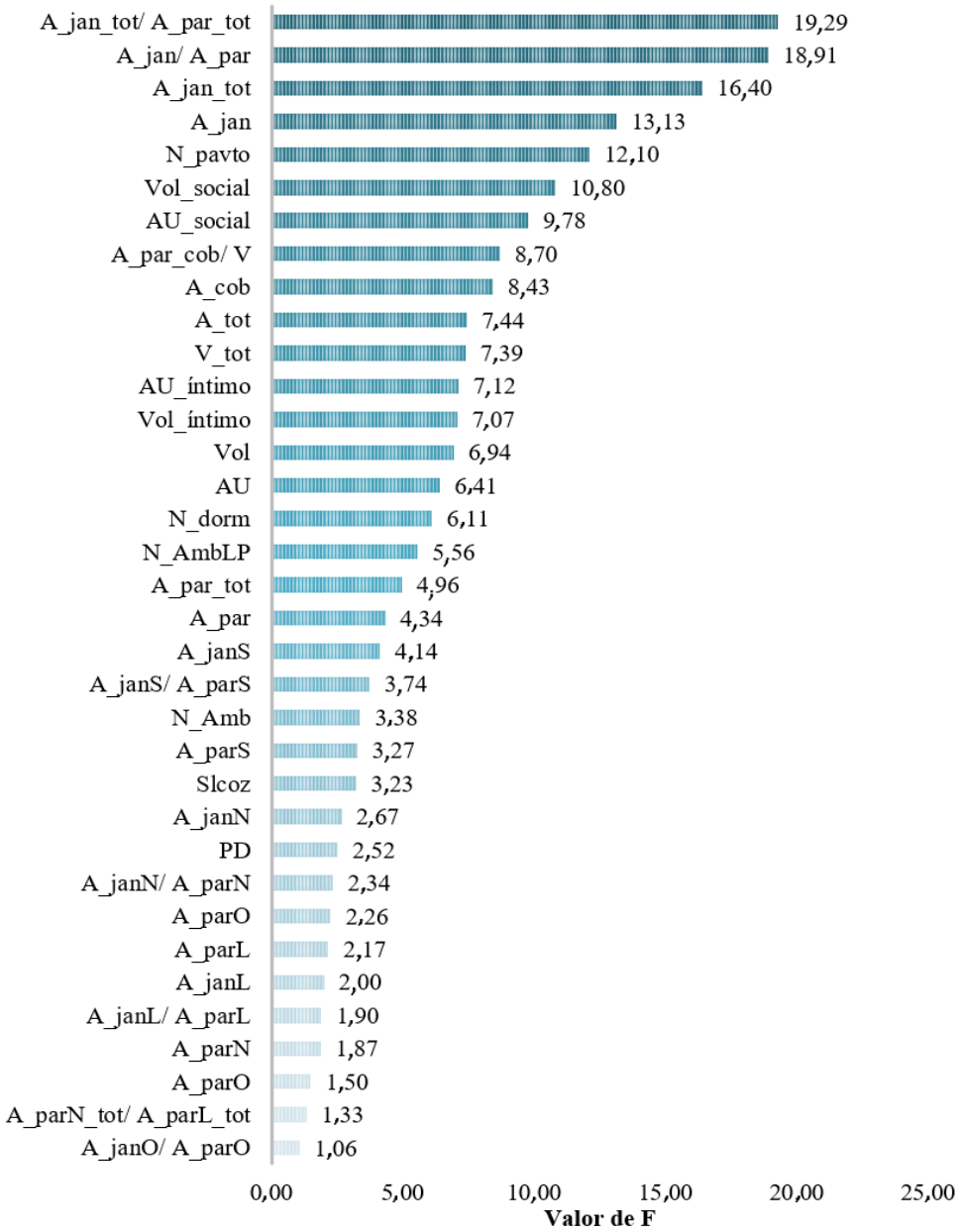
A detecção de objetos atípicos foi aplicada nas Matrizes B, D e E. Com os seis objetos excluídos, essas matrizes foram compostas por apenas 96 objetos.

4.2.3. Ponderação dos fatores

O último tratamento ao qual foram submetidos os dados foi a ponderação dos fatores. A Figura 33 apresenta os valores de F obtidos para cada variável, considerando os indicadores de desempenho do caso base (configuração A). As variáveis que apresentaram maior influência no desempenho térmico das habitações foram a razão da área de janela por área de parede (F=19,29), razão da área de janela por área de parede dos ambientes de permanência prolongada (F=18,91) e área de janela (F=16,40). As variáveis que apresentaram menor influência foram a razão entre área de janela e área de parede na fachada oeste (F=1,06), a proporção entre as dimensões da fachada norte e fachada leste (F=1,33) e a área de parede na fachada oeste (F=1,50).

É preciso ressaltar que este resultado se refere estritamente à amostra em análise, em função da variabilidade encontrada em cada variável, não devendo, portanto, ser extrapolada para outros estudos (a menos que seja para fins de comparação entre amostras). Em outras palavras, as variáveis que resultaram em baixo valor de F não necessariamente representam variáveis pouco importantes ou sem relevância para o desempenho térmico de edificações. Esse valor representa o quanto as variações ao longo da variável têm relação direta com as variações no desempenho térmico das edificações dessa amostra. Se não há grande variação na amostra em determinada variável, não vai haver grande influência dessa variável sobre o desempenho.

Figura 33 - Valores de F obtidos para cada variável.



A ponderação dos fatores foi aplicada nas Matrizes C, D e E.

4.3. Aplicação da análise de *cluster*

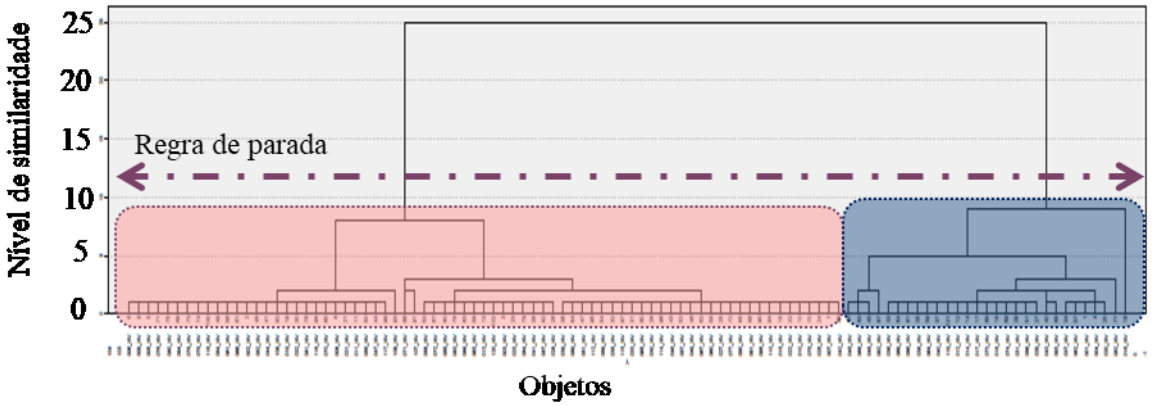
Os 105 diferentes métodos de agrupamento formados foram analisados e selecionados de acordo com a quantidade de agrupamentos mais adequada a cada método para as técnicas hierárquicas e não hierárquicas. Essas técnicas foram tratadas separadamente, pois os procedimentos para determinação da quantidade ideal de agrupamentos para cada uma delas diferem.

4.3.1. Formação e quantidade de agrupamentos a partir da técnica hierárquica

Primeiramente, a análise de agrupamentos foi realizada para os métodos que envolviam técnicas hierárquicas de agrupamento. Para cada um dos métodos, obteve-se um gráfico, denominado dendograma. O dendograma permite compreender passo a passo como se deu o processo de formação dos agrupamentos. Com o dendograma, é possível identificar visualmente uma solução prévia quanto à quantidade de grupos a serem formados. O critério utilizado para definir a quantidade do número de agrupamentos através do dendograma foi o nível de similaridade.

As Figuras 34 a 37 apresentam dendogramas obtidos a partir de três métodos que obtiveram diferentes soluções quanto à quantidade de grupos formados. A Figura 34 apresenta o dendograma obtido a partir do método M6_A, onde foram apresentadas todas as etapas do processo de aglomeração. No eixo horizontal, foram apresentados cada um dos objetos envolvidos na análise e no eixo vertical, o nível de similaridade obtido a cada união. A cada etapa, dois objetos foram unidos em um único agrupamento, até que no final do processo restou apenas um único grupo com todos os objetos. O nível de similaridade em que cada um dos grupos foi unido foi representado por uma linha horizontal, dentro do gráfico. Para determinar a formação dos agrupamentos, foi traçada uma linha de corte no momento em que uma dada junção provocou aumento do nível de similaridade relativamente maior que nas etapas anteriores. Na Figura 34, foi possível identificar esse aumento próximo ao nível de similaridade correspondente a 10. Abaixo dessa linha foi possível determinar a quantidade ideal de agrupamentos e quais objetos pertencem a cada agrupamento. No caso do método M6_A, a partir da análise do dendograma, a melhor divisão indica uma solução de dois agrupamentos.

Figura 34 - Dendograma obtido a partir do método M6_A.



As Figuras 35 a 37 apresentam exemplos de dendogramas obtidos cujas soluções de partição levam à sugestão quanto a divisão da amostra em 3, 4 e 5 grupos, respectivamente. A Figura 35 foi obtida a partir do método M19_D, a Figura 36 a partir do método M10_B e a Figura 37 corresponde ao método M10_E.

Figura 35 - Dendograma obtido a partir do método M19_D.

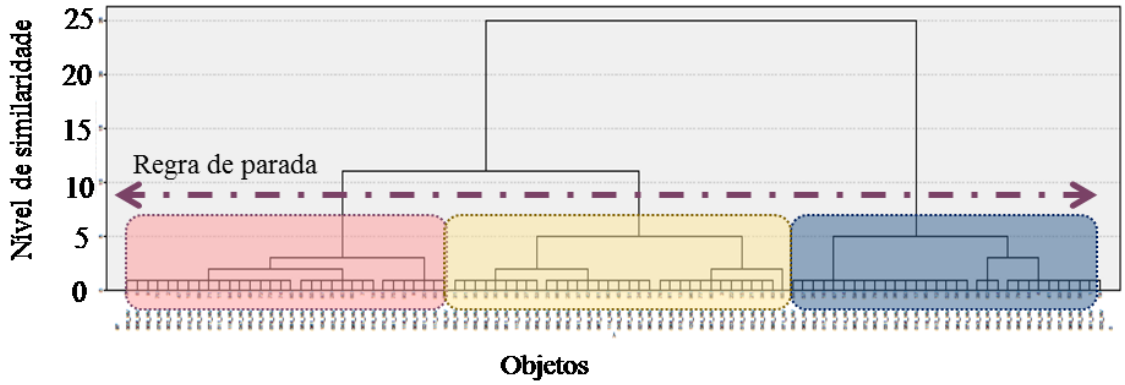


Figura 36 - Dendograma obtido a partir do método M10_B.

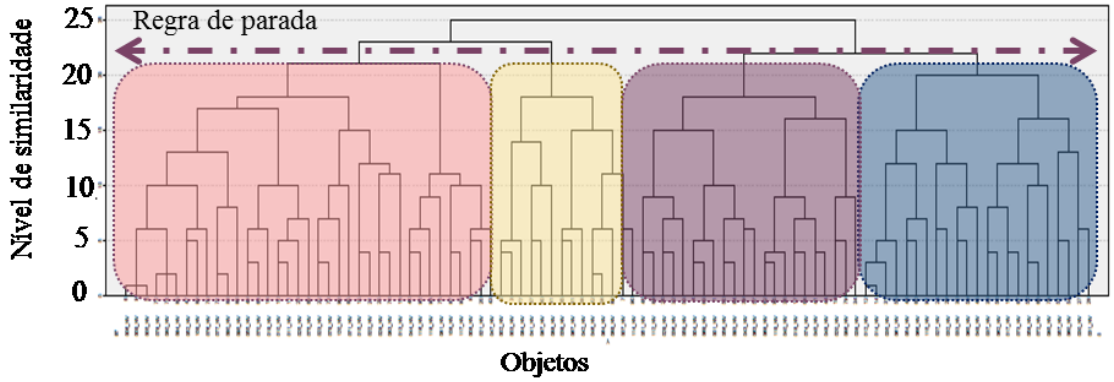
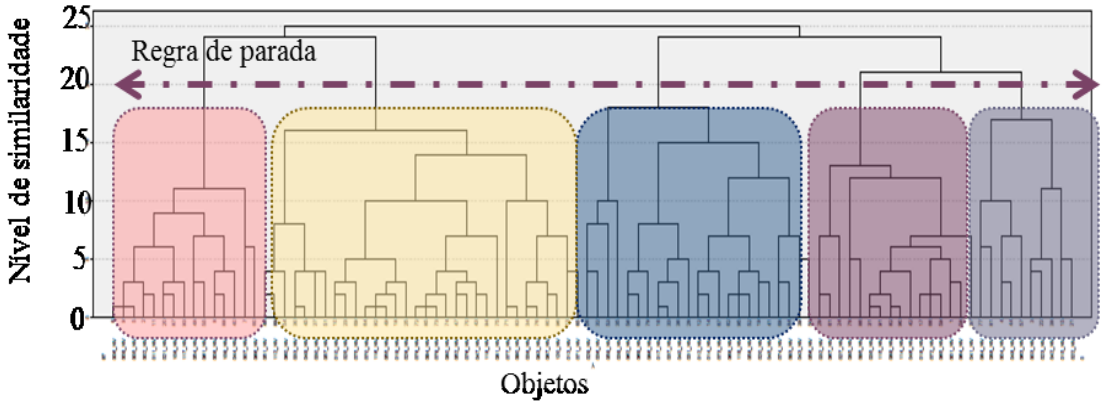
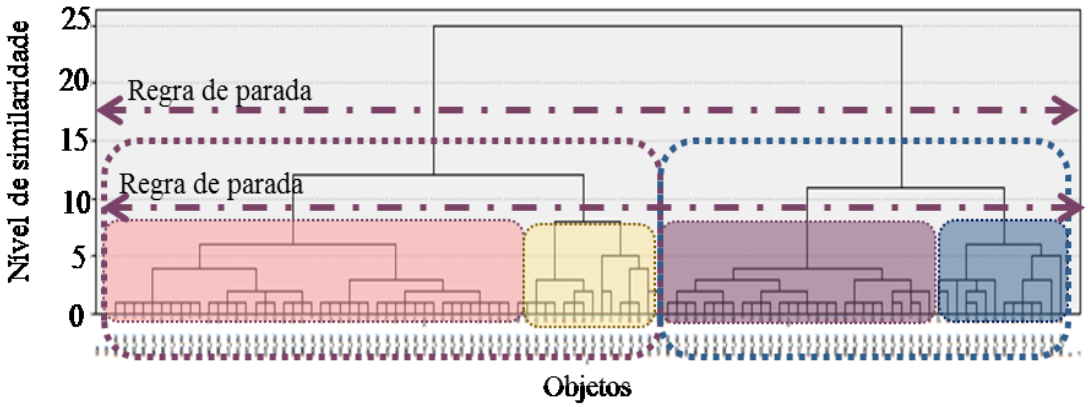


Figura 37 - Dendograma obtido a partir do método M10_E.



Nem sempre foi possível obter uma única solução quanto à quantidade de agrupamentos ideal a partir da leitura do dendograma. Em algumas situações, o dendograma indicou mais de uma solução, como é o caso da Figura 38.

Figura 38 - Dendograma obtido a partir do método M19_C.



Em casos como este, recorreu-se à análise da heterogeneidade. O valor de heterogeneidade foi obtido através do coeficiente de aglomeração. Nesse estudo, o programa de aglomeração foi obtido com o programa estatístico SPSS.

A Tabela 9 apresenta o programa de aglomeração obtido para o método M19_C, como exemplo. Na primeira coluna, foram listadas todas as etapas do processo de clusterização. Na segunda e terceira colunas, os objetos ou grupos que foram unidos na etapa correspondente. A quarta coluna apresentou-se o coeficiente de aglomeração, que indica o grau de heterogeneidade do grupo obtido a partir da união dos dois objetos (ou grupos). A quinta e sexta colunas indicam em que etapa os objetos unidos foram incorporados pela primeira vez a outro objeto ou

agrupamento. Por fim, a última coluna indica em qual etapa o agrupamento recém-formado volta a se unir com outro agrupamento. Por exemplo, na etapa 6, os objetos 73 e 89 foram unidos, alcançando um coeficiente de aglomeração correspondente a 12,6. O objeto 73 já havia se unido ao objeto 74 na etapa 3, mas o objeto 89 ainda formava um grupo unitário. Esse agrupamento irá sofreu uma nova junção na etapa 24.

Tabela 9 - Programa de aglomeração do método M19_C.

Etapa	**Agrupamentos combinados**		**Coeficiente de Aglomeração**	**Etapa em que o agrupamento apareceu pela primeira vez**		**Próxima etapa em que o novo agrupamento aparece**
	Agrupamento 1	Agrupamento 2		Agrupamento 1	Agrupamento 2	
1	6	9	0,6	0	0	31
2	10	48	2,3	0	0	7
3	73	74	4,1	0	0	6
4	53	101	6,9	0	0	25
5	72	85	9,7	0	0	56
6	73	89	12,6	3	0	24
7	10	24	15,5	2	0	30
8	12	14	184	0	0	25
9	2	5	21,4	0	0	31
10	46	47	24,4	0	0	32
(Etapas intermediárias omitidas)						
91	66	72	778,2	76	56	95
92	1	3	814,4	86	84	99
93	2	4	856,1	70	80	96
94	18	87	898,3	83	0	97
95	16	66	943,0	82	91	98
96	2	10	996,2	93	88	100
97	18	21	1053,6	94	90	99
98	16	19	1135,7	95	87	100
99	1	18	1245,5	92	97	101
100	2	16	1363,6	96	98	101
101	1	2	1620,9	99	100	0

A partir do coeficiente de aglomeração, calculou-se o aumento percentual a cada nova etapa, valor a partir do qual foi determinada a regra de parada. A Tabela 10 apresenta a Regra de Parada do método M19_C, a fim de exemplificar esse procedimento. Na primeira e segunda

colunas, apresenta-se as últimas etapas de clusterização e a quantidade de grupos formados naquela etapa. O coeficiente de aglomeração é apresentado na terceira coluna. O aumento percentual é apresentado na quarta coluna. Observa-se que o maior aumento acontece na última etapa, resultando em um valor que representa quase o dobro do obtido na etapa anterior. Isso mostra que aquela união provocou um aumento muito grande na heterogeneidade do grupo, indicando que a união entre os dois últimos agrupamentos não é adequada. Dessa forma, adota-se dois como quantidade ideal de agrupamentos.

Tabela 10 - Regra de parada do método M19_C a partir do coeficiente de aglomeração.

Etapa	**Regra de Parada**		
	Quantidade de agrupamentos formados em cada etapa	**Coeficiente de aglomeração**	**Aumento percentual do coeficiente de aglomeração em relação a etapa anterior (%)**
	(Etapas anteriores omitidas)		
92	10	814,4	4,7
93	9	856,1	5,1
94	8	898,3	4,9
95	7	943,0	5,0
96	6	996,2	5,6
97	5	1053,6	5,8
98	4	1135,7	7,8
99	3	1245,5	9,7
100	2	1363,6	9,5
101	1	1620,9	18,9

Alguns métodos combinaram medidas de similaridade e algoritmos de partição cujos resultados não permitiram uma separação clara da amostra. Essa dificuldade pode ser observada nas Figuras 39 a 41, correspondendo aos métodos M3_A, M4_ B e M20_E, respectivamente. Nas Figuras 39 e 40, os últimos agrupamentos são agregados de uma forma que seria necessário dividir a amostra em muitos subgrupos para se obter uma divisão sem que houvesse a formação de um ou mais grupos unitários. Na Figura 41, essa situação é ainda mais problemática, pois nem sequer é possível fazer qualquer divisão, visto que todos os objetos são unidos simultaneamente em um mesmo agrupamento, a um mesmo nível de similaridade. Os métodos que apresentaram dendograma com esse tipo de comportamento foram excluídos do estudo e estão listados no Quadro 6.

Nenhum dos métodos que foram configurados combinando o algoritmo de Ligação Simples originaram dendogramas capazes de permitir a identificação de uma separação clara da amostra. Uma das características desse algoritmo é a capacidade de identificar formas irregulares. Isso evidencia a probabilidade de a amostra em análise ser composta por grupos com características regulares e que assumem algum padrão ao longo das variáveis. Esse fato também já foi evidenciado pelos resultados apresentados na seção 4.2. O estudo de Geyer et al. (2017) também não obteve bons resultados ao utilizar esse algoritmo.

O método do centroide também não apresentou bons resultados a partir da análise do dendograma, com exceção quando combinado com o coeficiente de correlação de Pearson. Por sua vez, essa medida de similaridade também não apresentou bons resultados com o algoritmo de Ward. O coeficiente de Pearson apresenta capacidade melhor de identificar correlações entre os objetos (HAIR et al., 2009). A partir disso, pode-se supor que a amostra de habitações se caracteriza mais por relações de proximidade do que por padrões.

Figura 39 - Dendograma obtido a partir do método M3_A.

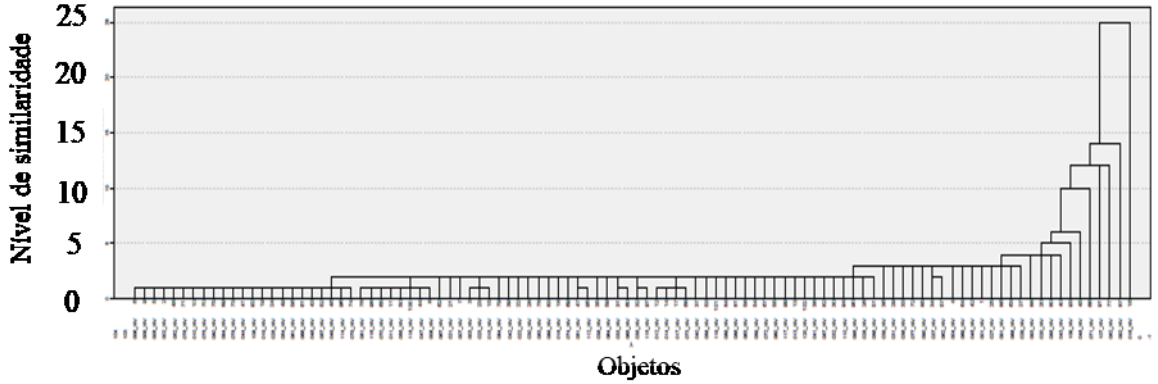


Figura 40 - Dendograma obtido a partir do método M4_B.

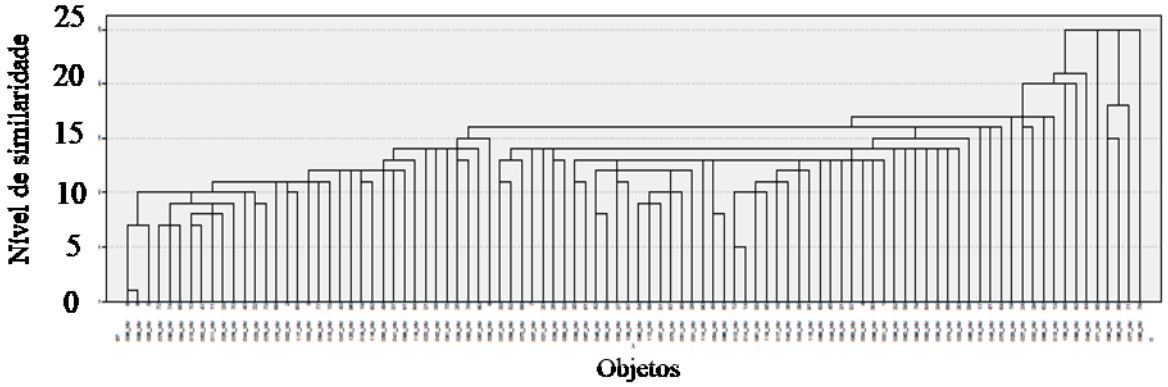
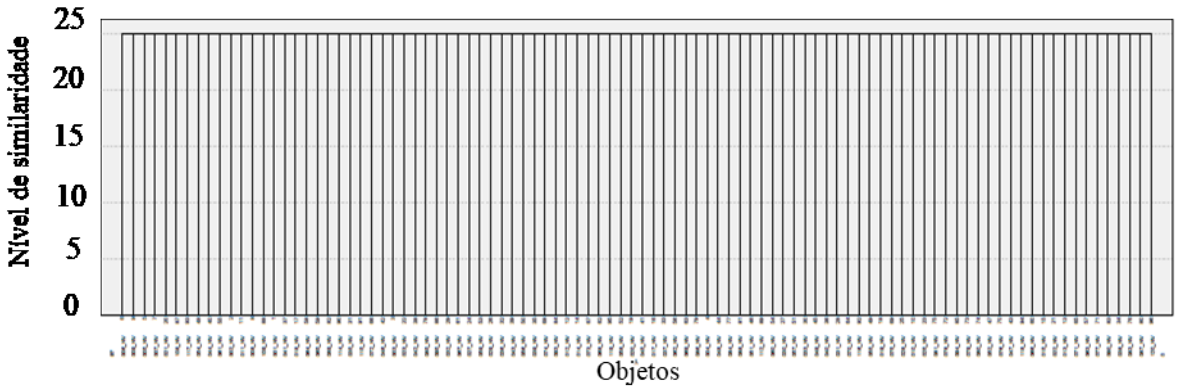


Figura 41 - Dendograma obtido a partir do método M20_E.



Quadro 6 - Métodos eliminados do estudo devido à configuração do dendograma.

Métodos		**Matriz de dados**				
Algoritmo de partição	**Medida de similaridade**	A	B	C	D	E
Ligação simples	City-Block	M01_A	M01_B	M01_C	M01_D	M01_E
	Euclidiana	M02_A	M02_B	M02_C	M02_D	M02_E
	Euclidiana quadrada	M03_A	M03_B	M03_C	M03_D	M03_E
	Chebyshev	M04_A	M04_B	M04_C	M04_D	M04_E
	Pearson	M05_A	M05_B	M05_C	M05_D	M05_E
Ligação Completa	Chebyshev	-	-	M09_C	-	M09_E
	Pearson	M10_A	-	M10_C	-	-
Centroide	City-Block	M11_A	M11_B	M11_C	M11_D	M11_E
	Euclidiana	M12_A	M12_B	M12_C	M12_D	M12_E
	Euclidiana quadrada	M13_A	M13_B	M13_C	-	M13_E
	Chebyshev	M14_A	M14_B	M14_C	M14_D	M14_E
	Pearson	-	M15_B	-	-	-
Ward	Pearson	M20_A	-	-	M20_D	-

Os dendogramas e programas de aglomeração apresentados foram selecionados como exemplos para compreensão dos resultados obtidos a partir da análise de agrupamento. Os demais resultados foram suprimidos a fim de não estender essa seção demasiadamente. Os dendogramas foram obtidos para todos os métodos apresentados na seção 3.3.3 e o programa de aglomeração foi calculado para todos os métodos para os quais não foi possível definir a quantidade ideal de grupos a partir do dendograma.

4.3.2. Formação e quantidade de agrupamentos a partir da técnica não hierárquica

Na sequência, foram realizadas as análises da técnica não hierárquica de agrupamento. Nesse tipo de procedimento, a designação dos objetos aos grupos é feita de forma interativa, a partir de pontos sementes pré-definidos. Por esse motivo, para a aplicação desse tipo de algoritmo é informada, previamente, a quantidade de grupos a serem formados. Como não há um valor preestabelecido, considerou-se nesse estudo todas as soluções quanto à quantidade de agrupamentos obtidas a partir das técnicas hierárquicas. Nesse sentido, as análises não hierárquicas foram realizadas com o algoritmo K-means adotando soluções com 2, 3, 4 e 5 agrupamentos. Essas configurações foram aplicadas para todas as cinco matrizes de dados. Os pontos sementes foram definidos aleatoriamente a partir de algoritmos do programa SPSS. Ao todo, foram formados agrupamentos a partir de 20 métodos (cinco matrizes de dados x quatro soluções quanto à quantidade de grupos formados).

Para prosseguir, foi necessário então decidir qual das soluções (quanto a quantidade de agrupamentos) deveria prosseguir, para cada matriz de dado.

As Tabelas 11 a 15 apresentam o histórico de interação para todos os métodos aplicando o algoritmo K-means, considerando até 10 interações. Pode-se observar que para a Matriz A, método k igual a 2, a convergência dos *clusters* 1 e 2 aconteceu na 9ª interação (momento em que a mudança da posição do centroide é igual a zero, ou seja, momento a partir do qual não houve alteração na estrutura do agrupamento). Da mesma forma se leem os demais casos.

É possível observar que a convergência não foi alcançada em todos os métodos (ao menos não até a 10ª interação). Na Matriz A, a convergência só foi alcançada para os métodos com k igual a 2 e k igual a 3 (9ª interação). Para os demais métodos utilizando essa matriz, pelo menos 1 dos grupos não alcançou a convergência. Adotando-se k igual a 4, nenhum dos grupos convergiu e adotando-se k igual a 5, o *cluster* 1 só convergiu na 10ª interação. Nas demais matrizes, todos os métodos chegaram à convergência antes das dez interações, com exceção do método que dividiu a amostra em 5 grupos. A única matriz que alcançou a convergência com 5 grupos foi a Matriz B. A Matriz B foi a única que alcançou a convergência em todos os 4 métodos.

No histórico de interação também fica em evidência a presença de objetos atípicos, o que ocorreu principalmente nas matrizes que não foram submetidas ao tratamento de dados para retirá-los (Matriz A e Matriz C). Objetos atípicos podem ser detectados em grupos que se

formaram com um único objeto (como o *cluster* 2 do método que formou quatro grupos a partir da Matriz A). Ele é facilmente identificado, pois no seu histórico de interação, a convergência já se dá na primeira interação (por ser um objeto bem distinto, nenhum outro objeto é designado a esse agrupamento, cujo centroide passa a ser ele mesmo). É pouco provável que um grupo com mais objetos já alcance a convergência na primeira interação, pois isso implicaria em acertar o centroide do grupo antes de qualquer interação e nenhum objeto ser designado a outro agrupamento.

Na Matriz A, esses agrupamentos unitários foram formados pelo objeto Hab_016. Ao observar a Tabela 7 (D^2), pode-se verificar que esse objeto já havia sido designado como atípico. Grupos pequenos que não alcançam a convergência também podem identificar a presença de objetos atípicos, como é o caso do método que formou cinco grupos a partir da Matriz C. Nesse método, o agrupamento 1 é formado por um agrupamento unitário, composto pelo objeto Hab_092. O agrupamento 5 é formado por apenas dois objetos, sendo eles o Hab_016 e Hab_082. Esses objetos também foram identificados como atípicos através do método D^2. Na Matriz C, observa-se claramente a presença de objetos atípicos, pois, para k igual a 3, k igual a 4 e k igual a 5, a convergência já é alcançada na 1ª interação em pelo menos um dos grupos.

Tabela 11 - Histórico de interação a partir dos centroides dos grupos para a Matriz A.

Matriz A		**Histórico de interação (alterações nos centroides dos grupos)**									
Quantidade k de grupos	**Nome do grupo**	1	2	3	4	5	6	7	8	9	10
k=2	**1**	87,312	79,957	68,781	3,731	0,098	1,737	0,045	0,001	**0,000**	
	2	160,185	7,099	25,783	1,373	0,021	1,089	0,017	**0,000**		
k=3	**1**	28,551	5,113	0,109	0,002	**0,000**					
	2	105,320	9,363	0,177	0,003	**0,000**					
	3	71,422	32,697	6,539	1,308	0,262	0,052	0,010	0,002	**0,000**	
k=4	**1**	112,581	1,535	1,420	0,028	2,578	8,034	4,175	2,715	0,075	0,002
	2	**0,000**									
	3	51,050	48,912	9,968	23,068	5,065	3,877	3,829	0,225	0,013	0,001
	4	121,256	11,001	2,956	6,947	4,310	8,187	3,809	2,017	0,040	0,001
k=5	**1**	51,050	30,197	5,033	0,839	0,140	0,023	0,004	0,001	0,001	**0,000**
	2	93,460	3,075	0,081	0,002	**0,000**					
	3	**0,000**									
	4	113,669	7,464	0,257	0,009	**0,000**					
	5	86,174	4,184	0,131	0,004	**0,000**					

Tabela 12 - Histórico de interação a partir dos centroides dos grupos para a Matriz B.

Matriz B		**Histórico de interação (alterações nos centroides dos grupos)**									
Quantidade k de grupos	**Nome do grupo**	1	2	3	4	5	6	7	8	9	10
k=2	**1**	8,129	0,305	0,008	**0,000**						
	2	7,658	0,116	0,002	**0,000**						
k=3	**1**	6,908	0,187	0,162	0,005	**0,000**					
	2	2,854	0,713	1,351	0,270	0,054	0,011	0,002	**0,000**		
	3	7,500	0,129	0,002	**0,000**						
k=4	**1**	7,466	0,131	0,002	**0,000**						
	2	2,854	0,713	1,351	0,270	0,054	0,011	0,002	**0,000**		
	3	**0,000**									
	4	6,908	0,187	0,162	0,005	**0,000**					
k=5	**1**	6,980	1,963	0,214	0,006	**0,000**					
	2	**0,000**									
	3	2,854	0,713	0,178	0,045	0,011	0,003	0,001	**0,000**		
	4	2,295	1,792	0,179	0,018	0,002	**0,000**				
	5	5,494	2,197	0,155	0,003	**0,000**					

Tabela 13 - Histórico de interação a partir dos centroides dos grupos para a Matriz C.

Matriz C		**Histórico de interação (alterações nos centroides dos grupos)**									
Quantidade k de grupos	**Nome do grupo**	1	2	3	4	5	6	7	8	9	10
k=2	**1**	106,774	21,203	6,949	0,193	0,005	**0,000**				
	2	69,994	4,705	3,567	0,052	0,001	**0,000**				
k=3	**1**	**0,000**									
	2	60,398	3,351	0,049	0,001	**0,000**					
	3	73,623	6,844	0,196	0,006	**0,000**					
k=4	**1**	**0,000**									
	2	74,556	14,624	2,782	0,070	0,002	**0,000**				
	3	43,986	17,749	2,055	0,035	0,001	**0,000**				
	4	53,305	5,374	1,075	0,215	0,043	0,009	0,002	**0,000**		
k=5	**1**	**0,000**									
	2	75,270	13,726	4,480	0,115	0,003	**0,000**				
	3	16,801	5,600	1,867	0,622	0,207	0,069	0,023	0,008	0,003	0,001
	4	43,986	18,013	3,359	0,056	0,001	**0,000**				
	5	25,774	8,591	2,864	0,955	0,318	0,106	0,035	0,012	0,004	0,001

Tabela 14 - Histórico de interação a partir dos centroides dos grupos para a Matriz D.

Matriz D		**Histórico de interação (alterações nos centroides dos grupos)**									
Quantidade k de grupos	**Nome do grupo**	1	2	3	4	5	6	7	8	9	10
k=2	**1**	972,692	140,826	19,640	0,401	0,008	**0,000**				
	2	887,701	76,054	18,940	0,387	0,008	**0,000**				
k=3	**1**	848,006	35,038	0,973	0,027	0,001	**0,000**				
	2	641,088	11,119	0,347	0,011	**0,000**					
	3	925,262	38,379	1,238	0,040	0,001	**0,000**				
k=4	**1**	237,385	15,394	13,919	0,366	0,010	**0,000**				
	2	706,608	74,353	1,582	0,034	0,001	**0,000**				
	3	**0,000**									
	4	449,189	213,983	44,898	3,454	0,266	0,020	0,002	**0,000**		
k=5	**1**	307,405	61,481	12,296	2,459	0,492	0,098	0,020	0,004	0,001	0,001
	2	479,287	19,228	39,108	12,367	0,458	0,017	0,001	**0,000**		
	3	256,836	68,285	96,980	10,700	0,306	0,009	**0,000**			
	4	**0,000**									
	5	439,429	163,705	55,914	1,747	0,055	0,002	**0,000**			

Tabela 15 - Histórico de interação a partir dos centroides dos grupos para a Matriz E.

Matriz E		**Histórico de interação (alterações nos centroides dos grupos)**									
Quantidade k de grupos	**Nome do grupo**	1	2	3	4	5	6	7	8	9	10
k=2	**1**	75,662	14,574	2,349	0,057	0,001	**0,000**				
	2	43,497	18,465	1,807	0,032	0,001	**0,000**				
k=3	**1**	75,774	13,899	4,488	0,118	0,003	**0,000**				
	2	16,801	5,600	1,867	0,622	0,207	0,069	0,023	0,008	0,003	0,001
	3	43,497	18,121	3,476	0,060	0,001	**0,000**				
k=4	**1**	50,456	3,350	1,840	1,721	0,051	0,001	**0,000**			
	2	49,664	4,805	6,702	0,268	0,011	**0,000**				
	3	**0,000**									
	4	39,458	7,238	6,553	1,723	0,044	0,001	**0,000**			
k=5	**1**	20,348	18,591	13,627	6,050	0,432	0,031	0,002	**0,000**		
	2	**0,000**									
	3	16,801	5,600	1,867	0,622	0,207	0,069	0,023	0,008	0,003	0,001
	4	50,747	6,711	6,959	2,439	0,072	0,002	**0,000**			
	5	56,619	1,241	3,865	0,081	0,002	**0,000**				

Um indício de que determinada divisão da amostra formou agrupamentos com boa estabilidade é a convergência nas primeiras etapas do histórico de interação. Dessa forma, é possível ter um indicativo que a melhor formação para a maioria dos métodos é aquela que divide a amostra em dois grupos. Foram selecionados para prosseguir na análise todos aqueles que alcançaram a convergência antes da 10ª interação.

Uma forma adicional para identificar a quantidade ideal de agrupamentos a serem formados por cada método é através da análise das variâncias obtidas a cada aumento da quantidade de agrupamentos. Na Figura 42, são apresentados alguns valores de F para todos os métodos e soluções de agrupamento. Uma redução no valor de F de uma determinada variável em uma solução de k grupos para outra com k mais 1 grupos significa que essa mudança provocou menor impacto na diferenciação dos grupos. Por exemplo, tomando por base a variável área total (A_tot), considerando a formação de dois agrupamentos com a Matriz B, obteve-se F igual a 106,5. Esse valor cai para 69,7 quando a amostra é dividida em três agrupamentos e 46,7 quando dividida em quatro agrupamentos. Isso significa que a diferença entre os agrupamentos formados, ao considerar essa variável, é mais representativa ao assumir k igual a 2 do que para as demais formações. Em outras palavras, a formação com maior valor de F indica uma solução com maior qualidade.

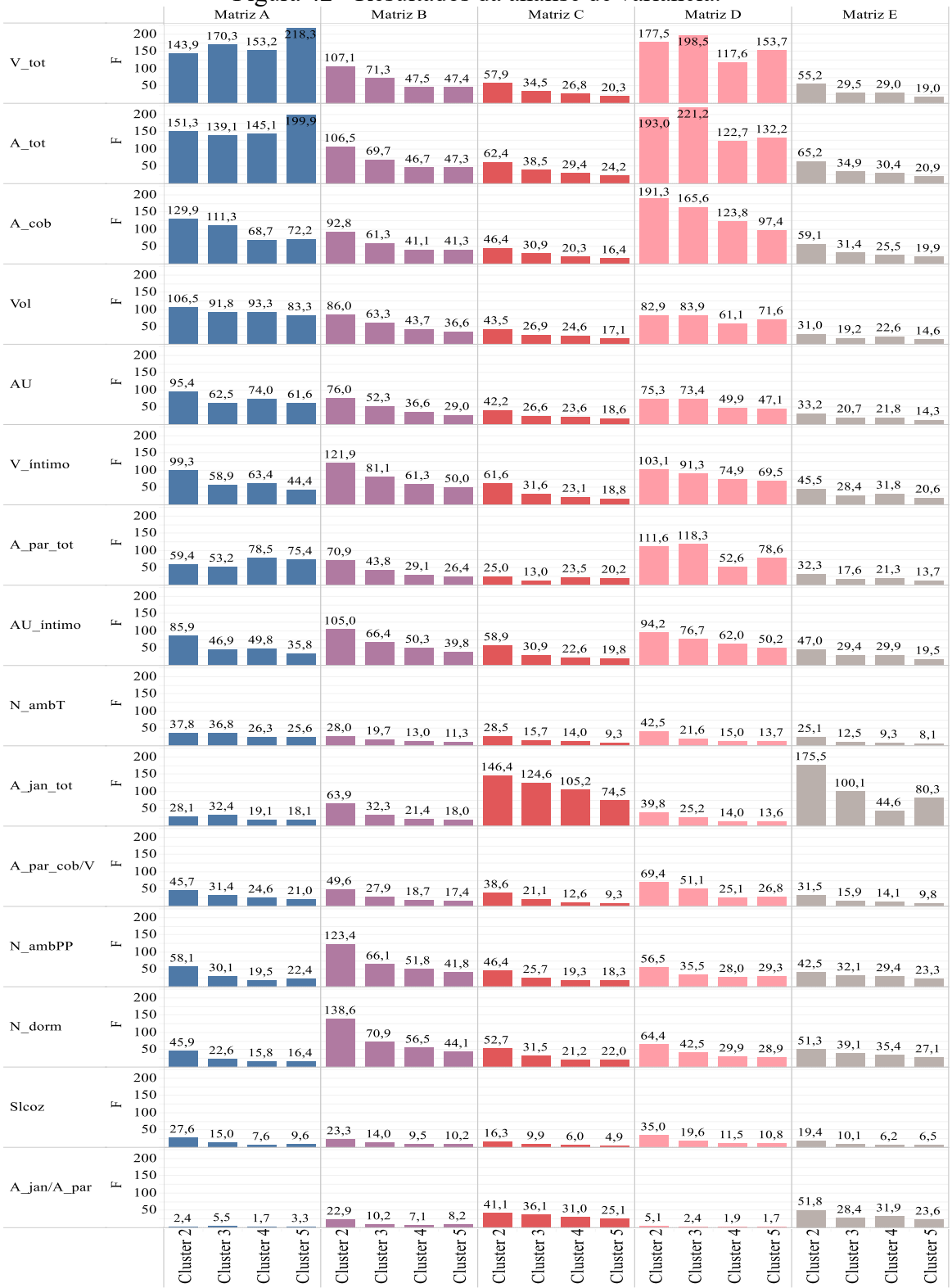
Além do valor de F, a significância estatística deve ser observada. A significância (ou p_{valor}) representa se a variabilidade entre os objetos de uma amostra, para determinada variável, ao longo dos grupos, é estatisticamente significativa para a diferenciação dos grupos. Nesse estudo, quando a significância é maior que 0,05, significa que para aquela formação, a variável analisada não teve impacto na formação dos grupos, ou seja, não há diferença significativa entre os objetos de cada grupo. É o caso, por exemplo, da razão entre a área de janela e a área de parede, na Matriz A. Todas as demais variáveis apresentadas na Figura 42 apresentaram p_{valor} menor que 0,05.

A relação completa das variáveis, o valor de F e a significância estatística está disponibilizada no Apêndice C.

Como pode ser visto na Figura 42, para a maioria das variáveis, a formação k igual a 2 foi a que apresentou o maiores valores de F. Essa solução fica evidente para as formações a partir das Matrizes B, C e E. A Matriz A apresentou maior quantidade de variáveis com F maior para a formação k igual a 2, enquanto as demais formações alternaram-se quanto qual apresentava maior valor de F. A Matriz D foi a matriz de dados para a qual foi mais difícil

identificar a melhor solução quanto à divisão dos agrupamentos, sendo k igual a 2 e k igual a 3 similarmente relevantes.

Figura 42 - Resultados da análise de variância.



4.3.3. Conjunto final de soluções

Após a interpretação dos resultados da análise de *cluster*, foi possível definir um conjunto de soluções final quanto aos métodos que apresentaram resultados válidos e quanto à quantidade de grupos ideal para cada método. Os Quadros 7 a 11 apresentam o conjunto de soluções obtidos para os métodos de técnicas hierárquicas, para cada matriz. Na primeira coluna estão listados os métodos resultantes do processo de seleção inicial. Na segunda coluna, apresenta-se as soluções possíveis a partir da leitura do dendograma. Na terceira coluna, são apresentadas as soluções obtidas a partir da análise do aumento percentual do grau de heterogeneidade com os coeficientes de aglomeração. Por fim, é apresentada na última coluna a decisão quanto à quantidade de grupos. Ao todo, foram selecionadas 60 soluções de agrupamentos.

Quadro 7 - Decisão quanto à quantidade de grupos para cada método da Matriz A.

Método	**Dendograma**	**Coeficiente de aglomeração**	**Decisão quanto à quantidade de grupos**
M6_A	2, 3 ou 4 grupos	4 grupos	k=4
M7_A	2 grupos	2 grupos	k=2
M8_A	2 grupos	2 grupos	k=2
M9_A	2, 3 ou 4 grupos	4 grupos	k=4
M15_A	3 grupos	2 grupos	k=2 e k=3
M16_A	2 ou 3 grupos	2 ou 3 grupos	k=2 e k=3
M17_A	2 ou 3 grupos	2 grupos	k=2
M18_A	3 grupos	3 grupos	k=3
M19_A	2, 3 ou 5 grupos	2 ou 3 grupos	k=2 e k=3

Quadro 8 - Decisão quanto à quantidade de grupos para cada método da Matriz B.

Método	**Dendograma**	**Coeficiente de aglomeração**	**Decisão quanto à quantidade de grupos**
M6_B	2 grupos	2 grupos	k=2
M7_B	2 grupos	2 grupos	k=2
M8_B	2 grupos	2 grupos	k=2
M9_B	2 ou 4 grupos	2 grupos	k=2
M10_B	4 grupos	2 ou 4 grupos	k=4
M10_B	4 grupos	2 ou 4 grupos	k=4

Quadro 8 - Decisão quanto à quantidade de grupos para cada método da Matriz B (continuação).

Método	Dendograma	Coeficiente de aglomeração	Decisão quanto à quantidade de grupos
M16_B	2 ou 3 grupos	2 grupos	k=2
M17_B	2 ou 3 grupos	2 grupos	k=2
M18_B	2 ou 3 grupos	2 grupos	k=2
M19_B	3 grupos	2 grupos	k=2
M20_B	2 grupos	2 grupos	k=2

Quadro 9 - Decisão quanto à quantidade de grupos para cada método da Matriz C.

Método	Dendograma	Coeficiente de aglomeração	Decisão quanto à quantidade de grupos
M6_C	2 grupos	5 grupos	k=2 e k=5
M7_C	2 grupos	2 ou 4 grupos	k=2 e k=4
M8_C	3 grupos	3 grupos	k=3
M15_C	2 grupos	2 grupos	k=2
M16_C	2 grupos	2 grupos	k=2
M17_C	2 ou 4 grupos	4 grupos	k=4
M18_C	2 ou 4 grupos	2 grupos	k=2
M19_C	2 ou 4 grupos	2 grupos	k=2
M20_C	2 grupos	2 grupos	k=2

Quadro 10 - Decisão quanto à quantidade de grupos para cada método da Matriz D.

Método	Dendograma	Coeficiente de aglomeração	Decisão quanto à quantidade de grupos
M6_D	2, 3 ou 5 grupos	2 grupos	k=2
M7_D	2 ou 3 grupos	2 ou 3 grupos	k=2 e k=3
M8_D	2, 3 ou 5 grupos	2, 3 ou 5 grupos	k=2, k=3 e k=5
M9_D	2, 3 ou 4 grupos	4 grupos	k=4
M10_D	3 ou 4 grupos	2 grupos	k=2
M13_D	3 ou 5 grupos	3 grupos	k=3
M15_D	5 grupos	2 ou 5 grupos	k=5
M16_D	2, 3 ou 4 grupos	2 grupos	k=2
M17_D	2 ou 3 grupos	2 ou 3 grupos	k=2 e k=3
M18_D	2 grupos	2 grupos	k=2
M19_D	2 ou 3 grupos	2 ou 3 grupos	k=2 e k=3

Quadro 11 - Decisão quanto à quantidade de grupos para cada método da Matriz E.

Método	Dendograma	Coeficiente de aglomeração	Decisão quanto à quantidade de grupos
M6_E	2 grupos	2 grupos	k=2
M7_E	2 grupos	2 grupos	k=2
M8_E	2 grupos	2 ou 3 grupos	k=2
M10_E	2 ou 4 grupos	2, 4 ou 5 grupos	k=2 e k=4
M15_E	2 grupos	2 grupos	k=2
M16_E	2 ou 3 grupos	2 grupos	k=2
M17_E	2 ou 4 grupos	2 grupos	k=2
M18_E	2 grupos	2 grupos	k=2
M19_E	2ou 4 grupos	2 grupos	k=2

Com o histórico de interação e a análise de variância, foi possível definir um conjunto de soluções a partir da análise de *cluster* usando o algoritmo K-means. O Quadro 12 apresenta a decisão quanto à quantidade de grupos para cada método. Na primeira coluna, estão listados os métodos analisados. Na segunda e terceira colunas, apresenta-se a solução quanto ao histórico de interação e ANOVA. Na última coluna, apresenta-se a decisão final quanto à quantidade de agrupamentos. Foram selecionadas seis soluções de agrupamentos.

Quadro 12 - Decisão quanto à quantidade de grupos para cada método.

Método	Histórico de interação	ANOVA	Decisão final quanto à quantidade de grupos
M21_A	2, 3 ou 5 grupos	2 grupos	k=2
M21_B	2, 3, 4 ou 5 grupos	2 grupos	k=2
M21_C	2, 3 ou 4 grupos	2 grupos	k=2
M21_D	2, 3 ou 4 grupos	2 ou 3 grupos	k=2 e k=3
M21_E	2 ou 4 grupos	2 grupos	k=2

4.4 Determinação dos modelos de referência

O Quadro 13 apresenta os objetos designados como modelos para todos os 60 métodos de agrupamento resultantes da seção 4.3. Nas colunas à esquerda, são apresentados os modelos encontrados, enquanto na coluna à direita, os métodos para os quais foram obtidos tais modelos.

Observa-se que a variedade de modelos encontrados é grande, e diferenciam-se por matriz de dados, medidas de similaridade e algoritmos de partição, ou seja, não é possível determinar um padrão único para a obtenção de tais modelos. Esse fato evidencia a necessidade de, ao aplicar a análise de agrupamentos, testar a combinação entre os diferentes parâmetros antes de definir o método final.

Os modelos de referência de maior ocorrência foram os objetos Hab_052 e a Hab_058, ocorrendo em dez métodos. Esses modelos foram resultantes da combinação das matrizes C e E e dos algoritmos City-block, distância Euclidiana e distância Euclidiana Quadrada. O par de modelos voltou a ocorrer em formação com quatro agrupamentos, para as mesmas matrizes e medidas de similaridade. Outros três objetos de maior ocorrência que fizeram par com o objeto Hab_052 foram os objetos Hab_012, Hab_119 e Hab_088, os dois primeiros ocorrendo principalmente com o algoritmo Ward e o último, com o algoritmo Ligação Simples. A combinação entre esses modelos também voltou a ocorrer em agrupamentos com três, quatro e cinco grupos.

A presença de objetos atípicos como modelo de referência também pode ser observada, como nos métodos M7_C e M6_C (Hab_092). Essa é uma séria consequência de não utilizar medidas para detecção de objetos atípicos, especialmente quando combinadas com o algoritmo Ligação Completa. Por formar grupos a partir das medidas entre os objetos mais distantes, objetos atípicos, que normalmente estão mais afastados da amostra, acabam formando grupos e agregando outros objetos que não se assemelham.

Quadro 13 – Modelos de referência obtidos para todos os métodos resultantes da análise de agrupamentos.

Modelos					
***Cluster* 1**	***Cluster* 2**	***Cluster* 3**	***Cluster* 4**	***Cluster* 5**	**Método**
Hab_052	Hab_058	-	-	-	M16_C, M17_C, M18_C, M6_E, M7_E, M8_E, M15_E, M16_E, M18_E, M21_E
Hab_052	Hab_012	-	-	-	M18_D, M21_D, M10_E, M17_E
Hab_052	Hab_119	-	-	-	M16_B, M18_B, M20_B, M21_B
Hab_052	Hab_088	-	-	-	M6_B, M7_B, M8_B, M17_B
Hab_010	Hab_086	-	-	-	M16_A, M17_A, M19_A, M21_A
Hab_052	Hab_058	Hab_048	Hab_092	-	M7_C, M8_C

Quadro 13 – Modelos de referência obtidos para todos os métodos resultantes da análise de agrupamentos (continuação).

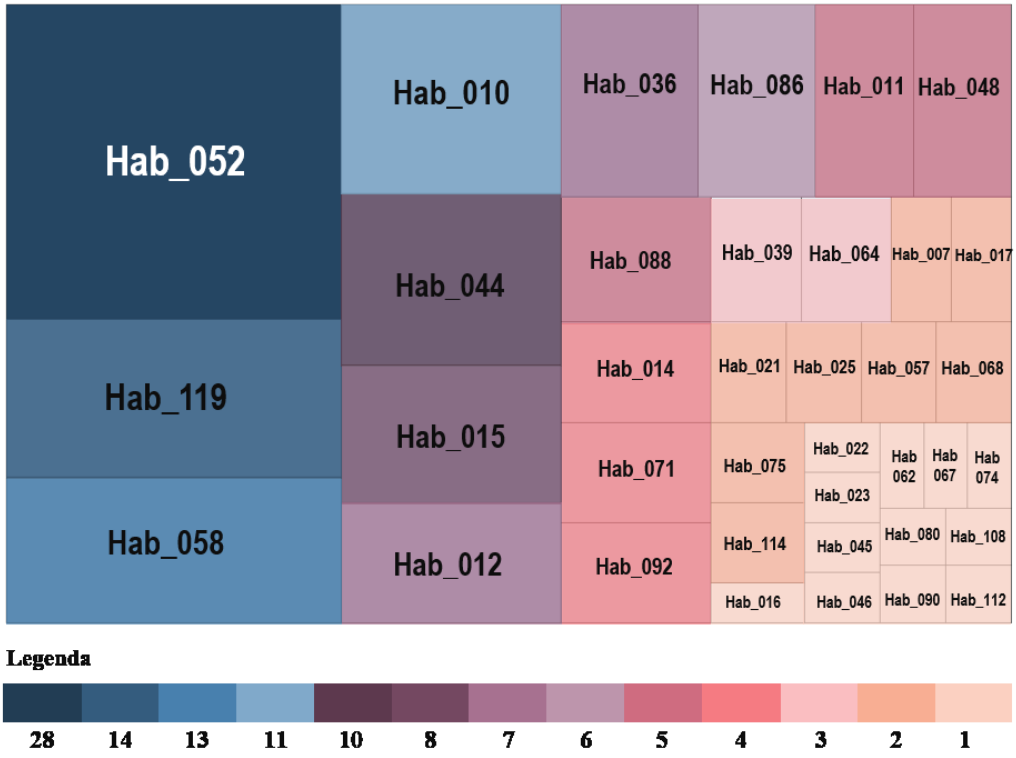
Modelos					
***Cluster* 1**	***Cluster* 2**	***Cluster* 3**	***Cluster* 4**	***Cluster* 5**	**Método**
Hab_119	Hab_044	Hab_015	-	-	M19_D, M21_D
Hab_119	Hab_007	-	-	-	M6_D, M19_D
Hab_044	Hab_036	Hab_039	-	-	M7_D, M8_D
Hab_036	Hab_025				M7_D, M8_D
Hab_036	Hab_057				M7_A, M8_A
Hab_052	Hab_119	Hab_023	-	-	M17_B
Hab_052	Hab_119	Hab_088	Hab_108	-	M10_B
Hab_052	Hab_012	Hab_048	Hab_092	Hab_021	M6_C
Hab_052	Hab_064	-	-	-	M21_C
Hab_119	Hab_010	Hab_017	-	-	M15_A
Hab_119	Hab_044	Hab_071	Hab_039	Hab_068	M8_D
Hab_119	Hab_015	Hab_011	Hab_016	-	M6_A
Hab_119	Hab_011	Hab_64	Hab_075	-	M10_E
Hab_058	Hab_010	-	-	-	M19_E
Hab_010	Hab_014	-	-	-	M15_A
Hab_010	Hab_071	-	-	-	M9_B
Hab_010	Hab_064	-	-	-	M19_C
Hab_010	Hab_090	-	-	-	M19_B
Hab_010	Hab_017	Hab_67	Hab_074	Hab_080	M15_D
Hab_044	Hab_015	Hab_036	-	-	M17_D
Hab_044	Hab_015	Hab_086	-	-	M19_A
Hab_044	Hab_012	-	-	-	M16_D
Hab_044	Hab_014	-	-	-	M17_D
Hab_044	Hab_071	Hab_062	Hab_112	-	M9_D
Hab_015	Hab_086	Hab_011	-	-	M16_A
Hab_015	Hab_048	-	-	-	M6_C
Hab_015	Hab_092	-	-	-	M7_C
Hab_012	Hab_075	-	-	-	M20_C
Hab_011	Hab_014	-	-	-	M6_A
Hab_011	Hab_114	Hab_045	-	-	M18_A
Hab_048	Hab_114	-	-	-	M9_A
Hab_014	Hab_021	Hab_022	-	-	M10_D
Hab_071	Hab_068	Hab_046	-	-	M13_D

A Figura 43 apresenta um mapa em árvore referente à frequência de ocorrência dos modelos. Mapas em árvore oferecem uma forma relativamente simples de visualizar dados hierarquicamente estruturados. Cada retângulo representa uma habitação, enquanto os tamanhos e cores representam a quantidade de vezes que essa habitação foi selecionada como modelo de referência de um método de agrupamento (indicada na legenda, abaixo do quadro).

O objeto com maior ocorrência foi a Hab_052, como visto também no Quadro 13Quadro 13. Essa habitação foi selecionada como modelo em 28 dos 60 métodos propostos, ou seja, quase metade do montante. Há, portanto, forte evidência de que esta é de fato uma edificação representativa da amostra. Essa mesma habitação foi determinada como modelo de referência no estudo realizado por Schaefer e Ghisi (2016), que utilizou o mesmo banco de dados.

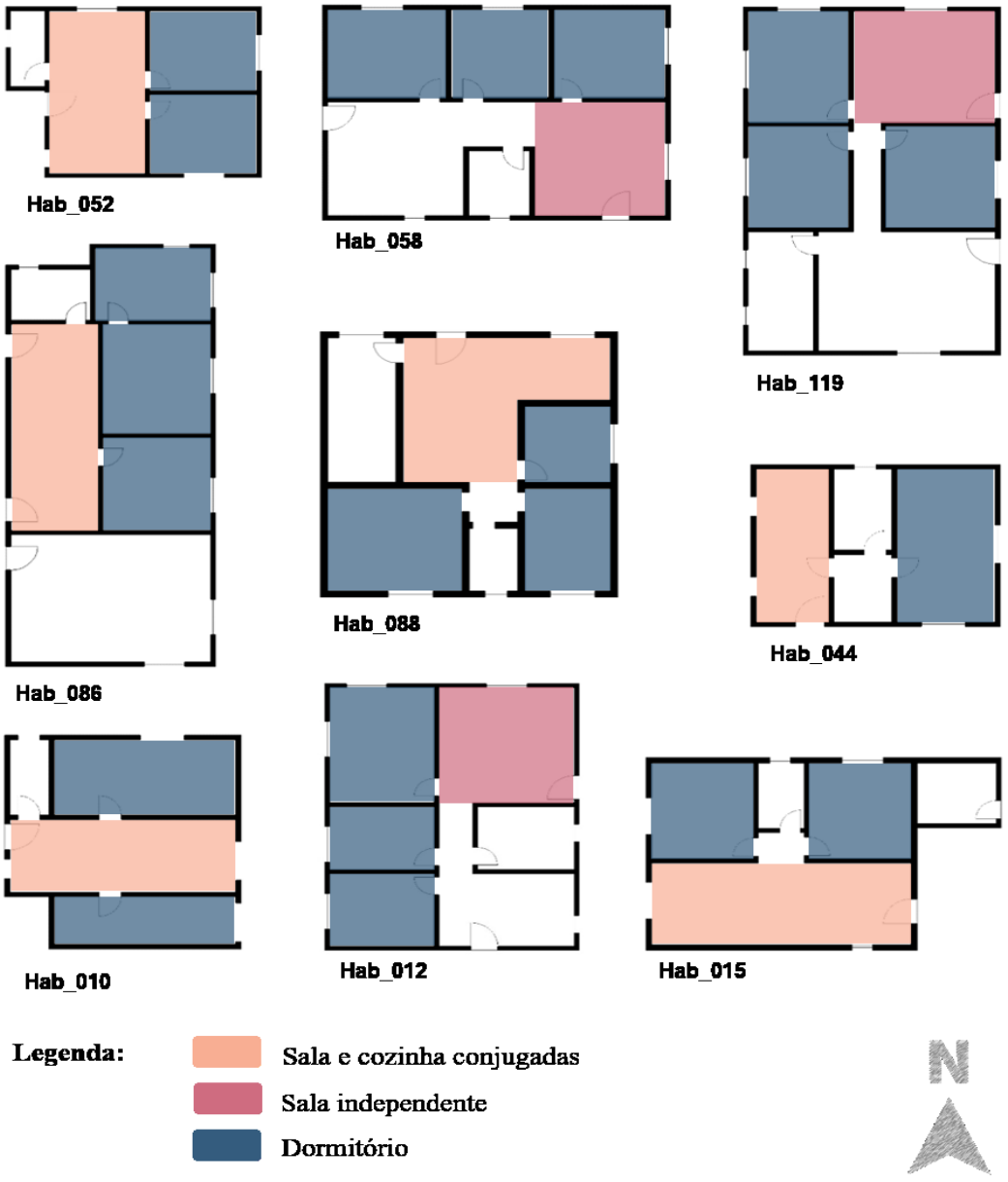
A segunda maior ocorrência foi o objeto Hab_119, seguida do objeto Hab_058, ocorrendo em 14 e 13 métodos, respectivamente. Nota-se que esses dois objetos formaram o conjunto de pares mais frequentes com o objeto Hab_052, como apresentado no Quadro 13. O objteo Hab_012 teve frequência de ocorrência um pouco abaixo (sete vezes), mas formou o segundo par mais frequente com o objeto Hab_052. Essas observações indicam forte tendência desses modelos serem resultantes da melhor solução de agrupamento.

Figura 43 – Mapa em árvore da frequência de ocorrência dos modelos.



Na Figura 44 encontra-se a representação gráfica em planta baixa dos modelos de maior ocorrência, indicados no Quadro 13 e na Figura 43. De forma geral, os modelos diferenciam-se pelas suas dimensões e pela orientação solar dos seus ambientes. Por exemplo, quanto ao par de modelos Hab_052 e Hab_058, o primeiro é bem menor que o segundo, e possui dormitórios orientados a leste e sala com orientação solar oeste. O segundo possui dormitórios orientados a norte e sala com orientação sudeste. Os dois modelos também se diferenciam pela quantidade de dormitórios e existência de sala e cozinha conjugadas no modelo Hab_052. Com essas informações, é possível criar expectativas sobre o desempenho diferente a partir de cada modelo.

Figura 44 - Representação gráfica em planta baixa dos modelos de maior ocorrência.



4.5. Validação dos métodos

A validação correspondeu à etapa em que a qualidade da formação dos grupos foi avaliada. A validação interna permitiu o estabelecimento de critérios mínimos de qualidade, sendo possível selecionar quais métodos atendiam a esses critérios e eliminar os demais. A validação relativa permitiu avaliar a capacidade de cada método em selecionar modelos capazes de representar o desempenho térmico dos demais objetos do grupo, auxiliando no processo de definição do melhor método de clusterização para essa amostra.

4.5.1. Validação: medida interna

A Tabela 16 apresenta os valores de inércia global, inércia *inter-cluster* e inércia *intra-cluster* para vinte dos métodos resultantes da seção 4.3 que atenderam os critérios estabelecidos na seção 3.5.1. Ao todo, resultaram do processo 36 métodos de clusterização.

A inércia global representa o grau de dispersão dos dados e permanece constante independente da forma como seus dados foram agrupados, se aplicada a uma mesma base de dados. Na Tabela 16, há dois valores encontrados para essa variável. Essa diferença ocorre devido às Matrizes B, D e E terem sido submetidas à detecção de objetos atípicos e estes, excluídos da amostra. O aumento de 30% no valor da inércia global nas Matrizes A e C evidencia a presença de objetos atípicos e o impacto na dispersão de seus dados, ou seja, estende os limites da área ocupada pela nuvem de dados. Esse comportamento tem efeito negativo, pois vai influenciar diretamente a mudança do centro de gravidade da nuvem. Como consequência, desestabiliza as relações de proximidade entre os objetos e podem também levar a modelos que não representam bem o grupo. Por esse motivo, pode-se observar que os agrupamentos formados por essas matrizes e que foram validados nessa etapa foram aqueles que dividiram a amostra em maior quantidade de grupos.

A inércia global representa a soma das inércias *inter* e *intra-cluster*. A inércia *intra-cluster* tende a diminuir à medida que a quantidade de grupos aumenta, enquanto a inércia *inter-cluster* tende a aumentar, pois os agrupamentos ficam mais homogêneos internamente e distantes um do outro. Por isso, pode-se observar na Tabela 16 que os melhores índices de qualidade Q foram obtidos para soluções com maior quantidade de grupos. Pelo mesmo motivo, os valores de inércia *inter-cluster* tendem a ser maiores que os valores de inércia *intra-cluster* em boas formações de agrupamento. Os algoritmos Ligação Completa e Ward apresentaram

boa diferenciação quanto aos valores de inércia *inter-cluster* e inércia *intra-cluster*, tendo a inércia *inter-cluster* assumido quase o dobro do valor da inércia *intra-cluster* na maioria dos métodos. O algoritmo K-means, por sua vez, apresentou esses valores bem equilibrados, o que se deve ao fato desse algoritmo tender a formar grupos com quantidades similares de objetos.

Na última coluna, são apresentados em ordem decrescente os valores de Q, medida de qualidade da clusterização. O valor de Q foi obtido a partir da razão entre inércia *inter-cluster* e inércia *intra-cluster*, ponderada pela quantidade de objetos em cada grupo. A ponderação se fez importante para evitar, por exemplo, que métodos com grupos muito pequenos, que naturalmente seriam muito homogêneos, tivessem o mesmo peso que os demais. O resultado seria um agrupamento não representativo da amostra exercendo grande influência no valor da medida de qualidade Q. Os melhores índices de Q foram obtidos com os métodos que utilizaram a Matriz D. A Tabela completa está disponibilizada no Apêndice D.

Tabela 16 – Inércia global, inércia *inter-cluster* e inércia *intra-cluster* por método.

Método	**k**	**Inércia Global**	**Inércia *Inter-cluster***	**Inércia *Intra-cluster***	**Q**
M09_D	4	557.661	376.633	181.133	2,14
M17_D	3	557.661	365.089	192.572	2,13
M07_D	3	557.661	362.428	195.232	2,11
M08_D	3	557.661	362.428	195.232	2,11
M08_D	5	557.661	408.697	149.069	2,09
M06_A	4	735.832	284.866	450.966	2,04
M16_A	3	735.832	433.347	302.486	2,04
M19_D	3	557.661	366.268	191.393	2,03
M18_A	3	735.832	453.298	282.534	2,00
M19_A	3	735.832	429.875	305.957	1,91
M09_A	4	735.832	449.514	286.319	1,79
M16_D	2	557.661	257.111	300.550	1,67
M17_D	2	557.661	251.962	305.698	1,66
M10_E	4	557.661	273.841	283.819	1,33
M18_D	2	557.661	272.247	285.414	1,26
M17_C	4	735.832	278.714	457.119	1,22
M06_B	2	557.661	245.307	312.354	1,14
M21_D	2	557.661	288.163	269.498	1,13
M17_B	2	557.661	248.223	309.438	1,01
M17_B	3	557.661	248.223	309.438	1,01

4.5.2. Validação: medida relativa

A validação relativa buscou avaliar a qualidade do agrupamento a partir de variáveis que descrevam o desempenho térmico dos seus objetos e, em especial, do modelo de referência. Para alcançar esse objetivo, seis índices estatísticos para a quantificação de erros amostrais foram aplicados aos 36 métodos resultantes da seção 4.5.1. Os resultados estão apresentados na Tabela 17.

A primeira e segunda coluna apresentam os métodos e quantidade de agrupamentos. A última coluna, o índice de erro global. As colunas intermediárias apresentam os valores normalizados de cada índice estatístico, para cada método. Os valores precisaram ser normalizados de forma que as diferentes amplitudes de cada índice não influenciassem o agrupamento desses valores em um índice global.

O método que obteve o melhor índice foi o Método M21_D, cuja configuração combinou a Matriz D e o algoritmo K-means. Esse método é frequentemente citado na literatura como um bom método de clusterização, especialmente por sua característica não-hierárquica, ou seja, permite a redesignação dos objetos. Ao comparar os valores encontrados para cada índice nesse método, observa-se que todos eles representam valores baixos se comparados com os dos demais métodos.

A necessidade de utilizar índices diferentes para validar o método fica evidente com os dados apresentados na Tabela 17. Por exemplo, os índices erro máximo (ME) e erro médio absoluto (EAM) do método M21_D ficaram acima daqueles obtidos pelo método M17_B. Entretanto, o índice de concordância (d) deste representa mais que o triplo do valor de M21_D. Destaca-se aqui que os índices erro máximo (ME) e erro médio absoluto (EAM) são índices que avaliam o desempenho de todo o grupo, enquanto o índice de concordância (d) compara também os desvios em relação ao modelo. A lista completa dos índices obtidos encontra-se no Apêndice E.

Tabela 17 – Índices estatísticos ponderados para quantificação de erros por método.

Método	**K**	**d**	**ME**	**EAM**	**RMSE**	**EF**	**CRM**	**E_{global}**
M21_D	2	0,165	0,337	0,230	0,248	0,138	0,126	1,24
M18_D	2	0,395	0,148	0,136	0,159	0,179	0,386	1,40
M17_B	3	0,604	0,170	0,167	0,166	0,318	0,509	1,44
M10_B	4	0,336	0,192	0,140	0,154	0,273	0,349	1,44
M10_E	4	0,378	0,162	0,127	0,293	0,073	0,373	1,60

Tabela 17 – Índices estatísticos ponderados para quantificação de erros por método (continuação).

Método	k	D	ME	EAM	RMSE	EF	CRM	E_{global}
M07_D	2	0,557	0,162	0,237	0,216	0,242	0,197	1,61
M08_D	2	0,557	0,162	0,237	0,216	0,242	0,197	1,61
M21_B	2	0,402	0,290	0,247	0,274	0,221	0,344	1,77
M18_B	2	0,362	0,320	0,250	0,288	0,196	0,374	1,79
M06_B	2	0,589	0,128	0,122	0,122	0,360	0,496	1,82
M16_B	2	0,396	0,320	0,255	0,290	0,186	0,371	1,82
(Valores intermediários omitidos)								
M09_D	4	0,409	0,535	0,639	0,659	0,537	0,008	2,78
M17_D	3	0,113	0,621	0,690	0,733	0,582	0,230	2,96
M16_A	3	0,500	0,701	0,454	0,594	0,205	0,544	2,99
M17_C	4	0,546	0,546	0,434	0,519	0,451	0,538	3,03
M17_D	2	0,690	0,512	0,596	0,610	0,355	0,677	3,44
M08_D	5	0,349	0,720	0,699	0,760	0,652	0,518	3,69
M07_D	3	0,191	0,784	0,770	0,846	0,753	0,720	4,06
M08_D	3	0,191	0,784	0,770	0,846	0,753	0,720	4,06
M18_A	3	0,474	0,959	0,793	0,930	0,886	0,135	4,17
M09_A	4	0,504	1,001	0,813	0,961	1,000	0,028	4,30
M16_D	2	0,336	0,870	0,953	0,998	0,985	0,213	4,35

4.6. Caracterização do objeto final

A caracterização do objeto final corresponde à apresentação do produto final do método proposto, que é a obtenção de grupos distintos quanto às características geométricas e desempenho térmico das habitações que os compõem e a determinação de modelos de referência que os represente.

4.6.1. Perfis a partir da geometria

Primeiramente, foi realizada a caracterização do melhor método a partir das suas características geométricas, que foram os dados utilizados no processo de clusterização. O método identificado como melhor para esse estudo foi o método M21_D, que segregou a amostra em dois grupos. Os objetos determinados como modelos de referência foram a Hab_012 e a Hab_052.

As Tabelas 18 e 19 apresentam algumas das características descritivas de cada agrupamento, juntamente com as características dos modelos que os representam, para todas as variáveis envolvidas na análise. As variáveis qualitativas foram apresentadas na Tabela 18 a partir da porcentagem de ocorrência de habitações para cada categoria. As variáveis quantitativas, por sua vez, foram apresentadas em termos de média e desvio padrão, para cada agrupamento. A caracterização completa dos grupos está apresentada no Apêndice F.

O método do modelo de referência adotado nesse estudo baseia-se no conceito de edifício real, ou seja, trata-se da edificação mais próxima à média do grupo (e não corresponde à média, propriamente). Ao analisar os valores apresentados nas Tabelas 18 e 19, verifica-se que os modelos adotados para esse método representam bem a média do grupo. Para as variáveis qualitativas, o modelo assumiu valor correspondente à categoria de maior ocorrência. Por exemplo, quanto à quantidade de dormitórios, o modelo de referência 1 possui três, que corresponde a 57% dos casos da amostra. O modelo de referência do agrupamento 2 possui dois dormitórios, que corresponde a 64% dos casos da sua amostra de habitações.

Quanto às variáveis quantitativas, verificou-se que o valor do modelo estava dentro do intervalo de até um desvio padrão (acima ou abaixo) de todas as variáveis. Dessa forma, conclui-se que os modelos representaram bem seus grupos.

De forma geral, os modelos mostraram-se bastante satisfatórios. O modelo 1 ficou definido por uma habitação de 64m^2, com sala e cozinha independentes e três dormitórios. O modelo 2 possui sala e cozinha integradas e dois dormitórios. Os modelos diferiram em características importantes da geometria, que influenciam no desempenho térmico, como as suas dimensões e orientação solar dos ambientes de permanência prolongada. O modelo 2 (Hab_052) também foi obtido como modelo a partir do estudo realizado por Schaefer e Ghisi (2016) e aproxima-se bastante do modelo encontrado por Triana (2015).

A representação gráfica dos dois modelos está apresentada nas Figuras 45-47.

Tabela 18 – Perfil dos agrupamentos a partir das variáveis qualitativas.

Variável		**Agrupamento**	
		1	2
Existência de sala e cozinha conjugadas	Sim	20%	72%
	Não	80%	28%
	Modelo de referência	**Não**	**Sim**
Quantidade de dormitórios	1	4%	32%
	2	27%	64%
	3	57%	4%
	4 ou mais	12%	0%
	Modelo de referência	**3**	**2**

Tabela 19 – Perfil dos agrupamentos a partir das variáveis quantitativas.

Variável		Agrupamento	
		1	2
Área total (m^2)	Média	68,98	39,02
	Desvio padrão	9,33	7,90
	Modelo de referência	**64,04**	**37,08**
Área útil dos ambientes sociais (m^2)	Média	17,47	15,51
	Desvio padrão	4,22	4,50
	Modelo de referência	**16,32**	**16,41**
Área útil dos ambientes íntimos (m^2)	Média	28,74	15,57
	Desvio padrão	5,68	4,75
	Modelo de referência	**28,12**	**17,68**
Volume total (m^3)	Média	174,77	96,16
	Desvio padrão	25,14	18,62
	Modelo de referência	**172,90**	**89,00**
Razão entre área de parede e janela dos ambientes de permanência prolongada	Média	9,30	8,05
	Desvio padrão	2,32	1,72
	Modelo de referência	**9,51**	**7,93**

Figura 45 – Perspectivas da maquete eletrônica utilizada nas simulações: (a) Vista leste e norte do modelo de referência 1. (b) Vista leste e norte do modelo de referência 2. (c) Vista oeste e sul do modelo de referência 1. (d) Vista oeste e sul do modelo de referência 2.

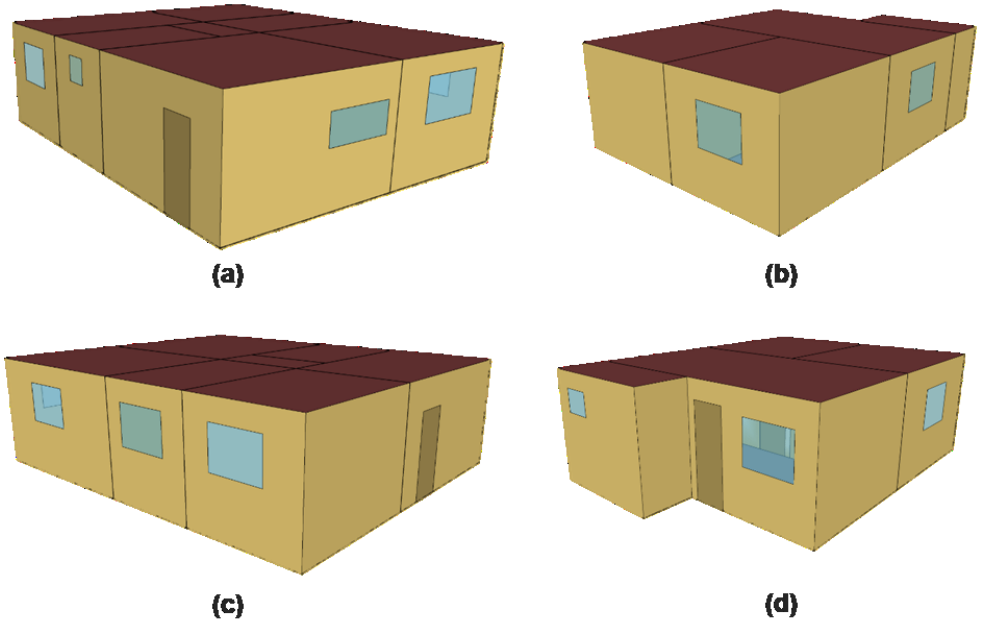


Figura 46 – Modelo de referência do agrupamento 1.

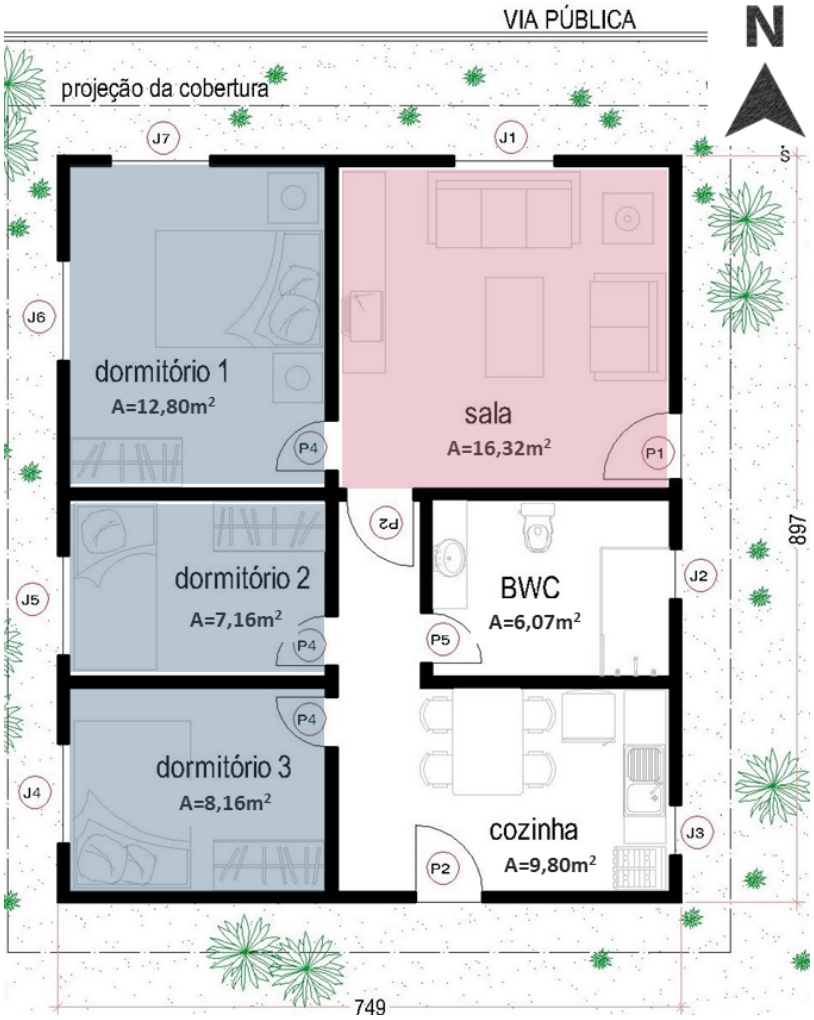
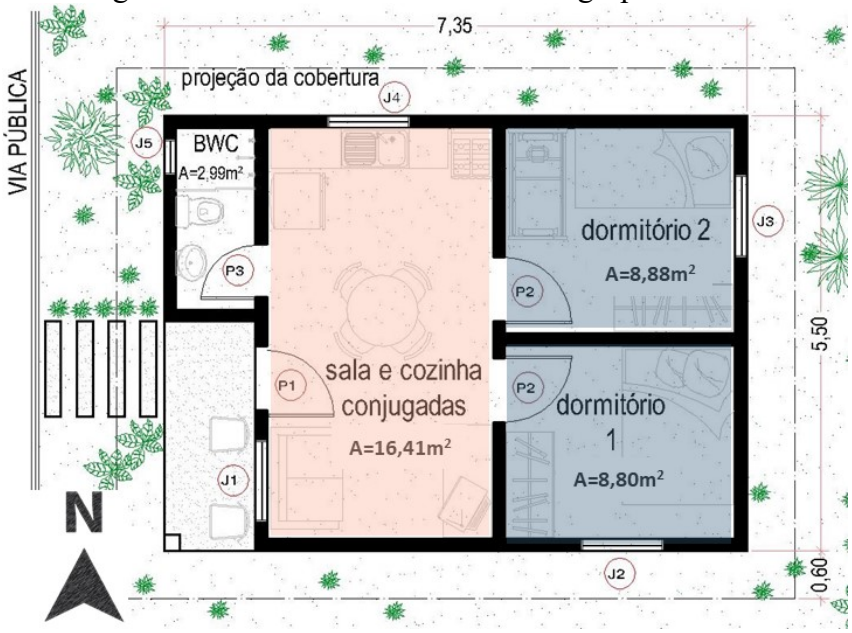


Figura 47 – Modelo de referência do agrupamento 2.



4.6.2. Perfis a partir do desempenho térmico

A principal função do modelo de referência é a aplicabilidade deste em estudos de desempenho térmico de edificações similares. Por esse motivo, além da geometria, é importante que os agrupamentos formados apresentem características diferentes quanto ao seu desempenho térmico e que os modelos encontrados representem de forma aproximada as características do seu grupo.

As Figuras 48 e 49 mostram o desempenho para aquecimento e resfriamento anual de cada modelo e a Tabela 20 os valores de média, mediana e erro padrão de cada grupo, para cada configuração de simulação e ambiente.

Observa-se que os dois modelos obtiveram desempenho para resfriamento inferior ao aquecimento, o que já era esperado para o clima de Florianópolis e a partir do que foi apresentado na Figura 31. Observa-se também que o modelo 1 apresentou desempenho pior para resfriamento quando comparado ao modelo 2, enquanto este teve desempenho inferior para aquecimento.

Com as Figuras 48 e 49 também se confirma a expectativa dos indicadores de desempenho dos modelos estarem no intervalo de maior concentração de casos visto na Figura 31 (resumo dos indicadores de desempenho da amostra). Este é um indicador de que os modelos representam bem a amostra. O mesmo pode ser observado para os valores de média e mediana na Tabela 20.

Figura 48 – Indicadores de desempenho do Modelo 1 para aquecimento e resfriamento (ºCh).

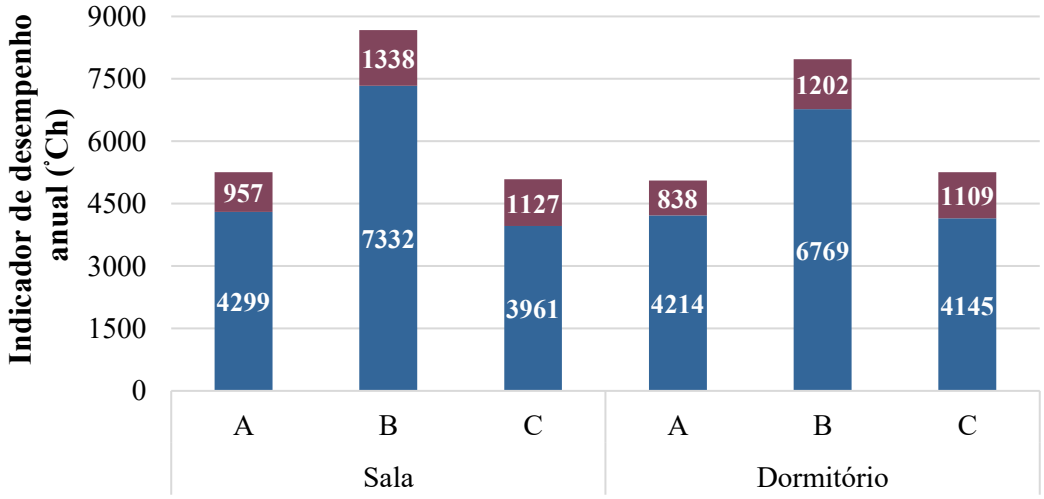


Figura 49 – Indicadores de desempenho do Modelo 2 para aquecimento e resfriamento (ºCh).

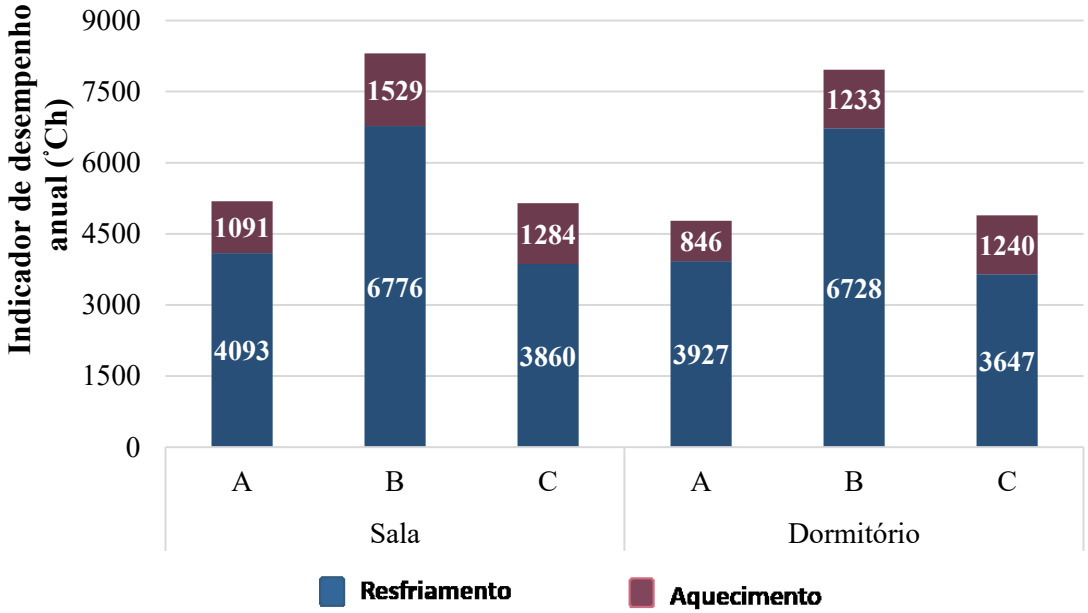


Tabela 20 – Valores de média, mediana e erro padrão dos grupos para cada indicador de desempenho.

Variável	Agrupamento 1 (n=49)			Agrupamento 2 (n=47)		
	Mediana (ºCh)	Média (ºCh)	Erro (ºCh)	Mediana (ºCh)	Média (ºCh)	Erro (ºCh)
ID_R_dorm_A	4.152	4.158	38	4.039	4.034	50
ID_R_dorm_B	6.936	6.928	51	3.919	3.891	57
ID_R_dorm_C	3.947	3.966	39	3.919	3.891	57
ID_A_dorm_A	886	873	19	907	932	29
ID_A_dorm_B	1.195	1.200	22	1.346	1.343	38
ID_A_dorm_C	1.150	1.147	19	1.260	1.270	24
ID_R_sala_A	4.029	4.049	58	4.061	4.037	75
ID_R_sala_B	6.577	6.612	95	6.828	6.655	132
ID_R_sala_C	3.765	3.766	56	3.855	3.784	3860
ID_A_sala_A	954	969	25	1135	1083	35
ID_A_sala_B	1.293	1.274	36	1.565	1.502	51
ID_A_sala_C	1.127	1.149	24	1.316	1.261	32

A Figura 50 apresenta o diagrama de caixas referente aos indicadores de desempenho de todas as habitações da amostra, obtidos para cada agrupamento. Nas colunas estão apresentados os ambientes e configurações de simulação e, nas linhas, os indicadores de desempenho. Dentro de cada quadro, encontram-se os diagramas de caixa de cada agrupamento.

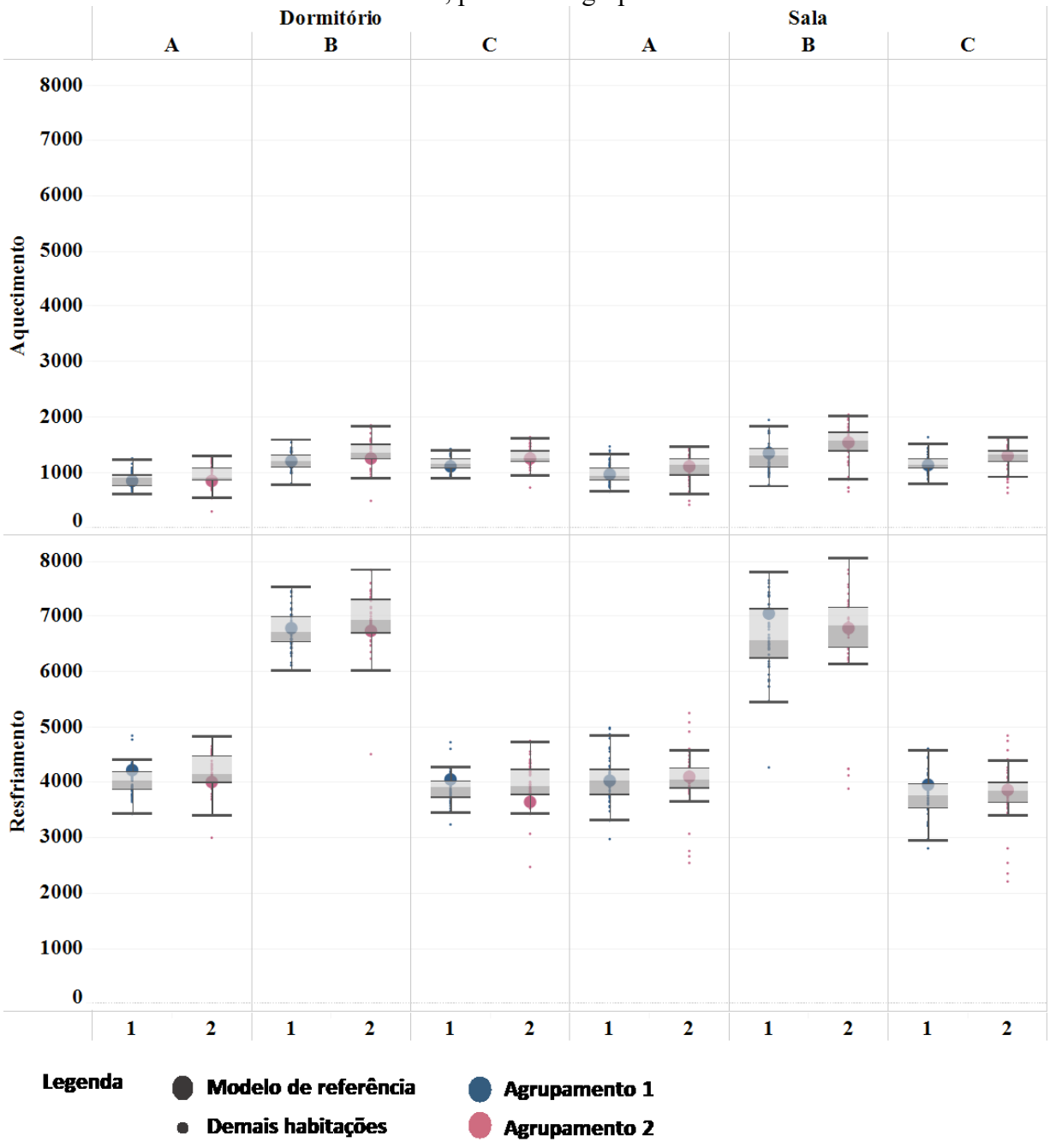
No diagrama de caixas, é possível analisar visualmente a distribuição amostral dos valores de desempenho e também a posição do modelo de referência na distribuição. Como já observado em outras figuras e tabelas anteriores, o desempenho quanto à condição de resfriamento é bastante inferior ao aquecimento, para todas as variáveis e configurações de simulação. Além dos valores, a distribuição dos dados é mais dispersa. A presença de objetos atípicos pode ser percebida pelos pontos que estão acima ou abaixo dos limites extremos de cada caixa. Esses pontos também foram observados na Figura 31.

Através do diagrama de caixas, foi analisada a posição do modelo em relação à distribuição amostral. Um bom indicativo de que o modelo representa bem a amostra é quando ele se encontra entre o primeiro e terceiro quartis, o que foi observado para a maioria das variáveis. Na Figura 50, a única variável para a qual não foi possível estabelecer essa relação foi o indicador de desempenho de resfriamento do dormitório para a configuração C, do agrupamento 2. Em algumas variáveis, o modelo se encontrou junto à extremidade dos quadrantes.

Outra observação que pode ser feita em um diagrama de caixas é comparar a posição das medianas na distribuição. A igualdade entre elas é melhor quantificada através de testes de hipóteses, mas já é possível obter um indicativo a partir da Figura 50. A mediana é representada pela linha horizontal localizada entre as linhas do primeiro e terceiro quadrante. Na Figura 50, essa diferença fica bem clara para a maioria das variáveis, em especial aquelas que foram obtidas a partir da configuração de simulação B. Configurações como a C de resfriamento da sala, entretanto, apresentam as medianas próximas e distribuição similar, indicando que para essa configuração as amostras não obtiveram muita diferenciação.

Os gráficos de caixa são interessantes pois conseguem resumir em uma só imagem muitas informações da amostra, como os valores encontrados, as variâncias, mediana e posição do modelo de referência na distribuição amostral. Entretanto, apresentam os dados de forma apenas descritiva. Para complementar a análise, as Tabela 22 e 23 mostram os resultados dos testes de hipótese, aplicados com o propósito de comprovar estatisticamente que os grupos formados se diferem. Por se tratar de teste inferencial, a interpretação dos seus resultados é mais confiável.

Figura 50 – Diagrama de caixas referente ao indicador de desempenho de todas as habitações da amostra, para cada agrupamento.



Foi testada a igualdade dos agrupamentos a partir da distribuição amostral (Tabela 21) e também a partir das medianas (Tabela 22), para nível de significância de 0,05. Os dois testes apresentaram resultados similares, com exceção do indicador de desempenho de resfriamento das salas, cuja hipótese de igualdade foi rejeitada pelo teste de medianas.

Observou-se que os testes de hipótese aplicados rejeitaram a hipótese de igualdade para a maioria dos indicadores de desempenho dos dormitórios, tanto para resfriamento quanto

para aquecimento. Esse comportamento era esperado devido à orientação solar dos ambientes: em um grupo, orientados a leste, em outro, a oeste.

Quanto às salas, foram rejeitadas as hipóteses de igualdade apenas para aquecimento, não havendo diferença significativa entre os grupos quanto ao resfriamento. Como esses ambientes também são voltados a orientações diferentes (nordeste em um grupo e oeste em outro), esse comportamento não era esperado, em um primeiro momento. Entretanto, ao analisar mais criticamente os modelos, pode-se supor que alguns fatores podem estar contribuindo para neutralizar os efeitos da orientação solar, como, por exemplo, a relação entre a área de fachada exposta à insolação e o volume interno da sala da Hab_052. Outros possíveis fatores seriam a existência do banheiro em uma das fachadas e a existência de ventilação cruzada, também na Hab_052.

Todos os indicadores de desempenho obtidos a partir da configuração B rejeitaram a hipótese de igualdade. Essa configuração possui sistemas construtivos com maior transmitância e menor capacidade térmicas. Tornam, portanto, a edificação mais sensível aos efeitos climáticos sobre a sua geometria, o que confere com resultado encontrado. Em outras palavras, nessa situação o desempenho térmico é muito mais influenciado pelas condicionantes da geometria da edificação.

Tabela 21 - Hipótese de igualdade quanto à distribuição de valores do indicador de desempenho ao longo das categorias de agrupamento através do teste U de Mann-Whitney (nível de significância $p<0,05$).

Ambiente	**Indicador de desempenho**	**Configuração de simulação**	**Distribuição dos valores de desempenho**	
			p_{valor}	Decisão
Dormitório	Resfriamento	A	0,029	Rejeita-se a hipótese de igualdade
		B	0,010	Rejeita-se a hipótese de igualdade
		C	0,150	Aceita-se a hipótese de igualdade
	Aquecimento	A	0,114	Aceita-se a hipótese de igualdade
		B	0,002	Rejeita-se a hipótese de igualdade
		C	<0,000	Rejeita-se a hipótese de igualdade
Sala	Resfriamento	A	0,517	Aceita-se a hipótese de igualdade
		B	0,021	Aceita-se a hipótese de igualdade

Tabela 21 - Hipótese de igualdade quanto à distribuição de valores do indicador de desempenho ao longo das categorias de agrupamento através do teste U de Mann-Whitney (nível de significância p<0,05) (continuação).

Ambiente	**Indicador de desempenho**	**Configuração de simulação**	**Distribuição dos valores de desempenho**	
			p_{valor}	Decisão
Sala	Resfriamento	C	0,220	Aceita-se a hipótese de igualdade
	Aquecimento	A	0,001	Rejeita-se a hipótese de igualdade
		B	<0,000	Rejeita-se a hipótese de igualdade
		C	0,001	Rejeita-se a hipótese de igualdade

Tabela 22 -Teste de Medianas para amostras independentes ao longo das categorias de agrupamento (nível de significância p<0,05).

Ambiente	**Indicador de desempenho**	**Configuração de simulação**	**Distribuição dos valores de desempenho**	
			p_{valor}	Decisão
Dormitório	Resfriamento	A	0,041	Rejeita-se a hipótese de igualdade
		B	0,014	Rejeita-se a hipótese de igualdade
		C	0,683	Aceita-se a hipótese de igualdade
	Aquecimento	A	0.683	Aceita-se a hipótese de igualdade
		B	0,001	Rejeita-se a hipótese de igualdade
		C	0,004	Rejeita-se a hipótese de igualdade
Sala	Resfriamento	A	1,000	Aceita-se a hipótese de igualdade
		B	0,041	Rejeita-se a hipótese de igualdade
		C	0,102	Aceita-se a hipótese de igualdade
	Aquecimento	A	0,001	Rejeita-se a hipótese de igualdade
		B	<0,000	Rejeita-se a hipótese de igualdade
		C	<0,000	Rejeita-se a hipótese de igualdade

Baseando-se nos resultados obtidos e nas considerações acima expostas, considera-se que o método M21_D conseguiu segregar a amostra em grupos distintos e que os modelos de referência obtidos são adequados para representar seus grupos.

5. Conclusão

O objetivo desse trabalho foi desenvolver um método de aplicação da análise de agrupamentos visando a obtenção de modelos de referência de edificações para uso em estudos de desempenho térmico, a partir de diferentes configurações de clusterização. Para alcançar esse objetivo, seis objetivos específicos foram traçados.

Primeiramente, um estudo de caso envolvendo uma amostra de habitações de interesse social da região de Florianópolis foi submetida ao método proposto. As habitações da amostra, nomeadas no estudo como objetos, foram descritas a partir das suas características geométricas e submetidas à análise de agrupamentos através de diferentes métodos de clusterização. Cada método de clusterização foi elaborado a partir da combinação de diferentes tratamentos de dados, medidas de similaridade e algoritmos de partição.

Nem todas as combinações formaram métodos para os quais foi possível obter formações adequadas de agrupamentos. Dos 125 métodos propostos, apenas 60 (ou seja, menos da metade) apresentaram soluções capazes de segregar a amostra em grupos robustos. Alguns deles não permitiram sequer a elaboração do dendograma (que registra o processo de formação dos grupos etapa por etapa) ou da identificação da regra de parada através da medida de heterogeneidade (calculada a partir do coeficiente de aglomeração). Quanto ao método não hierárquico, algumas formações não chegaram a convergência até a 10ª interação. Quando isso acontece, há um indicativo de que a divisão do grupo provocou certa instabilidade, ou seja, agrupamento de baixa qualidade.

Com a maioria dos métodos a amostra foi segregada em dois ou três grupos, mas formações com divisão de até cinco grupos foram identificadas. De forma geral, as formações diferiram bastante entre si (com exceção de algumas poucas combinações), seja quanto à quantidade de agrupamentos formados, quanto à quantidade de objetos por grupo e também a designação de cada objeto a cada grupo. Com esses resultados, foi possível concluir que a seleção de diferentes parâmetros pode resultar em formações bem diferentes, mesmo para uma mesma amostra. Dessa forma, não se indica a aplicação da análise de agrupamentos da forma como ela tem sido aplicada em boa parte das pesquisas, ou seja, adotando configurações de forma indiscriminada. Pelo contrário, sugere-se que a escolha dos algoritmos, medidas e tratamento dos dados seja feita de forma muito crítica, embasada em critérios de seleção conhecidos a partir da literatura ou, preferencialmente, a partir da aplicação de um método

preliminar de tomada de decisão como o apresentado nesta tese, visto que atualmente a literatura ainda é limitada na área de pesquisa em questão.

Na sequência, para cada um dos grupos formados pelos 60 métodos resultantes, definiu-se um modelo de referência, correspondendo ao objeto mais próximo ao centroide. A diversidade de modelos resultantes desse processo foi grande. De forma geral, observou-se repetição mais frequente de pares de modelos dentro de uma mesma matriz de dados do que a partir de medidas de similaridade ou de algoritmos de partição. Cada matriz teve um conjunto de modelos mais frequente diferente das demais. À parte da matriz, algumas combinações de algoritmos e medidas de similaridade também resultaram em modelos iguais ou similares. Por exemplo, a combinação do algoritmo Ligação Completa com as medidas City-block e distância Euclidiana resultaram sempre em um mesmo conjunto de modelos, quando combinados com a mesma matriz de dados (os modelos mudavam ao mudar também a matriz de dados). O mesmo foi observado considerando o algoritmo de Ward com as medidas distância Euclidiana e distância Euclidiana Quadrada.

Respondendo ao terceiro objetivo específico, a partir da aplicação dos critérios de validação interna e relativa, foi possível encontrar o melhor método para a amostra em análise. No caso dessa pesquisa, foi o método M21_D, resultante da combinação do algoritmo K-means com a matriz de dados D (formada a partir da detecção de objetos atípicos e ponderação dos fatores). O método M18_D, formado pela combinação do algoritmo de Ward, distância Euclidiana Quadrada e Matriz D, apresentou a segunda melhor solução, baseando-se nos resultados da validação relativa. Esses dois métodos, na verdade, apresentaram formações muito parecidas e com os mesmos modelos de referência, diferindo apenas pela designação de poucos objetos a grupos diferentes.

Quanto à sua geometria, os dois agrupamentos formados pelo método selecionado mostraram-se bastante distintos. O primeiro é formado por habitações maiores, compostas em sua maioria por ambiente da sala independente da cozinha e três dormitórios, enquanto o segundo é mais propriamente descrito por habitações menores, de dois dormitórios e sala e cozinha integrada. O modelo de referência do primeiro grupo é uma edificação de 64m^2, sala independente com orientação solar norte e leste, e três dormitórios, todos com parede externa com orientação oeste. O segundo modelo caracteriza-se por uma edificação com apenas 37m^2. A sala e cozinha são conjugadas com orientação oeste, e dois dormitórios, ambos com orientação leste.

Quanto ao desempenho térmico, os testes de hipótese mostraram que os grupos diferem para a maioria dos indicadores de desempenho. Os dormitórios foram os ambientes que apresentaram maior diferença, tanto para o indicador de desempenho para resfriamento quanto para aquecimento. As salas, entretanto, aceitaram a hipótese de igualdade apenas para o indicador de desempenho para aquecimento, com exceção do indicador de desempenho para resfriamento da configuração que alterou a configuração dos materiais. Para essa configuração, todos os indicadores rejeitaram a hipótese de igualdade. Esse era um resultado esperado, pois as alterações propostas deixam a envoltória mais sensível às condições externas, ressaltando a influência das características geométricas sobre o desempenho. Quanto à comparação dos modelos, o diagrama de caixas mostrou que o indicador de desempenho dos modelos manteve-se dentro do intervalo do primeiro ao terceiro quartil, aproximando-se da mediana na maioria das variáveis.

É importante destacar que o método se direciona à descrição das características de habitações com propriedades similares e não à descrição detalhada de habitações individuais. Dessa forma, o resultado do *cluster* não deve ser considerado ao nível individual de cada habitação, mas a um grupo de habitações e seu potencial com respeito às medidas combinadas. Ele também não deve ser utilizado como modelo a ser replicado, visto que sua obtenção não está relacionada à qualidade projetual e sim à representatividade da amostra.

Quanto à influência dos diferentes parâmetros de clusterização, pode-se concluir que as diferentes configurações impactaram sobremaneira os resultados de agrupamentos. O algoritmo K-means apresentou, de forma geral, resultados razoáveis, independente dos demais parâmetros com os quais foi combinado. Apresentou, entretanto, desempenho inferior para as matrizes sem detecção de objetos atípicos (Matriz A e Matriz C). Esse comportamento já havia sido apontado na literatura. O algoritmo Ligação Simples apresentou o pior desempenho, não gerando sequer uma única formação válida. O estudo desenvolvido por Geyer et al. (2017), mencionado na revisão de literatura, já havia obtido resultado semelhante. Isso deve ocorrer pois a estrutura de dados que se referem à geometria de edificações parece ter variações mais lineares, enquanto o algoritmo mencionado destaca-se por conseguir separar grupos com formações mais orgânicas (por isso, muitas vezes usado em estudos da área biológica). Os algoritmos Ward e Ligação Completa também apresentaram resultados razoáveis, este último surpreendendo, visto que não foi encontrado em nenhum outro estudo da área. Por serem as habitações de interesse social de Florianópolis um grupo pequeno, bem delimitado e pouco heterogêneo (quanto às suas características), é provável que a habilidade do algoritmo de obter

as distâncias a partir de objetos mais distantes tenha auxiliado no processo de segregação da amostra em grupos bem distintos.

As medidas de similaridade também mostraram desempenhos diferentes, especialmente quando combinadas com alguns algoritmos. As distâncias City-block, distância Euclidiana e distância Euclidiana Quadrada resultaram em boa formação, enquanto a medida de correlação de Pearson não teve bom desempenho. Isso deve ter ocorrido por essa medida de similaridade ter maior habilidade para separar objetos que apresentam o mesmo padrão ao longo das variáveis (HAIR et al., 2009), enquanto a estrutura dos dados quanto a geometria das habitações parece apresentar relação maior de proximidade. A medida City-block, embora tenha tido bons resultados, não foi encontrada em nenhum estudo da área.

O tratamento de dados (e formação das diferentes matrizes) foi a etapa que apresentou as maiores diferenças. Isso ficou evidente pela determinação dos modelos de referência, que permaneceram similares dentro de cada matriz, mesmo quando combinada com diferentes algoritmos e medidas de distância, mas diferentes ao mudar a matriz de dados. O tratamento que apresentou a maior influência (positiva) nos resultados foi a ponderação dos fatores. Por ter ressaltado as variáveis de entrada com maior impacto nas variáveis dependentes, resultou em boa formação tanto a partir da validação interna quanto validação relativa. Quanto à padronização dos dados, esperava-se melhor desempenho desse tratamento estatístico, visto que todos os estudos apontam essa técnica como adequada e até necessária ao realizar a análise de agrupamentos. Entretanto, esse resultado deve ser interpretado com cautela, pois é possível que seja fruto de coincidência em função dos dados da amostra: como as variáveis mais influentes no desempenho foram também aquelas com maior dispersão, padronizar os dados, exclusivamente nessa amostra, pode ter minimizado a influência de variáveis importantes no processo de clusterização.

É preciso ressaltar que a análise de agrupamentos é uma análise extremamente dependente dos dados envolvidos, portanto, as conclusões aqui apresentadas servem apenas como referência a outros estudos para comparação, não devendo os resultados serem extrapolados. Em outras palavras, embora há um indicativo de parâmetros e combinações que resultaram em melhor e pior desempenho, é possível obter resultados diferentes para outras amostras. Infelizmente, não há estudos na área suficientes para que alguma conclusão mais geral possa ser feita.

A partir do trabalho realizado, concluiu-se que o objetivo geral desse estudo foi atingido e deu-se resposta a todos os objetivos específicos. A amostra foi dividida em

agrupamentos com formação distintas e modelos de referência de cada agrupamento puderam ser determinados. Provou-se a representatividade dos modelos a partir de análises descritivas e inferenciais. As medidas de validação utilizadas permitiram a identificação do método com melhor habilidade para dividir a amostra em grupos distintos, tanto para as variáveis de entrada quanto para variáveis dependentes (os indicadores de desempenho). Provou-se a necessidade de realizar a análise de agrupamentos a partir de um método como o proposto, composto por sucessivas etapas de tomada de decisão quanto aos tratamentos estatísticos a serem aplicados, as medidas de similaridade e algoritmos adotados e procedimentos para validação dos dados de cada método. Estudos futuros podem ser desenvolvidos com a replicação do método apresentado para diferentes áreas de estudo, desde que as variáveis que descrevem os objetos e os indicadores de desempenho estejam diretamente relacionadas com os objetivos da nova pesquisa.

5.1. Limitações do trabalho

Durante esse trabalho, alguns fatores foram observados como limitações:

- Esse estudo foi desenvolvido a partir de um banco de dados existente. Portanto, a amostra foi composta a partir da quantidade de habitações disponíveis, e não resultante de cálculos para dimensionamento amostral. Dessa forma, seria necessário adicionar outros procedimentos para confirmar sua representatividade quanto às habitações de interesse social de Florianópolis. Entretanto, isso seria necessário apenas para garantir que a amostra é estatisticamente significativa para a população. A análise de agrupamentos, por seu caráter exploratório e não inferencial, não é dependente de amostra com valor mínimo de elementos, ou seja, os resultados a partir dela são, necessariamente, representativos da amostra. Adicionalmente, amostras com pelo menos 100 casos são normalmente consideradas suficientes (HAIR et al., 2009);
- Quanto ao desempenho térmico, algumas variáveis importantes não foram incluídas na análise, como existência de ventilação cruzada, sombreamento e entorno. Essas considerações poderiam resultar em diferentes interpretações da análise;

- Devido às diferentes configurações espaciais das habitações, variando a quantidade de ambientes, dimensionamento e disposição na edificação, foi necessário adotar simplificações para comparar os indicadores de desempenho. Essas simplificações podem mascarar alguns resultados;
- A ausência de estudos similares que possibilite a comparação dos resultados também é uma limitação. Os resultados são muito dependentes da base de dados, motivo pelo qual os resultados obtidos aqui são válidos apenas para essa amostra. Outras análises, com outras amostras e variáveis, seriam necessárias para poder inferir se as conclusões obtidas aqui podem ser estendidas a outros estudos ou se as combinações que resultaram em bons agrupamentos são reflexo apenas dessa amostra.

5.2. Sugestões para trabalhos futuros

Ainda há poucos estudos desenvolvidos na área sobre a aplicação da análise de agrupamentos e a implicação do uso de diferentes medidas de similaridade e algoritmos sobre o resultado final. Esse contexto impossibilita a conclusão sobre a existência de um ou mais métodos ideais para a aplicação em edificações, de forma generalizada, considerando indicadores de desempenho. Outros estudos poderiam dar embasamento para tais conclusões, servindo como medida de comparação.

Dessa forma, como meio de contribuir para a complementação desse estudo, são sugestões para trabalhos futuros:

- A replicação deste método para outras amostras, considerando também outras variáveis (inclusive, padrões de comportamento) e outros tipos de edificação (comerciais, multifamiliares, etc.);
- Aplicar o método proposto adicionando outros tratamentos de dados, como análise de PCA, por exemplo, além de outras medidas de similaridade e algoritmos de partição;
- Considerar a utilização de outras formas de validação, por meio de medidas internas, externas ou relativas;

- A análise de agrupamentos baseia-se em critérios matemáticos para separação dos grupos. Aplicar métodos baseados em probabilidade, como o *Fuzzy*, ou em modelos caixa preta, como os Mapas de Kohonen, para a mesma amostra de dados, podem gerar resultados interessantes para comparação;
- Adicionar uma etapa de classificação de edificações externas à análise, a fim de verificar se os grupos formados atenderiam à adição de novas edificações.

REFERÊNCIAS

ALAIDROOS, A.; KRARTI, M. Optimal design of residential building envelope systems in the Kingdom of Saudi Arabia, **Energy and Buildings**, v. 86, p. 104-117, 2015.

AMERICAN SOCIETY OF HEATING, REFRIGERATING AND AIR-CONDITIONING ENGINEERS. **Thermal Environmental Conditions for Human Occupancy. ANSI/ASHRAE Standard 55**. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. Atlanta, 2010.

ATTIA, S.; EVRARD, A.; GRATIA, E. Development of Benchmark models for the Egyptian residential building sector. **Applied Energy**, v.94, n. 2012, p. 270-284, jun. 2012.

BALLARINI, I.; CORGNATI, S.; CORRADO, V. Use of reference buildings to assess the energy saving potentials of the residential building stock: the experience of TABULA project. **Energy Policy**, v. 66, p. 273-284, mai. 2014.

BHATNAGAR, M.; MATHUR, J.; GARG, V. Development of reference building models for India. **Jornal of Building Engeneering**, v. 21, p. 267–277, 2019.

BODACH, S.; HAMHABER, J. Energy efficiency in social housing: opportunities and barriers from a case study in Brazil. **Energy Policy**, v. 38, n. 12, p. 7898-7910, dez. 2010.

BRANDÃO, D. Q. Tipificação e aspectos morfológicos de arranjos espaciais de apartamentos no âmbito da análise do produto imobiliário brasileiro. **Ambiente Construído**, v. 3, p. 35–53, 2003.

BROWN, N.; UBBELOHDE, M. S.; LOISOS, G.; PHILIP, S. Quick Design Analysis for Improving Building Energy Performance. **Energy Procedia**, v. 57, p. 1868–1877, 2014.

BUILDINGS PERFORMANCE INSTITUTE EUROPE. **Europe's buildings under the microscope: A country-by-country review of the energy performance of buildings.** Buildings Performance Institute Europe. Bruxelas, 2011. Disponivel em: < http://bpie.eu/publication/europes-buildings-under-the-microscope/>. Acesso em: nov/2017.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. F. **Introdução à análise de agrupamentos.** IX Simpósio Nacional de Probabilidade e Estatística. **São Paulo, SP**: 1990.

CB3E. **Proposta de Instrução Normativa Inmetro para a Classe de Eficiência Energética de Edificações Residenciais**. 2018. Disponível em: <http://cb3e.ufsc.br/sites/default/files/2018-09-25-INI-R%20-%20Vers%C3%A3o02.pdf>. Acessado em: fev, 2019.

CHARISI, S. The role of the building envelope in achieving nearly-zero energy buildings (nZEBs), **Procedia Environmental Science**, v. 38, p; 115-120, 2017.

CICELSKY, A.; MEIR, I. A. Parametric analysis of environmentally responsive strategies for building envelopes specific for hot hyperarid regions. **Sustainable Cities and Society**, 2014.

CORGNATI, S. P.; FABRIZIO, E.; FILIPPI, M.; MONETTI, V. Reference buildings for cost optimal analysis: Method of definition and application. **Applied Energy**, v. 102, p. 983–993, jul. 2013.

DASCALAKI, E. G.; DROUTSA, K.; GAGLIA, A. G.; KONTOYIANNIDIS, S.; BALARAS, C. Data collection and analysis of the building stock and its energy performance—An example for Hellenic buildings. **Energy and Buildings**, v. 42, n. 8, p. 1231–1237, ago. 2010.

DASCALAKI, E. G.; DROUTSA, K. G; BALARAS, C. A; KONTOYIANNIDIS, S. Building typologies as a tool for assessing the energy performance of residential buildings – A case study for the Hellenic building stock. **Energy and Buildings**, v. 43, n. 12, p. 3400–3409, dez. 2011.

DOE. **Buildings Energy Data Book**. 2011 Disponível em: <http://buildingsdatabook.eren.doe.gov/ChapterIntro1.aspx>. Acesso em: nov. 2012.

DOE. **Residential Prototype Building Models**. Disponível em: <https://www.energycodes.gov/development/residential/iecc_models>. Acesso em: dez, 2019.

É, M.; KALAGASIDIS, S.; JOHNSSON, F. Building-stock aggregation through archetype buildings: France, Germany, Spain and the UK. Building and Environment, v.81, p. 270-282, nov. 2014.

EC. **Energy Efficiency Status Report**. Disponível em: <http://iet.jrc.ec.europa.eu/energyefficiency/sites/energyefficiency/files/energy-efficiency-status-report-2016.pdf>. Acesso em: jun. 2018.

EL-DARWISH, I. GOMAA, M. Retrofitting strategy for building envelopes to achive energy efficiency, **Alexandria Engineering Journal.** 2017.

EPE - EMPRESA DE PESQUISA ENERGÉTICA. Balanço Energético Nacional 2019: Ano base 2018. Rio de Janeiro: EPE, 2019. Disponível em: <http://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-377/topico-494/BEN%202019%20Completo%20WEB.pdf>. Acesso em: dez. 2019.

GAITANI, N.; LEHMANN, C.; SANTAMOURIS, M.; MIHALAKAKOU, G.; PATARGIAS, P. Using principal component and cluster analysis in the heating evaluation of the school building sector. **Applied Energy**, v. 87, n. 6, p. 2079–2086, jun. 2010.

GEYER, P.; SCHLÜTER, A.; CISAR, S. Application of clustering for the development of retrofit strategies for large building stocks. **Advanced Engineering Informatics**, v. 31, p. 32–47, 2017.

GHISI, E.; VIEIRA, A.S.; SCHAEFER, A.; MARINOSKI, A.K.; SILVA, A.S.; BALVEDI, B.F.; ALMEIDA, L.S.S. **Uso racional de água e eficiência energética em habitações de interesse social - Volume 1– Hábitos e indicadores de consumo de água e energia**. 2015. Relatório Técnico de Pesquisa. Florianópolis, 2015.

GIACOMIN, R.; CALMON, J. L.; VIEIRA, D.; CHAIN, M. C. Characterization of Representative Residential Buildings within a Neighborhood and Their Energy Efficiency Levels According to RTQ-R. **Applied Sciences**, v. 9, n. 18 , p. 3832-3854 . 2019.

GIGLIO, T., LAMBERTS, R; BARBOSA, M.; URBANO, M.. A procedure for analysing energy savings in multiple small solar water heaters installed in low-income housing in Brazil. **Energy Policy**, v. 72, p. 43–55, set. 2014.

HAIR, J. F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. ***Análise multivariada de dados***. 6. ed. Porto Alegre: Bookman, 2009.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster Validity Methods: Part I. **ACM SIGMOD Record**, v.31, n.2, p. 40-45, 2002a.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster Validity Checking Methods: Part II. **ACM SIGMOD Record**, v.31, n.3, p. 19-27, 2002b.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: concepts and techniques.** (3rd edition). 2011. Walford: Morgan Kaufmann.

IEA. **Key World Energy Statistics**. Paris. 2012. Disponível em: <http://www.iea.org/publications/freepublications/publication/kwes.pdf>. Acesso em: jun, 2018.

IPEA. INSTITUTO DE PESQUISA ECONÔMICA APLICADA. **Estimativas do déficit habitacional brasileiro (2007-2011) por municípios (2010)**. Brasília: maio. 2013. Disponível em: <http://www.ipea.gov.br/portal/images/stories/PDFs/nota_tecnica/130517_notatecnicadirur01.pdf>.

ISO Standard 13790. **Energy Performance of Buildings – Calculation of energy use for heating and cooling**. 2008.

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

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis.** 4 ed. New Jersey, Library of Congress:. 1998.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data:** an introduction to cluster analysis. New Jersey: John Wiley, 2005.

KOHLER, N.; HASSLER, U. The building stock as a research object. **Building Research and Information**, v. 30, p. 226-236, 2002.

KRAGH, J.; WITTCHEN, K. Development of two Danish building typologies for residential buildings. **Energy and Buildings**, v. 68, p. 79-86, jan. 2014.

LI, X.; YAO, R.; LIU, M.; COSTANZO, V.; YU, W.; WANG, W.; SHORT, A.; LI, B. Developing urban residential reference buildings using clustering analysis of satellite images. **Energy and Buildings**, v. 169, p. 417-429, 2018.

LIDDAMENT, M. W. **Air infiltration calculation techniques - an applications guide: AIVC**. Bracknell, UK, 1986.

LOGA, T.; DIEFENBACH, N.; DASCALAKI, E. G.; BALARAS, C. **Use of Building Typologies for Energy Performance Assessment of National Building Stocks. Existent Experiences in European Countries and Common Approach: First TABULA Synthesis Report**. Darmstadt: Institut Wohnen und Umwelt GmbH, 2008.

LOGA, T.; STEIN, B.; DIEFENBACH, N. TABULA building typologies in 20 European countries – Making energy-related features of residential building stocks comparable. **Energy and Buildings**, v. 132, p. 4-12, 2016.

LOUKAIDOU, K.; MICHOPOULOS, A.; ZACHARIADIS, T. Nearly-zero Energy Buildings: Cost-Optimal Analysis of Building Envelope Characteristics, **Procedia Environmental Science**, v.38, p. 20-27, 2017.

McKENNA, R.; MERKEL, E.; FEHRENBACH D.; MEHNE S.; FICHTNER, W. Energy efficiency in the German residential sector: A bottom-up building-stock-model-based analysis in the context of energy-political targets. **Building and Environment**, v. 62, p. 77-88, abr. 2013.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Ed. da UFMG, 2007.

MMA. **Projetando Edificações Energicamente Eficientes**. Disponível em: <http://projeteee.mma.gov.br/componentes-construtivos/>. Acessado em: nov, 2018.

PETCHARAT, S.; CHUNGPAIBULPATANA, S.; RAKKAWAMSUK, P. Assessment of potential energy saving using cluster analysis: A case study of lighting systems in buildings. **Energy and Buildings**, v. 52, p. 145-152, set. 2012.

ROSA, A.S. **Determinação de modelos de referência de habitações populares unifamiliares para Florianópolis através da análise de agrupamento**. 2014. Dissertação (Mestrado em Engenharia Civil) – Departamento de Engenharia Civil, Universidade Federal de Santa Catarina, Florianópolis, 2014.

SANCHES, T. B.; DAVID, N. **Levantamento das características tipológicas de edifícios de escritórios de Brasília.** IX Encontro Nacional e V Latino Americano de Conforto no Ambiente Construído. **Anais**...Ouro Preto: 2007.

SANDBERG, N.; SANTORI, I.; HEIDRICH, O.; DAWSON, R.; DASKALAKI, E.; et al. Dynamic building stock modelling: Application to 11 European countries to support the energy efficiency and retrofit ambitions of the EU. **Energy and Buildings**, v. 132, p. 26-38, nov. 2016.

SANGIREDDY, S. A. R.; BHATIA, A.; GARG, V. Development of a surrogate model by extracting top characteristic features vectors for building energy prediction. **Journal of Building Engineering**, v. 23, p. 38-52, mai. 2019.

SCHAEFER, A.; GHISI, E. Method for obtaining reference buildings. **Energy and Builings**, v. 128, p. 660–672, jul. 2016.

SELVI, T. S.; PARILAMA, R. Clustering Algorithms Validated Using Relative Index Validation. **International Journal of Computer Science and Engeneering**, v. 6, p. 85-95, 2018.

SERGHIDES, D. K.; DIMITRIOU, S.; KATAFYGIOTOU, M. C. Towards European targets by monitoring the energy profile of the Cyprus housing stock. **Energy and Buildings**, v. 132, p. 130–140, jul. 2016.

SOKOL, J.; CEREZO, D. C.; REINHART, C. Validation of a bayesian-based method for defining residential archetypes in urban building energy models. **Energy and Buildings**, v. 134, p. 11-24, jan. 2017.

TEIXEIRA, C. A.; MELO, A. P.; FOSSATI, M.; SORGATO, M. J.; LAMBERTS, R. **Estudo de modelos representativos para o setor residencial de edificações multifamiliares.** XIII Encontro Nacional e IX Encontro Latino Americano de Conforto no Ambiente Construído. **Anais**...Campinas: 2015.

THEODORIDOU, I.; PAPADOPOULOS, A. M.; HEGGER, M. A typological classification of the Greek residential building stock. **Energy and Buildings**, v.43, p. 2779-2787, jun. 2011a.

THEODORIDOU, I.; PAPADOPOULOS, A. M.; HEGGER, M. Statistical analysis of the Greek residential building stock. **Energy and Buildings**, v.43, p. 2779-2787, set. 2011b.

TORCELLINI, P.; DERU, M.; GRIFFITH, B.; BENNE, K. **DOE commercial building benchmark models.** ACEEE Summer Study on Energy Efficiency in Buildings. **Anais**...Washington: ACEEE, 2008.

TRIANA, M. A.; LAMBERTS, R..; SASSI, P. Characterization of representative building typologies for social housing projects in Brazil and its energy performance. **Energy Policy**, v. 87, p. 524–541, ago. 2015.

UE, 2010. **Diretiva 2010/31/EU do Parlamento Europeu e do Conselho (*council*) de 19 de maio de 2010 sobre o desempenho energético de edificações (*recast*)**. Jornal Oficial da União Europeia, 18 junho 2010. Disponível em: <http://europa.eu/legislation_summaries/other/l27042_en.htm>. Acesso em: dez. 2018.

UNEP. Buildings: Investing in energy and resource efficiency. In: UNEP (Ed.). **Towards a Green Economy: Pathways to Sustainable Development and Poverty Eradication**. 02.11.2011. United Nations Environment Programme, 2011. p. 330–373.

VERSAGE, R. **Metamodelo para estimar a carga térmica de edificações condicionadas artificialmente**. 2015. Tese (Doutorado em Engenharia Civil) – Departamento de Engenharia Civil, Universidade Federal de Santa Catarina, Florianópolis, 2015.

WITTEN, I. H.; FRANK, E.; **Data Mining: practical machine learning tools and techniques.** (2nd edition). 2005. Morgan Kaufmann.

YOUG, S. G. ; KIM, J. H. ; GIM, Y.; KIM, J.; CHO, H. HONG, J.; BAIK, Y. J. ; KOO, J. Impacts of building envelope design factors upon energy loads and their optimization in US standard climate zones using experimental design, **Energy and Buildings**, v. 141, p. 1-15, 2017.

YU, Z.; FUNG, B. C. M.; HAGHIGHAT, F.; YOSHINO, H.; MOROFSKY, E. A systematic procedure to study the influence of occupant behavior on building energy consumption. **Energy and Buildings**, v. 43, n. 6, p. 1409–1417, jun. 2011.

APÊNDICE A – Matriz A

Variáveis	Objeto																								
	Hab_001	Hab_002	Hab_003	Hab_004	Hab_005	Hab_006	Hab_007	Hab_008	Hab_009	Hab_010	Hab_011	Hab_012	Hab_013	Hab_014	Hab_015	Hab_016	Hab_017	Hab_018	Hab_019	Hab_020	Hab_021	Hab_022	Hab_023	Hab_025	Hab_027	Hab_028
Slcoz	0,00	1,00	1,00	0,00	1,00	1,00	0,00	1,00	1,00	1,00	1,00	0,00	1,00	0,00	1,00	0,00	0,00	0,00	1,00	1,00	0,00	1,00	1,00	1,00	1,00	1,00
N Dorm	4,00	2,00	2,00	3,00	2,00	2,00	2,00	1,00	2,00	2,00	2,00	3,00	2,00	3,00	2,00	4,00	3,00	3,00	2,00	4,00	2,00	2,00	2,00	2,00	1,00	2,00
N AmbT	3,00	1,00	1,00	3,00	1,00	1,00	2,00	2,00	1,00	1,00	1,00	3,00	2,00	3,00	3,00	7,00	2,00	4,00	2,00	2,00	2,00	1,00	2,00	1,00	2,00	1,00
N AmbPP	5,00	3,00	3,00	5,00	3,00	3,00	4,00	2,00	3,00	3,00	3,00	4,00	3,00	4,00	3,00	6,00	4,00	4,00	3,00	5,00	3,00	3,00	3,00	3,00	2,00	3,00
N Pavto	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	2,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
AU tot	66,57	37,82	45,03	74,25	33,24	33,20	47,18	27,44	34,53	38,04	28,80	64,04	52,34	64,47	49,71	135,64	63,88	80,35	33,93	83,56	32,07	46,85	45,5	43,58	59,95	49,71
A cob	66,57	37,82	45,03	74,25	33,24	33,20	47,18	27,44	34,53	38,04	28,80	64,04	52,34	64,47	49,71	76,54	63,88	80,35	33,93	83,56	32,07	46,85	45,5	43,58	59,95	49,71
PD	2,70	2,40	2,68	2,74	2,60	2,60	2,25	2,15	2,60	2,55	2,58	2,70	2,64	2,60	2,40	2,60	2,63	2,30	2,57	2,10	2,10	2,21	2,70	2,60	2,50	2,30
V tot	179,74	90,77	120,69	203,45	86,43	86,33	106,16	59,00	89,78	96,99	74,30	172,9	138,19	167,61	119,31	362,25	168,00	184,81	87,19	175,48	67,36	103,55	122,85	113,31	149,88	114,33
A par tot	95,88	59,04	72,41	103,35	60,35	60,32	62,45	47,01	61,42	64,47	55,52	86,78	92,92	86,31	87,58	209,56	84,73	95,94	64,04	83,87	51,91	64,88	77,17	69,97	79,67	88,77
A ian tot	9,42	4,98	6,75	7,36	5,06	5,58	5,22	4,20	5,58	4,47	4,68	7,53	6,07	7,03	6,55	12,20	5,72	10,70	1,29	7,75	6,03	5,13	6,92	4,57	11,94	4,62
A parN/ A parL	1,94	1,00	1,25	1,56	0,79	0,79	1,33	1,80	0,81	0,97	1,15	0,83	1,77	0,59	1,19	0,67	1,28	1,42	1,39	0,42	1,05	1,46	0,89	1,47	1,61	1,66
A par cob/ V	0,90	1,06	0,97	0,87	1,08	1,08	1,03	1,26	1,06	1,05	1,13	0,87	1,05	0,89	1,15	0,78	0,88	0,95	1,12	0,95	1,24	1,07	0,99	1,00	0,93	1,21
A jan tot/A par tot	9,82	8,43	9,32	7,13	8,38	9,25	8,37	8,93	9,09	6,93	8,43	8,68	6,53	8,15	7,48	5,82	6,76	11,15	2,02	9,24	11,61	7,91	8,97	6,53	14,99	5,21
AU	51,21	34,14	40,42	58,18	30,36	30,13	34,79	22,05	31,23	34,99	26,24	44,44	44,25	40,60	40,16	74,87	45,25	50,71	27,85	71,23	22,59	41,85	38,60	36,25	37,03	45,47
AU social	14,29	14,83	21,30	24,55	16,19	17,05	17,54	12,46	17,22	14,42	15,38	16,32	24,28	13,21	21,13	32,85	13,74	18,95	8,89	23,18	7,51	24,02	21,61	13,29	21,30	16,47
AU íntimo	36,92	19,31	19,11	33,63	14,17	13,07	17,25	9,59	14,01	20,57	10,86	28,12	19,97	27,39	19,03	42,02	31,51	31,76	18,96	48,05	15,08	17,83	16,99	22,96	15,73	29,00
Vol	138,27	81,93	108,31	159,41	78,95	78,34	78,27	47,40	81,20	89,24	67,69	119,98	116,83	105,55	96,38	194,67	119,02	116,62	71,58	149,59	47,44	92,49	104,21	94,25	92,58	104,58
Vol social	38,58	35,59	57,09	67,26	42,11	44,34	39,45	26,79	44,77	36,78	39,68	44,06	64,11	34,35	50,71	85,42	36,14	43,58	22,85	48,68	15,76	53,07	58,33	34,56	53,26	37,89
Vol íntimo	99,69	46,34	51,22	92,17	36,84	34,00	38,81	20,61	36,44	52,46	28,02	75,92	52,72	71,21	45,67	109,26	82,87	73,05	48,73	100,91	31,68	39,41	45,88	59,69	39,32	66,69
A par	74,01	49,75	68,49	77,02	50,66	50,62	39,69	32,68	51,27	55,05	46,67	62,80	77,53	58,28	62,26	94,67	60,68	55,65	47,26	67,66	33,26	50,86	60,11	55,08	35,59	77,23
A ian	9,12	4,68	6,40	5,94	4,76	5,28	3,60	2,64	5,28	4,17	4,42	5,97	5,76	5,30	5,46	7,28	4,45	7,14	0,99	6,41	4,62	4,67	4,83	4,37	1,80	4,29
A ian/ A par	12,32	9,41	9,34	7,71	9,40	10,43	9,07	8,08	10,30	7,57	9,48	9,51	7,43	9,10	8,77	7,69	7,34	12,83	2,09	9,47	13,89	9,18	8,04	7,93	5,06	5,56
A parN	21,33	14,76	16,25	22,77	6,52	6,86	13,14	12,14	6,90	12,78	12,02	19,74	25,92	16,12	14,28	31,77	23,80	20,38	15,93	12,54	5,29	15,55	15,60	11,00	19,14	24,23
A janN	1,68	0,00	0,00	1,06	0,80	1,32	0,80	0,00	1,32	1,32	1,23	2,55	5,04	1,68	1,42	1,44	1,01	5,94	0,00	0,00	0,00	1,33	0,00	1,42	1,80	0,00
A janN/ A parN	7,87	0,00	0,00	4,65	12,28	19,23	6,09	0,00	19,14	10,33	10,23	12,92	19,43	10,42	9,93	4,53	4,23	29,15	0,00	0,00	0,00	8,58	0,00	12,96	9,40	0,00
A parL	8,53	14,76	16,04	10,80	13,96	13,60	13,40	0,00	13,66	16,30	12,90	10,75	16,76	10,96	15,48	31,54	9,16	16,37	13,39	29,40	12,60	6,62	15,93	9,10	6,10	10,23
A janL	0,00	2,64	3,26	0,00	1,32	1,32	0,80	0,00	1,32	2,05	2,46	0,00	0,73	0,00	0,00	0,00	1,01	0,00	0,50	2,09	1,54	0,00	1,10	1,52	0,00	0,00
A janL/ A parL	0,00	17,89	20,33	0,00	9,45	9,71	5,97	0,00	9,66	12,55	19,06	0,00	4,33	0,00	0,00	0,00	11,00	0,00	3,70	7,12	12,22	0,00	6,92	16,71	0,00	0,00
A parS	27,88	10,70	20,17	26,72	13,37	13,36	13,14	12,14	13,83	15,93	14,86	8,67	18,08	7,97	18,12	2,89	9,16	6,23	10,61	6,33	10,50	15,55	12,65	20,87	0,00	26,10
A janS	4,08	0,72	1,51	2,14	2,64	2,64	2,00	2,64	2,64	0,00	0,00	0,00	0,00	0,00	1,21	0,00	0,00	1,20	0,00	1,10	3,08	0,00	1,10	0,00	0,00	2,60
A janS/ A parS	14,63	6,73	7,48	8,02	19,74	19,75	15,22	21,74	19,09	0,00	0,00	0,00	0,00	0,00	6,67	0,00	0,00	19,25	0,00	17,44	29,34	0,00	8,72	0,00	0,00	9,98
A parO	16,26	9,53	16,04	16,74	16,8	16,80	0,00	8,39	16,88	10,04	6,89	23,65	16,76	23,23	14,40	28,48	18,57	12,67	7,32	19,40	4,88	13,15	15,93	14,11	10,35	16,67
A janO	3,36	1,32	1,63	2,74	0,00	0,00	0,00	0,00	0,00	0,80	0,73	3,42	0,00	3,62	2,83	5,84	2,44	0,00	0,49	3,21	0,00	3,33	2,63	1,43	0,00	1,69
A janO/ A parO	20,66	13,85	10,17	16,35	0,00	0,00	0,00	0,00	0,00	7,97	10,66	14,46	0,00	15,59	19,69	20,52	13,13	0,00	6,76	16,55	0,00	25,36	16,48	10,10	0,00	10,14

Variáveis	Objeto																								
	Hab_029	Hab_030	Hab_031	Hab_032	Hab_033	Hab_035	Hab_036	Hab_037	Hab_038	Hab_039	Hab_040	Hab_041	Hab_042	Hab_043	Hab_044	Hab_045	Hab_046	Hab_047	Hab_048	Hab_049	Hab_051	Hab_052	Hab_054	Hab_055	Hab_056	Hab_057
Slcoz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	1,00	0,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00	1,00	0,00	1,00	0,00	0,00
N Dorm	1,00	2,00	3,00	1,00	2,00	3,00	2,00	3,00	2,00	2,00	3,00	3,00	2,00	3,00	1,00	4,00	3,00	1,00	4,00	2,00	2,00	2,00	2,00	1,00	3,00	2,00
N AmbT	3,00	2,00	3,00	3,00	4,00	4,00	4,00	2,00	3,00	2,00	2,00	1,00	2,00	4,00	2,00	4,00	2,00	2,00	5,00	2,00	4,00	1,00	2,00	2,00	4,00	2,00
N AmbPP	2,00	3,00	4,00	2,00	3,00	4,00	3,00	4,00	3,00	3,00	4,00	4,00	3,00	4,00	2,00	5,00	4,00	2,00	6,00	3,00	3,00	3,00	3,00	2,00	4,00	3,00
N Pavto	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
AU tot	73,78	65,09	70,30	41,30	60,15	56,95	81,21	71,19	71,93	54,87	71,25	54,56	50,87	92,78	34,84	93,84	54,32	32,31	91,89	50,51	60,47	37,08	42,66	28,51	81,15	46,78
A cob	73,78	65,09	70,30	41,30	60,15	56,95	81,21	71,19	71,93	54,87	71,25	54,56	50,87	92,78	34,84	93,84	54,32	32,31	91,89	50,51	60,47	37,08	42,66	28,51	81,15	46,78
PD	2,65	2,23	2,60	2,50	2,65	2,60	2,60	2,69	2,40	2,48	2,55	2,31	2,30	2,70	2,63	2,35	2,35	2,30	3,00	2,61	2,52	2,40	2,35	2,76	2,40	2,57
V tot	195,51	145,15	182,79	103,25	159,39	148,07	211,16	191,5	172,63	136,07	181,68	126,03	117,00	250,50	91,62	220,52	127,66	74,32	275,66	131,84	152,39	89,00	100,26	78,69	194,76	120,23
A par tot	101,33	90,48	107,10	65,24	83,73	83,18	93,80	93,44	82,61	76,38	90,06	75,74	66,28	109,81	62,79	105,83	72,52	52,34	126,89	78,89	82,35	61,67	61,66	63,75	103,76	76,06
A ian tot	7,70	5,74	5,92	9,61	3,93	8,48	7,65	5,23	3,63	5,29	4,52	7,13	6,67	7,96	4,62	8,64	6,48	3,90	10,37	4,37	4,73	4,26	5,40	2,83	6,62	4,98
A parN/ A parL	1,13	1,38	0,72	1,41	1,46	0,50	0,92	1,62	0,71	0,57	0,54	0,56	0,75	0,95	1,35	1,45	0,72	0,91	1,54	2,02	1,34	1,33	0,83	1,62	2,09	0,81
A par cob/ V	0,89	1,07	0,97	1,03	0,90	0,94	0,82	0,85	0,89	0,96	0,88	1,03	1,00	0,80	1,06	0,90	0,99	1,13	0,79	0,98	0,93	1,10	1,04	1,17	0,94	1,02
A jan tot/A par tot	7,60	6,34	5,53	14,72	4,70	10,19	8,16	5,60	4,40	6,93	5,02	9,41	10,07	7,25	7,36	8,16	8,93	7,44	8,17	5,54	5,75	6,91	8,76	4,44	6,38	6,54
AU	35,24	38,46	38,18	20,62	31,20	36,50	48,70	52,37	45,73	42,87	39,46	50,66	39,40	53,14	26,00	68,19	35,64	24,15	65,51	32,18	31,97	34,09	29,91	23,24	40,20	31,78
AU social	19,23	8,69	13,14	8,58	10,67	11,94	17,08	15,96	21,60	13,62	11,26	17,66	21,62	23,76	11,22	21,36	12,69	16,87	27,87	10,49	12,49	16,41	12,45	13,77	14,31	10,79
AU íntimo	16,02	29,77	25,04	12,04	20,54	24,57	31,61	36,42	24,12	29,25	28,20	33,01	17,78	29,38	14,78	46,84	22,95	7,28	37,64	21,69	19,48	17,68	17,46	9,47	25,89	20,99
Vol	93,40	85,77	99,27	51,55	82,68	94,89	126,61	140,88	109,75	106,33	100,62	117,03	90,61	143,48	68,38	160,26	83,76	55,54	196,53	84,00	80,57	81,82	70,30	64,15	96,47	81,67
Vol social	50,95	19,38	34,16	21,45	28,26	31,04	44,41	42,92	51,84	33,78	28,70	40,78	49,72	64,15	29,51	50,20	29,83	38,80	83,63	27,37	31,47	39,39	29,27	38,00	34,34	27,73
Vol íntimo	42,44	66,39	65,11	30,10	54,41	63,85	82,20	97,96	57,90	72,55	71,92	76,24	40,89	79,32	38,87	110,06	53,92	16,74	112,91	56,63	49,09	42,43	41,02	26,15	62,13	53,94
A par	44,67	61,14	68,14	31,87	44,49	49,76	51,24	65,66	55,79	50,59	48,16	71,52	50,48	42,46	53,64	74,31	47,38	33,67	91,58	43,88	35,65	49,91	44,6	42,09	43,03	55,63
A ian	5,20	3,30	5,20	3,31	2,59	7,20	5,67	4,22	2,47	5,29	3,20	6,94	4,96	5,56	4,40	6,90	5,12	2,38	8,94	2,87	2,38	3,96	4,32	1,31	4,83	4,13
A jan/ A par	11,64	5,40	7,63	10,37	5,81	14,47	11,07	6,43	4,43	10,46	6,64	9,70	9,82	13,09	8,20	9,28	10,80	7,06	9,76	6,54	6,67	7,93	9,69	3,12	11,23	7,43
A parN	0,00	6,62	9,42	12,80	17,57	13,91	8,37	28,89	10,02	0,00	15,91	13,64	7,00	18,71	13,46	27,66	7,00	3,61	38,49	19,21	4,01	14,88	14,00	8,64	13,37	7,99
A janN	0,00	1,12	1,32	2,20	2,35	1,80	1,43	1,80	0,00	0,00	2,40	1,84	0,00	1,39	0,00	4,69	0,00	0,00	5,96	0,95	0,00	0,99	2,88	0,00	2,40	0,00
A janN/ A parN	0,00	16,89	14,01	17,22	13,39	12,94	17,08	6,24	0,00	0,00	15,08	13,47	0,00	7,44	0,00	16,95	0,00	0,00	15,49	4,95	0,00	6,65	20,56	0,00	17,95	0,00
A parL	12,16	12,93	14,46	13,50	16,96	11,29	11,39	17,83	20,52	16,18	16,13	24,23	15,81	0,00	13,36	21,59	18,10	6,46	21,87	9,68	7,94	13,20	6,77	8,78	0,00	17,35
A janL	0,00	1,09	1,52	1,10	0,23	0,00	0,00	2,42	0,00	0,00	0,00	3,08	0,00	0,00	1,10	1,74	3,17	1,19	1,09	0,96	1,18	0,99	0,00	1,31	0,00	2,75
A janL/ A parL	0,00	8,45	10,51	8,16	1,38	0,00	0,00	13,55	0,00	0,00	0,00	12,73	0,00	0,00	8,23	8,08	17,54	18,38	5,00	9,92	14,84	7,50	0,00	14,94	0,00	15,88
A parS	20,35	22,65	22,55	5,57	9,96	0,00	22,51	9,15	17,16	13,89	0,00	13,64	14,24	0,00	13,46	16,4	15,25	12,53	19,98	7,22	13,43	14,88	7,00	16,86	16,04	17,11
A janS	5,20	0,00	1,60	0,00	0,00	0,00	4,24	0,00	1,13	2,64	0,00	1,67	3,19	0,00	1,10	0,47	1,94	1,19	1,89	0,00	1,20	0,99	0,00	0,00	1,11	0,00
A janS/ A parS	25,55	0,00	7,09	0,00	0,00	0,00	18,83	0,00	6,57	19,05	0,00	12,23	22,38	0,00	8,17	2,84	12,74	9,48	9,45	0,00	8,94	6,65	0,00	0,00	6,94	0,00
A parO	12,16	18,93	21,70	0,00	0,00	24,56	8,97	9,79	8,10	20,52	16,13	20,01	13,43	23,75	13,36	8,66	7,03	11,06	11,25	7,78	10,28	6,96	16,82	7,81	13,62	13,18
A janO	0,00	1,09	0,76	0,00	0,00	5,40	0,00	0,00	1,35	2,65	0,80	0,35	1,77	4,17	2,20	0,00	0,00	0,00	0,00	0,96	0,00	0,99	1,44	0,00	1,32	1,38
A janO/ A parO	0,00	5,77	3,50	0,00	0,00	21,98	0,00	0,00	16,61	12,89	4,96	1,74	13,18	17,54	16,47	0,00	0,00	0,00	0,00	12,35	0,00	14,23	8,56	0,00	9,69	10,45

Variáveis	Objeto																								
	Hab_058	Hab_059	Hab_060	Hab_061	Hab_062	Hab_063	Hab_064	Hab_065	Hab_066	Hab_067	Hab_068	Hab_069	Hab_070	Hab_071	Hab_072	Hab_073	Hab_074	Hab_075	Hab_076	Hab_077	Hab_078	Hab_079	Hab_080	Hab_081	Hab_082	Hab_083
Slcoz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
N Dorm	3,00	2,00	1,00	2,00	3,00	2,00	3,00	3,00	2,00	2,00	4,00	3,00	1,00	0,00	4,00	3,00	3,00	2,00	2,00	2,00	1,00	1,00	1,00	2,00	3,00	2,00
N AmbT	2,00	3,00	3,00	2,00	2,00	4,00	3,00	3,00	2,00	3,00	3,00	2,00	2,00	2,00	2,00	3,00	4,00	1,00	1,00	2,00	1,00	1,00	2,00	3,00	2,00	2,00
N AmbPP	4,00	3,00	2,00	3,00	4,00	4,00	4,00	4,00	4,00	3,00	5,00	4,00	2,00	1,00	5,00	4,00	4,00	3,00	3,00	3,00	2,00	2,00	2,00	3,00	4,00	3,00
N Pavto	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	2,00	1,00	1,00	1,00	1,00	3,00	1,00
AU tot	69,41	69,51	52,64	40,46	77,45	62,66	72,24	62,08	40,42	57,84	88,47	82,71	27,24	54,05	71,25	66,86	68,98	35,88	38,14	48,87	30,52	30,52	36,75	29,45	74,92	60,65
A cob	69,41	69,51	52,64	40,46	77,45	62,66	72,24	62,08	40,42	57,84	88,47	82,71	27,24	54,05	71,25	66,86	68,98	35,88	38,14	25,41	30,52	30,52	36,75	29,45	25,58	60,65
PD	2,45	2,34	2,65	2,62	2,70	2,67	2,40	2,60	2,30	2,30	2,85	2,40	2,50	2,40	2,18	2,66	2,55	2,96	2,20	2,00	2,68	2,68	2,68	2,45	2,35	2,37
V tot	170,05	162,66	139,51	106,01	209,12	167,32	173,36	161,41	92,96	133,03	252,15	198,49	68,11	129,73	155,33	177,86	175,91	106,20	83,92	101,64	81,78	81,78	98,50	72,16	180,34	143,75
A par tot	82,95	90,17	80,39	66,91	95,46	86,02	82,41	85,06	60,49	74,03	116,00	89,93	58,24	75,91	80,52	87,66	94,29	71,01	56,54	82,00	59,39	59,39	65,18	53,41	150,02	83,95
A ian tot	7,33	7,60	5,41	5,19	6,97	6,18	8,54	8,70	4,45	7,66	10,06	6,80	2,60	4,28	9,45	8,01	7,57	4,14	4,23	4,31	3,94	4,14	4,12	3,09	7,93	3,26
A parN/ A parL	1,42	1,71	0,95	0,84	1,20	1,45	0,75	1,73	0,59	0,68	0,53	1,63	2,59	0,46	1,94	1,27	0,94	0,90	0,84	0,69	0,85	1,16	0,85	1,20	0,52	1,92
A par cob/ V	0,89	0,98	0,95	1,01	0,82	0,88	0,89	0,91	1,08	0,99	0,81	0,86	1,25	1,00	0,97	0,86	0,92	1,00	1,12	1,05	1,09	1,09	1,03	0,82	0,97	1,00
A ian tot/A par tot	8,84	8,42	6,73	7,76	7,30	7,19	10,36	10,23	7,36	10,35	8,68	7,56	4,47	5,63	11,74	9,13	8,03	5,83	7,49	5,26	6,63	6,97	6,32	5,78	5,29	3,89
AU	48,98	32,71	26,18	27,60	55,94	43,41	47,17	36,97	30,72	34,87	59,89	56,94	20,57	45,44	53,07	43,36	42,99	32,71	35,10	42,23	28,28	28,28	29,54	20,91	65,63	52,63
AU social	18,50	8,99	14,20	10,39	16,46	23,94	19,47	11,76	14,92	12,12	19,83	17,76	10,27	45,44	12,81	11,71	18,08	14,16	18,70	21,69	19,28	19,28	19,11	10,84	21,84	25,86
AU íntimo	30,46	23,72	11,97	17,20	39,48	19,48	27,70	25,21	15,79	22,75	40,06	39,18	10,30	0,00	40,26	31,66	24,92	18,55	16,41	20,54	9,00	9,00	10,43	10,07	43,79	26,77
Vol	119,99	76,54	69,37	72,31	151,03	115,91	113,21	96,12	70,65	80,19	170,67	136,65	51,42	109,05	115,69	115,34	109,63	96,81	77,23	84,45	75,79	75,79	79,17	51,24	156,37	124,73
Vol social	45,33	21,03	37,64	27,23	44,45	63,91	46,73	30,57	34,31	27,88	56,52	42,62	25,67	109,05	27,92	31,14	46,10	41,91	41,13	43,38	51,67	51,67	51,21	26,56	51,33	61,29
Vol íntimo	74,65	55,52	31,73	45,09	106,58	51,99	66,47	65,56	36,33	52,32	114,16	94,03	25,75	0,00	87,77	84,20	63,53	54,91	36,10	41,07	24,12	24,12	27,96	24,68	105,04	63,45
A par	58,67	44,97	41,18	47,72	69,41	52,83	58,27	47,59	40,11	48,95	66,12	55,47	39,85	55,64	52,40	62,48	49,94	60,35	53,20	65,20	56,20	56,20	50,44	38,59	132,16	65,22
A ian	5,97	4,16	3,12	4,13	5,28	4,42	6,52	5,15	3,20	5,12	7,68	4,60	1,77	2,98	7,88	5,08	4,59	3,78	4,00	4,12	3,69	3,89	2,89	3,09	7,81	2,91
A ian/ A par	10,18	9,26	7,58	8,66	7,61	8,36	11,2	10,82	7,98	10,45	11,62	8,29	4,45	5,35	15,04	8,13	9,19	6,26	7,52	6,32	6,57	6,93	5,72	8,00	5,91	4,46
A parN	24,40	9,57	3,47	15,30	16,53	15,80	6,79	21,73	11,27	8,27	20,23	19,19	15,87	0,00	23,76	14,91	15,60	16,87	12,98	13,96	13,72	15,97	9,27	2,94	25,87	12,58
A ianN	2,22	0,00	0,00	1,19	1,33	1,10	0,00	3,75	1,60	0,00	3,60	0,23	0,00	0,00	2,55	1,07	2,75	0,00	2,00	1,71	1,54	0,00	1,35	0,00	4,16	0,00
A ianN/ A parN	9,11	0,00	0,00	7,76	8,04	6,99	0,00	17,26	14,2	0,00	17,79	1,18	0,00	0,00	10,73	7,21	17,66	0,00	15,41	12,25	11,20	0,00	14,54	0,00	16,09	0,00
A parL	17,07	0,00	10,76	6,62	21,65	0,00	23,49	15,56	14,42	9,37	25,93	0,00	8,10	21,83	13,67	8,47	24,22	18,63	11,95	24,20	12,78	13,72	8,58	12,13	38,68	10,67
A ianL	3,75	0,00	1,56	0,00	2,62	0,00	3,60	1,40	0,80	0,00	2,88	0,00	0,97	2,98	4,00	1,46	1,10	2,52	0,00	0,00	0,81	0,00	0,00	0,77	0,00	0,00
A ianL/ A parL	21,96	0,00	14,5	0,00	12,10	0,00	15,32	9,00	5,55	0,00	11,1	0,00	11,97	13,64	29,29	17,29	4,56	13,52	0,00	0,00	6,32	0,00	0,00	6,33	0,00	0,00
A parS	9,70	22,30	16,19	7,64	21,84	19,49	17,71	10,30	0,00	15,11	0,00	19,19	15,87	11,99	13,92	19,87	4,13	10,77	12,98	14,82	13,72	12,78	15,06	14,58	22,61	27,63
A ianS	0,00	3,15	0,00	0,00	0,00	1,10	2,92	0,00	0,00	2,01	0,00	1,40	0,81	0,00	1,33	0,00	0,73	0,00	1,00	2,41	1,35	0,81	1,54	1,24	3,65	2,91
A ianS/ A parS	0,00	14,14	0,00	0,00	0,00	5,66	16,52	0,00	0,00	13,32	0,00	7,29	5,08	0,00	9,53	0,00	17,66	0,00	7,70	16,27	9,83	6,32	10,20	8,53	16,16	10,53
A parO	7,50	13,10	10,76	18,15	9,38	17,54	10,27	0,00	14,42	16,20	19,95	17,09	0,00	21,83	1,06	19,23	5,98	14,08	15,29	12,22	15,97	13,72	17,53	8,94	45,00	14,34
A ianO	0,00	1,01	1,56	2,94	1,33	2,21	0,00	0,00	0,80	3,10	1,20	2,97	0,00	0,00	0,00	2,54	0,00	1,26	1,00	0,00	0,00	3,08	0,00	1,08	0,00	0,00
A ianO/ A parO	0,00	7,72	14,50	16,22	14,18	12,59	0,00	0,00	5,55	19,16	6,02	17,40	0,00	0,00	0,00	13,20	0,00	8,95	6,54	0,00	0,00	22,48	0,00	12,02	0,00	0,00

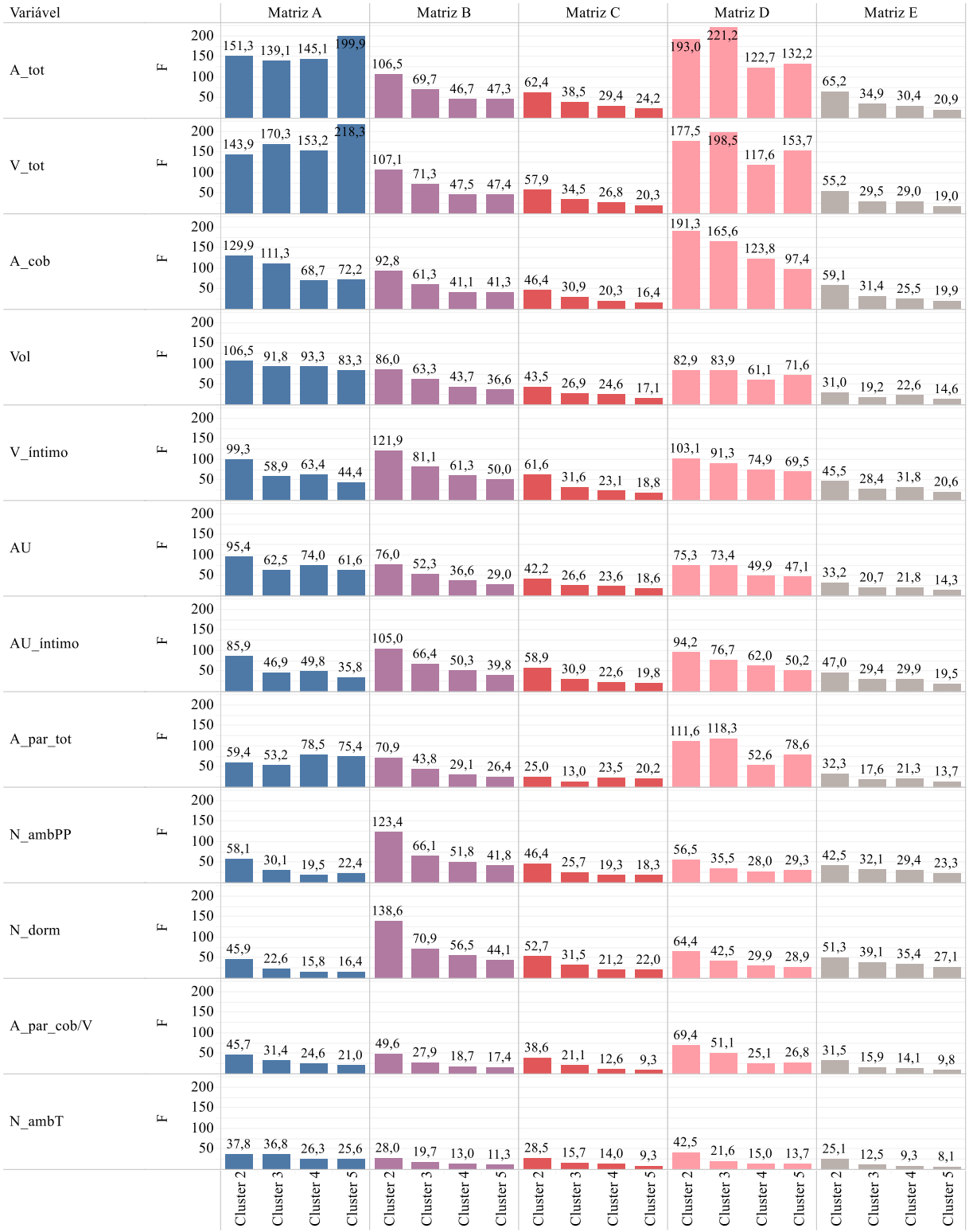
Variáveis	Objeto																							
	Hab_084	Hab_085	Hab_086	Hab_087	Hab_088	Hab_089	Hab_090	Hab_091	Hab_092	Hab_093	Hab_094	Hab_095	Hab_107	Hab_108	Hab_109	Hab_110	Hab_112	Hab_114	Hab_115	Hab_116	Hab_117	Hab_118	Hab_119	Hab_120
Slcoz	1,00	0,00	0,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00	1,00	1,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	1,00	1,00	1,00	0,00	1,00
N Dorm	3,00	3,00	3,00	2,00	3,00	3,00	2,00	2,00	2,00	3,00	1,00	2,00	1,00	2,00	1,00	5,00	2,00	3,00	3,00	1,00	3,00	1,00	3,00	2,00
N AmbT	1,00	3,00	2,00	1,00	2,00	3,00	2,00	3,00	5,00	5,00	1,00	2,00	3,00	2,00	0,00	2,00	3,00	3,00	2,00	1,00	1,00	1,00	2,00	2,00
N AmbPP	4,00	4,00	4,00	3,00	4,00	4,00	3,00	3,00	4,00	4,00	2,00	3,00	4,00	3,00	2,00	6,00	3,00	4,00	4,00	2,00	4,00	2,00	4,00	3,00
N Pavto	1,00	2,00	1,00	1,00	1,00	1,00	2,00	2,00	1,00	1,00	1,00	2,00	1,00	1,00	1,00	2,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
AU tot	54,30	64,18	82,89	55,39	63,03	58,14	47,94	46,60	105,55	87,63	34,27	48,81	66,48	38,24	12,51	54,02	59,70	56,76	60,33	22,29	29,23	32,20	76,14	48,59
A cob	54,30	42,83	82,89	55,39	63,03	58,14	25,41	31,71	105,55	87,63	34,27	35,26	66,48	38,24	12,51	36,26	59,70	56,76	60,33	22,29	29,23	32,20	76,14	48,59
PD	2,37	2,42	2,45	2,30	2,39	2,75	2,42	2,45	2,56	2,73	2,75	2,45	2,45	2,47	2,17	2,15	2,62	2,45	2,51	2,60	2,40	2,20	2,40	2,73
V tot	128,69	161,48	203,07	127,40	150,63	159,88	122,98	139,22	270,21	239,24	94,25	114,16	162,87	94,46	27,15	118,03	156,41	139,07	151,43	57,95	70,16	70,84	182,73	132,65
A par tot	71,34	111,85	98,88	83,90	75,90	87,67	99,22	107,46	114,63	106,2	67,49	85,87	163,67	65,31	31,68	78,91	81,90	80,07	89,56	53,67	56,02	52,98	85,73	91,78
A ian tot	6,24	12,36	7,17	4,46	6,01	9,98	5,18	4,02	17,16	8,48	4,16	8,25	4,14	6,36	2,71	8,33	5,68	7,95	9,94	2,01	5,16	3,67	7,24	3,42
A parN/ A parL	0,66	1,04	0,42	0,26	0,98	0,54	0,69	0,62	0,95	1,74	0,53	1,41	0,27	1,18	1,65	1,10	0,73	0,94	0,40	1,00	1,32	0,49	0,64	0,40
A par cob/ V	0,97	0,95	0,89	1,09	0,92	0,91	1,01	0,99	0,81	0,81	1,07	1,06	1,41	1,09	1,62	0,97	0,90	0,98	0,98	1,31	1,21	1,20	0,88	1,05
A ian tot/A par tot	8,75	11,05	7,25	5,32	7,91	11,38	5,22	3,74	14,97	7,98	6,16	9,61	2,53	9,73	8,55	10,55	6,93	9,93	11,10	3,75	9,21	6,93	8,44	3,73
AU	50,36	54,17	50,13	51,98	50,12	35,24	42,50	34,45	74,95	40,86	28,34	41,63	47,29	26,98	12,51	35,38	34,18	36,02	35,39	17,93	25,15	29,20	48,72	41,35
AU social	24,72	20,13	20,70	24,09	22,27	14,81	21,65	21,31	49,53	15,10	20,62	13,55	35,55	8,27	5,95	9,21	13,82	12,34	11,23	8,95	7,23	18,35	17,78	23,19
AU íntimo	25,65	34,03	29,43	27,88	27,85	20,42	20,85	13,14	25,41	25,75	7,72	28,09	11,73	18,72	6,57	26,16	20,37	23,68	24,16	8,98	17,92	10,86	30,95	18,16
Vol	119,36	131,08	122,81	119,56	119,78	96,91	102,84	84,39	191,87	111,54	77,92	96,57	115,85	66,65	27,15	77,95	89,55	88,25	88,84	46,62	60,36	64,25	116,93	112,88
Vol social	58,58	48,72	50,71	55,42	53,22	40,72	52,39	52,21	126,81	41,23	56,69	27,77	87,11	20,42	12,91	21,69	36,21	30,24	28,19	23,27	17,34	40,36	42,67	63,31
Vol íntimo	60,78	82,36	72,10	64,14	66,56	56,19	50,45	32,18	65,06	70,32	21,23	68,80	28,75	46,23	14,25	56,27	53,35	58,02	60,65	23,36	43,02	23,89	74,25	49,57
A par	66,79	80,47	50,57	81,14	56,28	49,78	78,70	58,68	61,15	42,72	62,26	66,37	87,69	44,43	31,68	52,76	43,21	51,47	47,64	38,18	42,17	45,33	50,14	78,05
A ian	5,95	7,09	3,65	4,22	4,75	9,68	4,82	4,02	13,42	3,80	3,89	7,98	3,71	4,31	2,71	7,44	3,38	6,63	7,20	1,84	4,95	3,48	5,16	3,24
A ian/ A par	8,90	8,81	7,22	5,20	8,44	19,44	6,12	6,85	21,94	8,89	6,25	12,02	4,23	9,71	8,55	14,10	7,82	12,88	15,11	4,83	11,74	7,68	10,29	4,15
A parN	14,22	14,28	8,45	8,85	13,81	0,00	16,87	10,29	0,00	22,88	11,83	23,54	12,39	7,64	9,87	10,66	0,00	10,04	12,80	6,70	12,05	4,68	16,87	13,19
A ianN	0,00	0,00	0,00	0,85	1,34	0,00	2,00	0,00	0,00	3,80	1,29	4,15	2,69	0,88	1,81	1,36	0,00	1,90	2,40	0,00	0,78	0,00	2,14	0,00
A ianN/ A parN	0,00	0,00	0,00	9,62	9,73	0,00	11,88	0,00	0,00	16,6	10,88	17,64	21,74	11,52	18,28	12,76	0,00	18,96	18,75	0,00	6,47	0,00	12,66	0,00
A parL	21,45	24,85	23,1	30,33	19,14	14,16	29,28	23,27	18,94	0,00	16,69	20,01	34,95	7,10	5,97	15,72	9,81	14,70	17,42	9,04	6,05	17,67	16,63	22,14
A ianL	3,42	4,46	2,29	2,52	1,34	2,19	0,00	1,20	1,53	0,00	0,00	1,50	0,00	0,83	0,90	1,47	1,20	1,75	2,40	0,00	1,39	1,16	1,12	0,00
A ianL/ A parL	15,92	17,96	9,90	8,29	7,02	15,48	0,00	5,16	8,07	0,00	0,00	7,50	0,00	11,74	15,12	9,34	12,23	11,88	13,78	0,00	22,98	6,57	6,73	0,00
A parS	14,22	25,19	0,00	8,85	15,27	15,54	17,91	10,29	23,94	9,75	11,83	10,09	5,39	14,75	9,87	10,66	17,40	19,40	0,00	13,42	12,05	8,82	0,00	13,19
A ianS	0,00	0,00	0,00	0,85	2,06	4,93	2,82	1,62	6,25	0,00	1,32	1,13	0,00	0,84	0,00	2,77	1,20	2,98	0,00	0,92	0,00	0,00	0,00	0,00
A ianS/ A parS	0,00	0,00	0,00	9,62	13,49	31,73	15,72	15,74	26,11	0,00	11,13	11,17	0,00	5,7	0,00	25,97	6,90	15,35	0,00	6,87	0,00	0,00	0,00	0,00
A parO	16,90	16,14	19,01	33,09	8,05	20,07	14,64	14,82	18,27	10,10	21,92	12,73	34,95	14,94	5,97	15,72	16,00	7,33	17,42	9,04	12,02	14,16	16,63	29,54
A ianO	2,53	2,62	1,36	0,00	0,00	2,55	0,00	1,20	5,64	0,00	1,29	1,20	1,01	1,76	0,00	1,85	0,98	0,00	2,40	0,92	2,78	2,32	1,90	3,24
A ianO/ A parO	14,97	16,25	7,17	0,00	0,00	12,72	0,00	8,10	30,85	0,00	5,87	9,42	2,90	11,78	0,00	11,74	6,13	0,00	13,78	10,21	23,12	16,38	11,42	10,96

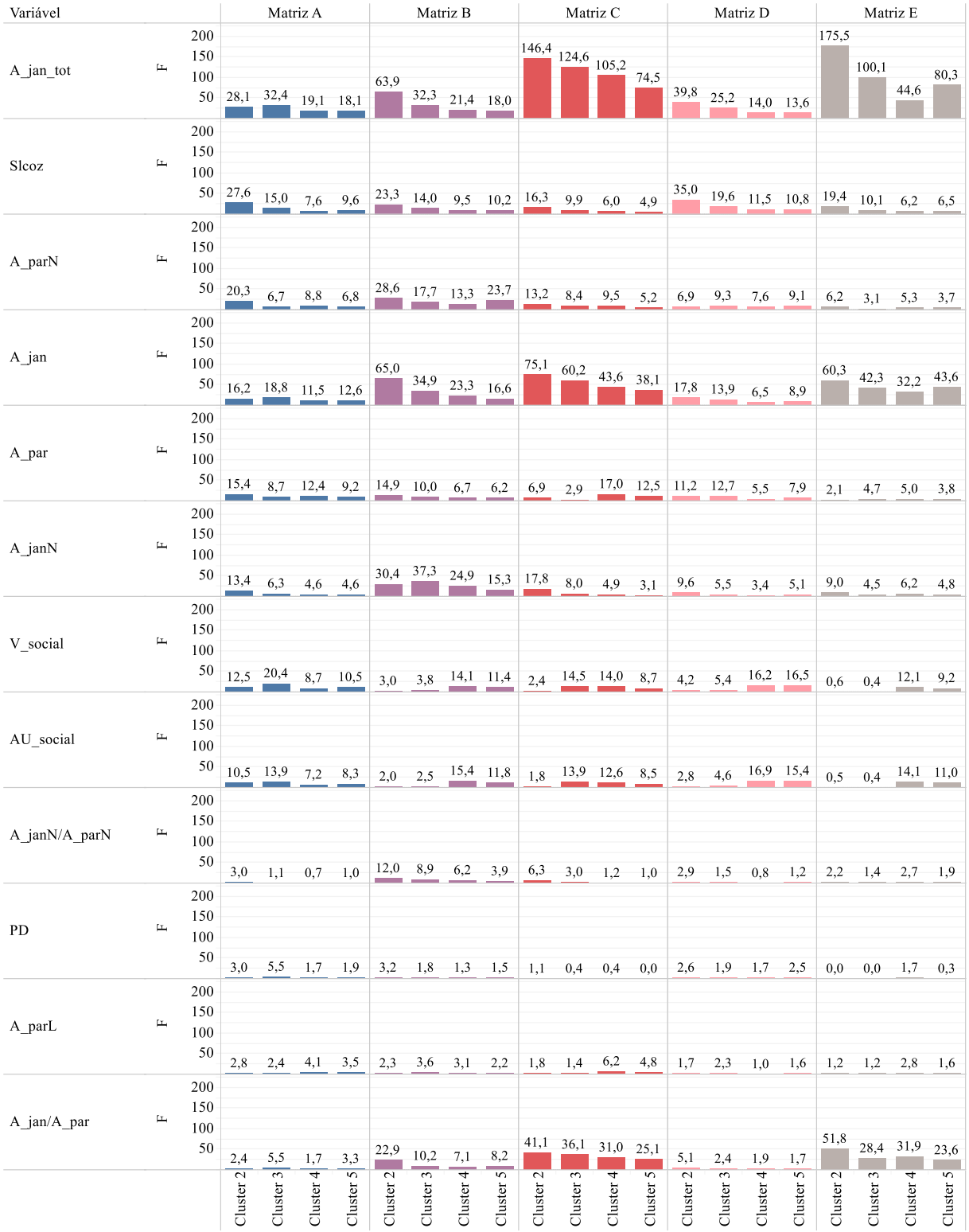
APÊNDICE B – Características dos objetos identificados como atípicos.

Variáveis envolvidas na análise	Identificação dos objetos						Estatísticas da amostra	
	Hab 016	Hab 107	Hab 082	Hab 092	Hab 091	Hab 110	Média	Desvio padrão
Slcoz	0	0	1	0	0	0	0	1
N_dorm	4	1	3	2	2	5	2	1
N_ambT	7	3	2	5	3	2	2	1
N_ambPP	6	4	4	4	3	6	3	1
N_pavto	2	1	3	1	2	2	1	0
A_tot (m^2)	135,64	66,48	74,92	105,55	46,60	54,02	55,86	20,31
A_cob (m^2)	76,54	66,48	25,58	105,55	31,71	36,26	53,68	19,62
PD (m)	2,60	2,45	2,35	2,56	2,45	2,15	2,49	0,19
V_tot (m^3)	362,25	162,87	180,34	270,21	139,22	118,03	140,36	54,24
A_parN/ A_parL	0,67	0,28	0,53	0,96	0,62	1,11	1,08	0,47
A_par_cob/ V	0,79	1,41	0,97	0,81	1,00	0,98	1,00	0,14
A_par_tot (m^2)	209,56	163,67	150,02	114,63	107,46	78,91	81,99	24,12
A_jan_tot (m^2)	12,20	4,14	7,93	17,16	4,02	8,33	6,28	2,50
A_jan_tot/ A_par_tot	5,82	2,53	5,29	14,97	3,74	10,55	7,74	2,39
AU_íntimo (m^2)	42,02	11,73	43,79	25,41	13,14	26,16	22,57	9,64
AU_social (m^2)	32,85	35,55	21,84	49,53	21,31	9,21	17,21	6,99
AU (m^2)	74,87	47,29	65,63	74,95	34,45	35,38	39,78	12,45
V_íntimo (m^3)	109,26	28,75	105,04	65,06	32,18	56,27	56,31	24,40
V_social (m^3)	85,42	87,11	51,33	126,81	52,21	21,69	42,97	17,92
Vol (m^3)	194,67	115,85	156,37	191,87	84,39	77,95	99,28	32,31
A_parN (m^2)	31,77	12,39	25,87	0,00	10,29	10,66	13,53	7,21
A_janN (m^2)	1,44	2,69	4,16	0,00	0,00	1,36	1,25	1,36

Variáveis envolvidas na análise	Identificação dos objetos						Estatísticas da amostra	
	Hab 016	Hab 107	Hab 082	Hab 092	Hab 091	Hab 110	Média	Desvio padrão
A_janN/ A_parN	4,53	21,74	16,09	0,00	0,00	12,76	7,96	7,42
A_parL (m^2)	31,54	34,95	38,68	18,94	23,27	15,72	14,58	7,57
A_janL (m^2)	0,00	0,00	0,00	1,53	1,20	1,47	1,10	1,13
A_janL/A_parL	0,00	0,00	0,00	8,07	5,16	9,34	7,25	6,99
A_parS (m^2)	2,89	5,39	22,61	23,94	10,29	10,66	13,29	6,63
A_janS (m^2)	0,00	0,00	3,65	6,25	1,62	2,77	1,16	1,35
A_janS/A_parS	0,00	0,00	16,16	26,11	15,74	25,97	7,55	8,19
A_parO (m^2)	28,48	34,95	45,00	18,27	14,82	15,72	14,30	7,27
A_janO(m^2)	5,84	1,01	0,00	5,64	1,20	1,85	1,29	1,38
A_janO/ A_parO	20,52	2,90	0,00	30,85	8,10	11,74	8,02	7,66
A_par (m^2)	94,67	87,69	132,16	61,15	58,68	52,76	55,71	15,35
AU (m^2)	7,28	3,71	7,81	13,42	4,02	7,44	4,81	1,92
A_jan/A_par	7,69	4,23	5,91	21,94	6,85	14,10	8,77	3,09

APÊNDICE C – Resultados da análise de variância para os métodos formados a partir de técnicas não hierárquicas de agrupamento.




Variável
Matriz A
Matriz B
Matriz C
Matriz D
Matriz E
A_jan_tot
28,1 32,4 19,1 18,1 63,9 32,3 21,4 18,0 146,4 124,6 105,2 74,5 39,8 25,2 14,0 13,6 175,5 100,1 44,6 80,3
Slcoz
27,6 15,0 7,6 9,6 23,3 14,0 9,5 10,2 16,3 9,9 6,0 4,9 35,0 19,6 11,5 10,8 19,4 10,1 6,2 6,5
A_parN
20,3 6,7 8,8 6,8 28,6 17,7 13,3 23,7 13,2 8,4 9,5 5,2 6,9 9,3 7,6 9,1 6,2 3,1 5,3 3,7
A_jan
16,2 18,8 11,5 12,6 65,0 34,9 23,3 16,6 75,1 60,2 43,6 38,1 17,8 13,9 6,5 8,9 60,3 42,3 32,2 43,6
A_par
15,4 8,7 12,4 9,2 14,9 10,0 6,7 6,2 6,9 2,9 17,0 12,5 11,2 12,7 5,5 7,9 2,1 4,7 5,0 3,8
A_janN
13,4 6,3 4,6 4,6 30,4 37,3 24,9 15,3 17,8 8,0 4,9 3,1 9,6 5,5 3,4 5,1 9,0 4,5 6,2 4,8
V_social
12,5 20,4 8,7 10,5 3,0 3,8 14,1 11,4 2,4 14,5 14,0 8,7 4,2 5,4 16,2 16,5 0,6 0,4 12,1 9,2
AU_social
10,5 13,9 7,2 8,3 2,0 2,5 15,4 11,8 1,8 13,9 12,6 8,5 2,8 4,6 16,9 15,4 0,5 0,4 14,1 11,0
A_janN/A_parN
3,0 1,1 0,7 1,0 12,0 8,9 6,2 3,9 6,3 3,0 1,2 1,0 2,9 1,5 0,8 1,2 2,2 1,4 2,7 1,9
PD
3,0 5,5 1,7 1,9 3,2 1,8 1,3 1,5 1,1 0,4 0,4 0,0 2,6 1,9 1,7 2,5 0,0 0,0 1,7 0,3
A_parL
2,8 2,4 4,1 3,5 2,3 3,6 3,1 2,2 1,8 1,4 6,2 4,8 1,7 2,3 1,0 1,6 1,2 1,2 2,8 1,6
A_jan/A_par
2,4 5,5 1,7 3,3 22,9 10,2 7,1 8,2 41,1 36,1 31,0 25,1 5,1 2,4 1,9 1,7 51,8 28,4 31,9 23,6
F
200
150
100
50
Cluster 2
Cluster 3
Cluster 4
Cluster 5


Variável
Matriz A
Matriz B
Matriz C
Matriz D
Matriz E
A_jan/A_.. F
A_parO F
A_janL/A.. F
A_janO F
A_parN/.. F
A_janS F
A_janS/A.. F
N_pavto F
A_parS F
A_janO/A.. F
A_janL F
200
150
100
50
2,43 5,52 1,72 3,26 22,92 10,25 7,12 8,18 41,13 36,06 31,02 25,07 5,13 2,41 1,87 1,67 51,80 28,39 31,87 23,61
2,41 5,57 4,09 3,15 2,08 1,22 1,56 1,20 0,96 0,50 3,55 7,02 5,07 3,44 2,98 2,14 0,03 1,89 1,21 2,24
1,60 0,92 1,28 0,53 0,20 0,09 0,31 0,24 0,03 0,15 0,11 0,75 0,71 0,13 0,83 0,40 0,11 0,34 0,43 1,23
1,43 7,00 5,17 4,56 3,02 4,25 3,05 2,15 6,22 7,60 4,90 4,82 1,01 2,07 0,87 1,38 1,75 2,24 1,71 1,94
0,65 0,26 0,69 0,36 0,04 0,23 0,81 1,39 0,14 0,04 0,15 0,97 0,00 0,25 0,71 0,62 0,00 0,79 0,65 0,86
0,26 0,30 1,07 0,57 0,31 0,07 0,32 12,84 1,94 8,55 6,94 6,02 0,09 0,13 0,56 0,31 3,24 2,93 1,69 1,73
0,25 0,28 0,47 0,28 0,51 0,13 0,38 8,04 0,93 2,89 3,20 3,02 0,16 0,23 0,55 0,47 2,90 2,83 2,67 1,84
0,23 0,48 3,11 3,19 0,19 0,19 0,14 0,24 2,74 1,26 16,78 16,21 1,12 0,14 0,52 0,14 0,09 0,10 0,51 1,29
0,19 0,95 1,47 1,17 0,90 0,53 0,38 1,61 0,18 1,80 1,89 2,01 0,00 1,18 0,19 0,54 0,60 2,84 0,45 1,49
0,01 2,46 1,08 0,98 0,09 2,23 1,83 1,29 1,33 4,99 3,24 3,10 0,44 0,45 0,43 0,54 0,01 1,20 0,90 1,13
0,01 0,07 0,98 0,68 5,03 1,60 2,19 1,55 3,31 2,91 1,45 2,24 0,61 1,09 1,01 1,13 4,92 3,10 2,99 3,27
Cluster 2
Cluster 3
Cluster 4
Cluster 5

APÊNDICE D – Inércia global, inércia *inter-cluster* e inércia *intra-cluster*.

Método	**k**	**Inércia Global**	**Inércia *Inter-cluster***	**Inércia *Intra-cluster***	**Q**
M09_D	4	557.661	376.633	181.133	2,14
M17_D	3	557.661	365.089	192.572	2,13
M07_D	3	557.661	362.428	195.232	2,11
M08_D	3	557.661	362.428	195.232	2,11
M08_D	5	557.661	408.697	149.069	2,09
M06_A	4	735.832	284.866	450.966	2,04
M16_A	3	735.832	433.347	302.486	2,04
M19_D	3	557.661	366.268	191.393	2,03
M18_A	3	735.832	453.298	282.534	2,00
M19_A	3	735.832	429.875	305.957	1,91
M09_A	4	735.832	449.514	286.319	1,79
M16_D	2	557.661	257.111	300.550	1,67
M17_D	2	557.661	251.962	305.698	1,66
M10_E	4	557.661	273.841	283.819	1,33
M18_D	2	557.661	272.247	285.414	1,26
M17_C	4	735.832	278.714	457.119	1,22
M06_B	2	557.661	245.307	312.354	1,14
M21_D	2	557.661	288.163	269.498	1,13
M17_B	2	557.661	248.223	309.438	1,01
M17_B	3	557.661	248.223	309.438	1,01
M07_B	2	557.661	259.321	298.430	1,00
M08_B	2	557.661	259.321	298.430	1,00
M19_D	2	557.661	274.202	283.459	0,97
M06_D	2	557.661	272.945	284.716	0,97
M21_A	2	735.832	349.594	386.238	0,84
M19_A	2	735.832	348.352	387.481	0,83
M21_B	2	557.661	251.946	305.714	0,81
M07_D	2	557.661	240.679	316.982	0,81
M08_D	2	557.661	240.679	316.982	0,81
M17_A	2	735.832	341.874	393.959	0,79
M16_A	2	735.832	341.874	393.959	0,79
M18_B	2	557.661	234.304	323.357	0,72
M15_D	5	557.661	225.678	331.983	0,71
M16_B	2	557.661	251.946	305.714	0,71
M16_E	2	557.661	229.442	328.219	0,70
M10_B	4	557.661	217.758	339.903	0,70

APÊNDICE E – Índices estatísticos ponderados para quantificação de erros de cada método de agrupamento.

Método	k	d	ME	EAM	RMSE	EF	CRM	E_{global}
M21_D	2	0,165	0,337	0,230	0,248	0,138	0,126	1,24
M18_D	2	0,395	0,148	0,136	0,159	0,179	0,386	1,40
M17_B	3	0,604	0,170	0,167	0,166	0,318	0,509	1,44
M10_B	4	0,336	0,192	0,140	0,154	0,273	0,349	1,44
M10_E	4	0,378	0,162	0,127	0,293	0,073	0,373	1,60
M07_D	2	0,557	0,162	0,237	0,216	0,242	0,197	1,61
M08_D	2	0,557	0,162	0,237	0,216	0,242	0,197	1,61
M21_B	2	0,402	0,290	0,247	0,274	0,221	0,344	1,77
M18_B	2	0,362	0,320	0,250	0,288	0,196	0,374	1,79
M06_B	2	0,589	0,128	0,122	0,122	0,360	0,496	1,82
M16_B	2	0,396	0,320	0,255	0,290	0,186	0,371	1,82
M16_E	2	0,394	0,252	0,202	0,241	0,250	0,498	1,83
M17_B	2	0,604	0,170	0,167	0,166	0,318	0,509	1,93
M07_B	2	0,591	0,181	0,167	0,167	0,325	0,515	1,94
M08_B	2	0,591	0,181	0,167	0,167	0,325	0,515	1,94
M19_A	3	0,326	0,707	0,724	0,792	0,559	0,304	2,03
M19_D	2	0,643	0,331	0,432	0,420	0,028	0,183	2,03
M06_D	2	0,649	0,339	0,435	0,425	0,041	0,186	2,07
M15_D	5	0,217	0,450	0,401	0,445	0,366	0,213	2,09
M21_A	2	0,314	0,393	0,372	0,424	0,098	0,545	2,14
M19_A	2	0,326	0,384	0,372	0,418	0,104	0,544	2,14
M17_A	2	0,316	0,389	0,368	0,420	0,100	0,549	2,14
M16_A	2	0,316	0,389	0,368	0,420	0,100	0,549	2,14
M06_A	4	0,436	0,553	0,362	0,457	0,077	0,386	2,27
M19_D	3	0,281	0,560	0,645	0,668	0,457	0,106	2,71
M09_D	4	0,409	0,535	0,639	0,659	0,537	0,008	2,78
M17_D	3	0,113	0,621	0,690	0,733	0,582	0,230	2,96
M16_A	3	0,500	0,701	0,454	0,594	0,205	0,544	2,99
M17_C	4	0,546	0,546	0,434	0,519	0,451	0,538	3,03
M17_D	2	0,690	0,512	0,596	0,610	0,355	0,677	3,44
M08_D	5	0,349	0,720	0,699	0,760	0,652	0,518	3,69
M07_D	3	0,191	0,784	0,770	0,846	0,753	0,720	4,06
M08_D	3	0,191	0,784	0,770	0,846	0,753	0,720	4,06
M18_A	3	0,474	0,959	0,793	0,930	0,886	0,135	4,17
M09_A	4	0,504	1,001	0,813	0,961	1,000	0,028	4,30
M16_D	2	0,336	0,870	0,953	0,998	0,985	0,213	4,35

APÊNDICE F – Características descritivas dos modelos e agrupamentos.

Existência de sala e cozinha conjugadas	**Agrupamento**	
	1	2
Sim	20%	72%
Não	80%	28%
Modelo de referência	**Não**	**Sim**

Quantidade de dormitórios	**Agrupamento**	
	1	2
1	4%	32%
2	27%	64%
3	57%	4%
4 ou mais	12%	0%
Modelo de referência	**3**	**2**

Quantidade de ambientes de permanência transitória	**Agrupamento**	
	1	2
1	6%	42%
2	37%	47%
3	33%	11%
4 ou mais	24%	0%
Modelo de referência	**3**	**1**

Quantidade de ambientes de permanência prolongada	**Agrupamento**	
	1	2
1	0%	2%
2	4%	30%
3	25%	60%
4 ou mais	71%	9%
Modelo de referência	**4**	**3**

Quantidade de pavimentos	**Agrupamento**	
	1	2
1	98%	94%
2	2%	6%
Modelo de referência	**1**	**1**

Área total (m^2)	**Agrupamento**	
	1	2
Média	68,98	39,02
Desvio padrão	9,33	7,90
Modelo de referência	**64,04**	**37,08**

Área cobertura (m^2)	**Agrupamento**	
	1	2
Média	68,55	37,76
Desvio padrão	9,76	7,70
Modelo de referência	**64,04**	**37,08**

Pé direito (m)	**Agrupamento**	
	1	2
Média	2,53	2,47
Desvio padrão	0,15	0,17
Modelo de referência	**2,70**	**2,40**

Volume total (m^3)	**Agrupamento**	
	1	2
Média	174,77	96,16
Desvio padrão	25,14	18,62
Modelo de referência	**172,90**	**89,00**

Relação entre dimensão da fachada norte e fachada leste	**Agrupamento**	
	1	2
Média	1,11	1,10
Desvio padrão	0,45	0,33
Modelo de referência	**0,83**	**1,34**

Relação entre área de fachada exposta e cobertura	**Agrupamento**	
	1	2
Média	0,92	1,09
Desvio padrão	0,05	0,08
Modelo de referência	**0,87**	**1,11**

Área de parede (m^2)	**Agrupamento**	
	1	2
Média	90,92	65,61
Desvio padrão	8,86	9,00
Modelo de referência	**86,78**	**61,67**

Área de janela (m^2)	**Agrupamento**	
	1	2
Média	7,27	4,89
Desvio padrão	1,62	1,13
Modelo de referência	**7,53**	**4,26**

Relação de área de janela por área de parede	**Agrupamento**	
	1	2
Média	8,02	7,52
Desvio padrão	1,70	1,64
Modelo de referência	**8,68**	**6,91**

Área útil dos ambientes íntimos (m^2)	**Agrupamento**	
	1	2
Média	28,74	15,57
Desvio padrão	5,68	4,75
Modelo de referência	**28,12**	**17,68**

Área útil dos ambientes sociais (m^2)	**Agrupamento**	
	1	2
Média	17,47	15,51
Desvio padrão	4,22	4,50
Modelo de referência	**16,32**	**16,41**

Volume dos ambientes íntimos (m^3)	**Agrupamento**	
	1	2
Média	72,42	38,27
Desvio padrão	14,41	11,40
Modelo de referência	**75,92**	**42,43**

Volume dos ambientes sociais (m^3)	**Agrupamento**	
	1	2
Média	44,22	38,11
Desvio padrão	10,44	11,00
Modelo de referência	**44,06**	**39,39**

Área útil dos ambientes condicionados (m^2)	**Agrupamento**	
	1	2
Média	46,20	31,08
Desvio padrão	7,65	6,03
Modelo de referência	**44,44**	**34,09**

Volume dos ambientes condicionados (m^3)	**Agrupamento**	
	1	2
Média	116,64	76,38
Desvio padrão	17,87	13,97
Modelo de referência	119,98	81,82

Área de parede da fachada norte (m^2)	Agrupamento	
	1	2
Média	15,20	11,59
Desvio padrão	6,07	4,13
Modelo de referência	**19,74**	**14,88**

Área de janela da fachada norte (m^2)	**Agrupamento**	
	1	2
Média	1,64	0,81
Desvio padrão	1,20	0,77
Modelo de referência	**2,55**	**0,99**

Relação entre área de parede e área de janela da fachada norte	**Agrupamento**	
	1	2
Média	9,13	6,60
Desvio padrão	6,37	6,28
Modelo de referência	**12,92**	**6,65**

Área de parede da fachada leste (m^2)	**Agrupamento**	
	1	2
Média	14,97	12,89
Desvio padrão	6,29	3,92
Modelo de referência	**10,75**	**13,20**

Área de janela da fachada leste (m^2)	**Agrupamento**	
	1	2
Média	1,22	1,03
Desvio padrão	1,08	0,75
Modelo de referência	**0,00**	**0,99**

Relação entre área de parede e área de janela da fachada leste	**Agrupamento**	
	1	2
Média	6,87	8,09
Desvio padrão	6,12	5,87
Modelo de referência	**0,00**	**7,50**

Área de parede da fachada sul (m^2)	**Agrupamento**	
	1	2
Média	13,34	13,32
Desvio padrão	6,81	2,93
Modelo de referência	**8,67**	**14,88**

Área de janela da fachada sul (m^2)	**Agrupamento**	
	1	2
Média	1,13	1,05
Desvio padrão	1,12	0,86
Modelo de referência	**0,00**	**0,99**

Relação entre área de parede e área de janela da fachada sul	**Agrupamento**	
	1	2
Média	6,84	7,48
Desvio padrão	6,65	6,09
Modelo de referência	**0,00**	**6,65**

Área de parede da fachada oeste (m^2)	**Agrupamento**	
	1	2
Média	14,93	12,12
Desvio padrão	5,53	4,08
Modelo de referência	**13,65**	**6,96**

Área de janela da fachada oeste (m^2)	**Agrupamento**	
	1	2
Média	1,34	1,08
Desvio padrão	1,22	0,86
Modelo de referência	**3,42**	**0,99**

Relação entre área de parede e área de janela da fachada oeste	**Agrupamento**	
	1	2
Média	7,26	8,27
Desvio padrão	6,56	6,35
Modelo de referência	**3,99**	**14,23**

Área de parede dos ambientes condicionados (m^2)	**Agrupamento**	
	1	2
Média	58,14	49,93
Desvio padrão	10,60	8,40
Modelo de referência	**62,80**	**49,91**

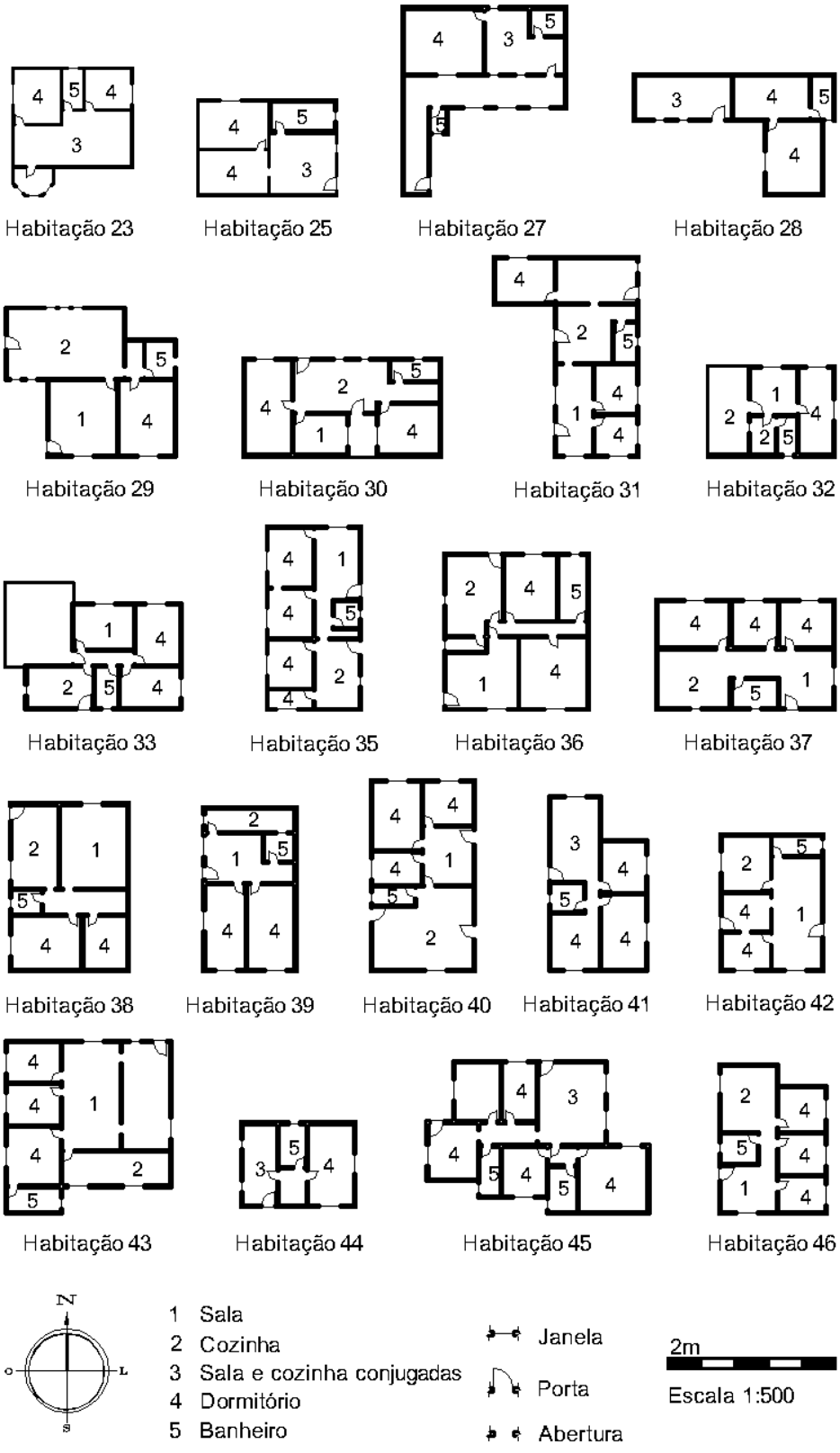
Área de janela dos ambientes condicionados (m^2)	**Agrupamento**	
	1	2
Média	5,32	3,98

Desvio padrão	1,40	0,94
Modelo de referência	5,97	3,96
Relação entre área de parede e área de janela dos ambientes condicionados	Agrupamento 1	2
Média	9,30	8,05
Desvio padrão	2,32	1,72
Modelo de referência	**9,51**	**7,93**

ANEXO A – Representação gráfica das habitações

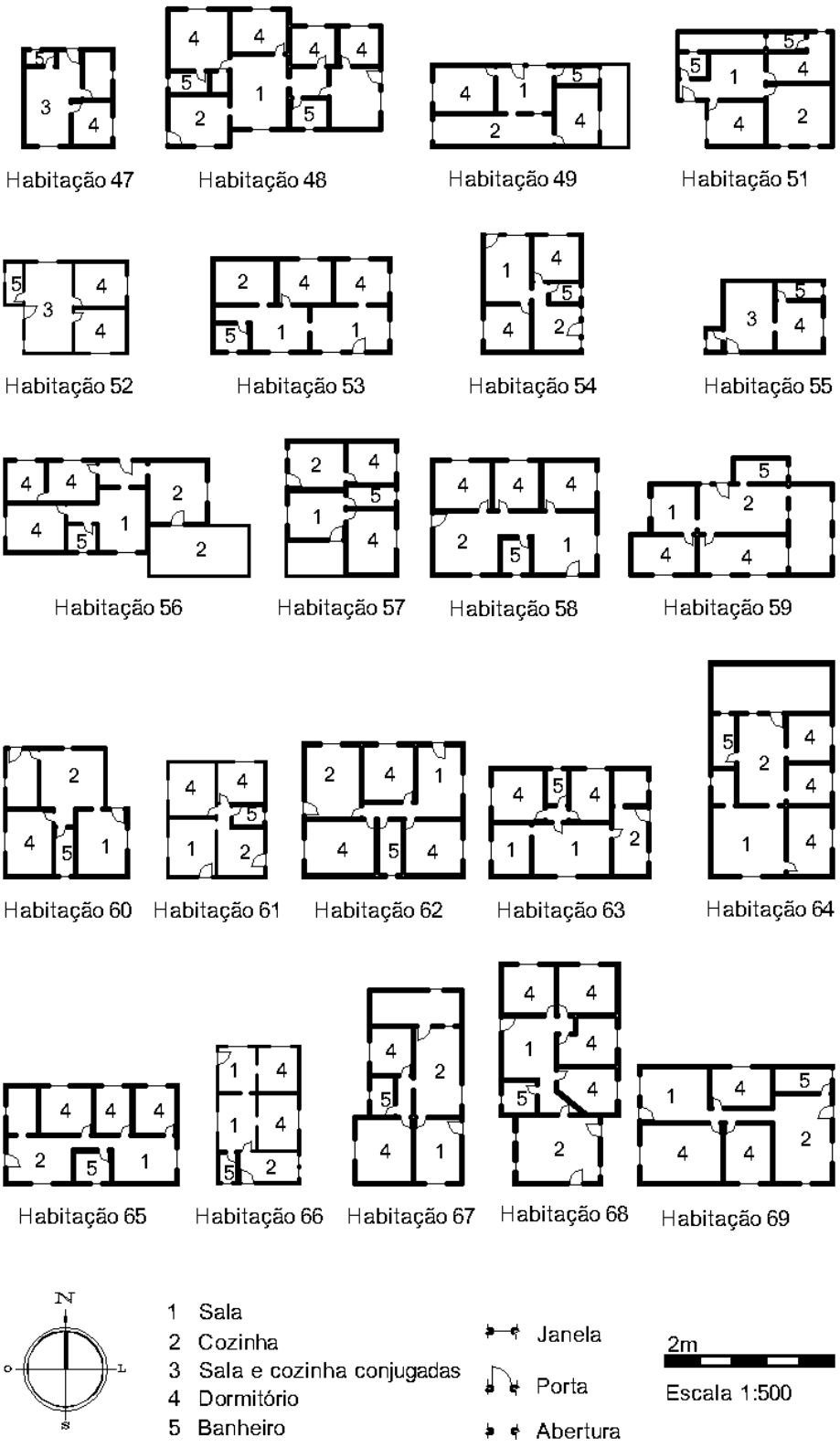
Obs.: As dimensões das aberturas nas plantas acima são meramente ilustrativas; não representam as dimensões reais.

Fonte: Rosa (2014).



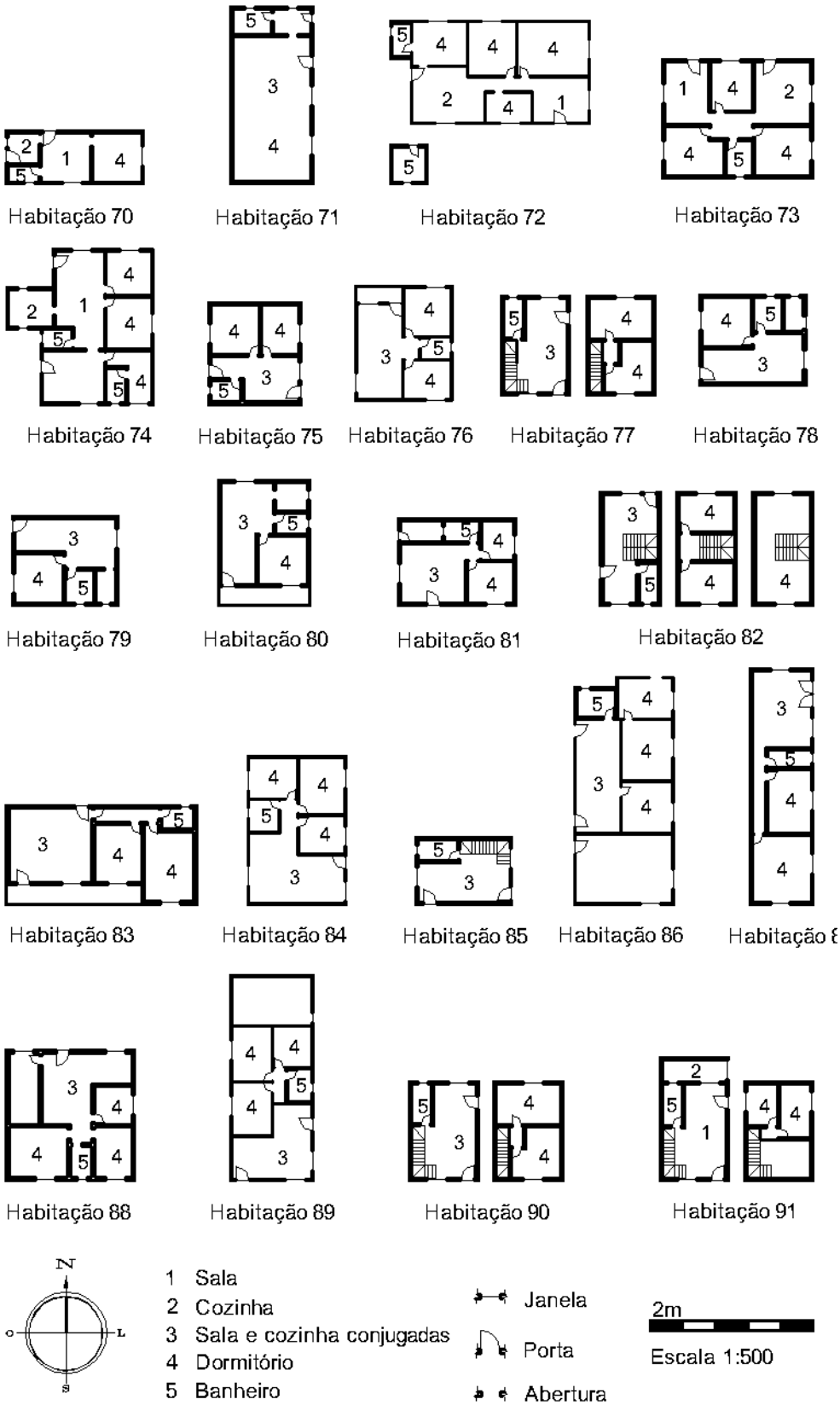
Obs.: As dimensões das aberturas nas plantas acima são meramente ilustrativas; não representam as dimensões reais

Fonte: Rosa (2014).



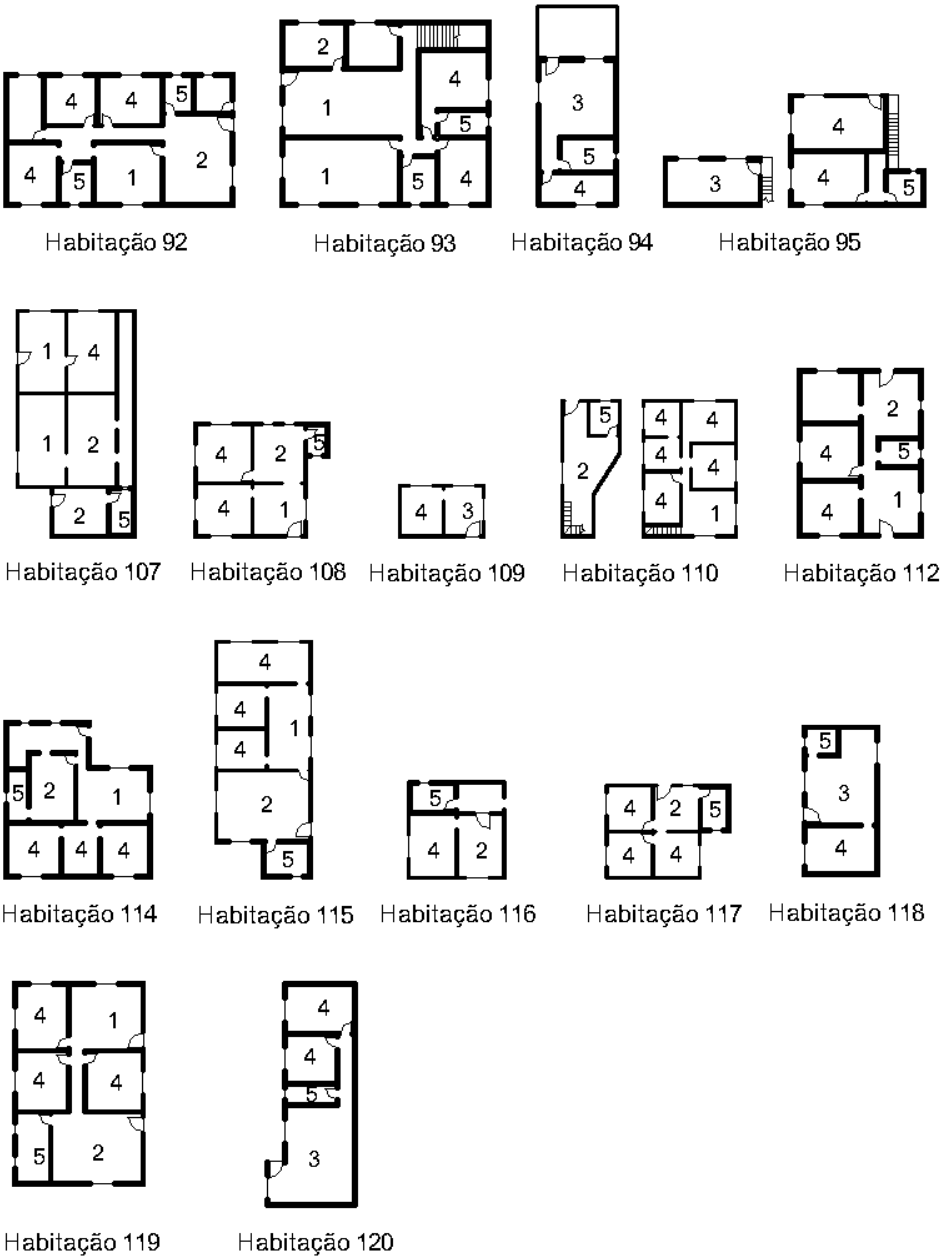
Obs.: As dimensões das aberturas nas plantas acima são meramente ilustrativas; não representam as dimensões reais

Fonte: Rosa (2014).



Obs.: As dimensões das aberturas nas plantas acima são meramente ilustrativas; não representam as dimensões reais

Fonte: Rosa (2014).



Obs.: As dimensões das aberturas nas plantas acima são meramente ilustrativas; não representam as dimensões reais

Fonte: Rosa (2014).