

Investigação dos Fatores Influenciadores da Qualidade do Ar em Habitações de Famílias com Crianças

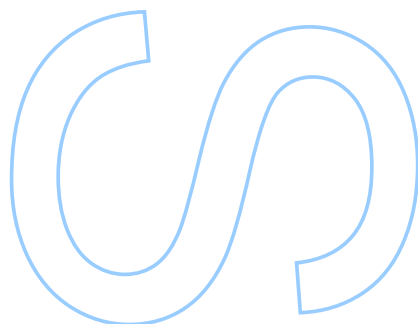
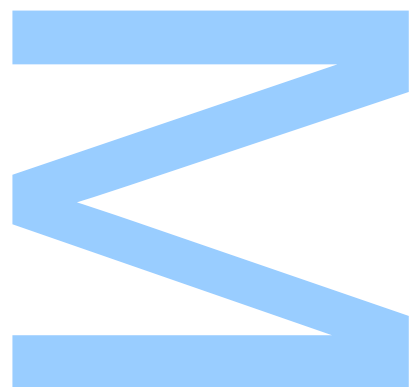
Beatriz Maria Fonseca Saraiva

Mestrado em Estatística Computacional e Análise de Dados

Departamento de Matemática

Faculdade de Ciências da Universidade do Porto

2025/2026



Investigação dos Fatores Influenciadores da Qualidade do Ar em Habitações de Famílias com Crianças

Beatriz Maria Fonseca Saraiva

Relatório de Estágio realizado no âmbito do Mestrado em
Estatística Computacional e Análise de Dados
Departamento de Matemática
2025/2026

Orientador

Ana Rita Pires Gaio, Professora Auxiliar, Departamento de
Matemática, Faculdade de Ciências da Universidade do Porto

Coorientador

Marta Sofia da Fonseca Gabriel, Investigadora Coordenadora, PhD,
Instituto de Ciência e Inovação em Engenharia Mecânica e
Engenharia Industrial (INEGI)



Agradecimentos

Gostaria de começar por expressar o meu agradecimento ao European project INQUIRE (Grant Agreement ID 101057499) pelo apoio ao desenvolvimento deste trabalho. Agradeço igualmente ao INEGI pela oportunidade de realizar este estágio e pela forma acolhedora com que me receberam. Em particular, à doutora Marta Gabriel e à doutora Fátima Felgueiras, pela disponibilidade constante e pela ajuda prestada durante todo o estágio. Um agradecimento especial à professora Ana Rita, que, para além da orientação incansável, me ensinou muito ao longo destes dois anos, sempre com grande dedicação. O meu maior obrigado vai para toda a minha família, pelo apoio incondicional ao longo de todo o meu percurso académico. Aos meus pais, que são os melhores do mundo, por todo o amor e compreensão. À minha irmã, que é a minha maior inspiração, por toda a ajuda e paciência. Aos meus avós, por todas as velinhas que nunca falharam. Agradeço também aos meus amigos pelos momentos felizes que tornaram este percurso mais fácil.

Resumo

A poluição do ar representa um dos principais problemas de saúde pública a nível global. Na Europa, a população passa grande parte do tempo em espaços interiores, em particular nas habitações. O dióxido de carbono (CO_2), apesar de não ser considerado um poluente, é frequentemente utilizado como indicador da qualidade das condições de ventilação, o que permite avaliar a qualidade do ar interior (QAI). Neste contexto, o presente trabalho tem como objetivo avaliar a relação entre as características habitacionais e as concentrações de CO_2 . O estudo baseou-se em dados de um questionário sobre características habitacionais aplicado a 205 famílias de oito países europeus, incluindo República Checa, Estónia, Itália, Países Baixos, Portugal, Suécia, Eslovénia e Reino Unido, bem como em medições contínuas de CO_2 obtidas através de sensores. Numa primeira análise, foram aplicados métodos de *clustering*, nomeadamente *partitioning around medoids* (PAM), hierárquico, modelos de mistura finita (FMM) e *self organizing map* (SOM), para identificar padrões estruturais nas habitações, sendo posteriormente utilizados métodos de *ensemble* para aumentar a robustez dos resultados. Numa segunda análise, a relação entre as características habitacionais e o CO_2 foi modelada através de um modelo de regressão função-sobre-escalar (FoSR), que permite descrever a evolução temporal como uma função contínua. O modelo foi ajustado aos dados horários com suavização por médias móveis (MM), sendo de seguida realizada uma análise de sensibilidade com dados de 5 em 5 minutos, com e sem aplicação da suavização. Os resultados indicam a existência de dois perfis habitacionais, nos quais as diferenças nas características construtivas e nos padrões de ventilação explicam variações nos níveis de CO_2 . Nos modelos FoSR, a presença de renovação, a tipologia da habitação, a presença de patologias de humidade e a densidade de ocupação mostraram-se associadas a essas variações, em concordância com a literatura. Os modelos apresentaram boa capacidade explicativa e os resultados mantiveram-se consistentes entre diferentes resoluções temporais e estratégias de suavização, reforçando a robustez das conclusões. Neste sentido, este estudo contribuiu para uma melhor compreensão dos fatores que influenciam a variabilidade das concentrações de CO_2 em habitações, bem como para a extensão destas metodologias a outros parâmetros da QAI em trabalhos futuros.

Palavras-chave: qualidade do ar interior, dióxido de carbono, *clustering*, métodos de *ensemble*, regressão função-sobre-escalar

Abstract

Air pollution represents one of the greatest threats to global public health. In Europe, the population spends a significant amount of time in indoor environments, mainly in residential dwellings. Carbon dioxide (CO₂), although not considered a pollutant, is frequently used as an indicator of ventilation conditions, allowing the assessment of indoor air quality. In this context, the present study aims to evaluate the relationship between housing characteristics and CO₂ concentrations. This study was based on questionnaire data on housing characteristics collected from 205 households across eight European countries, including the Czech Republic, Estonia, Italy, the Netherlands, Portugal, Sweden, Slovenia and the United Kingdom, as well as continuous CO₂ measurements obtained through sensors. In the first phase of the analysis, clustering methods were applied, namely partitioning around medoids (PAM), hierarchical clustering, finite mixture models (FMM) and self-organizing maps (SOM), to identify structural patterns in housing. Ensemble methods were subsequently used to improve the robustness of the results. In the second phase, the relationship between housing characteristics and CO₂ was modelled using function-on-scalar regression (FoSR), which allows the temporal evolution to be described as a continuous function. The model was fitted to hourly data after smoothing by moving average. A sensitivity analysis was then conducted using 5-minute interval data, with and without smoothing. The results indicate the existence of two housing profiles, in which differences in construction characteristics and ventilation patterns explain variations in CO₂ levels. In the FoSR models, the presence of renovation, dwelling typology, presence of dampness-related pathologies and occupancy density were associated with these variations, in agreement with the literature. The models showed good explanatory power and the results remained consistent across different temporal resolutions and smoothing strategies, reinforcing the robustness of the conclusions. Thus, this study contributes to a better understanding of the factors influencing the variability of CO₂ concentrations in dwellings, as well as to the extension of these methodologies to other indoor air quality parameters in future work.

Keywords: indoor air quality, carbon dioxide, clustering, ensemble methods, function-on-scalar regression

Índice

Lista de Tabelas	vi
Lista de Figuras	vii
Lista de Abreviaturas	ix
1 Introdução	1
1.1 Contexto geral	1
1.2 Qualidade do ar interior	2
1.2.1 Evolução da evidência científica	3
1.2.2 Exposição em habitações e grupos vulneráveis	4
1.2.2.1 Dióxido de carbono e fatores influenciadores	5
1.2.3 Monitorização e valores limite	7
1.3 Enquadramento metodológico	8
1.4 Objetivo	10
1.5 Organização do relatório	10
2 Métodos	12
2.1 Design do estudo	12
2.2 Caracterização dos dados e seleção de variáveis	12
2.3 Identificação de perfis habitacionais	15
2.3.1 Testes estatísticos	16
2.3.2 <i>Clustering</i>	17
2.3.2.1 <i>Partitioning around medoids</i>	17
2.3.2.2 <i>Clustering</i> hierárquico	18
2.3.2.3 Modelos de mistura finita	20
2.3.2.4 <i>Self organizing map</i>	21
2.3.2.5 Distância DAFI-Gower	23
2.3.3 Métodos de <i>ensemble</i> para <i>clustering</i>	24
2.3.4 Avaliação da qualidade dos <i>clusters</i>	26
2.3.4.1 Índices de validação interna	26
2.3.4.2 Estabilidade	27
2.3.4.3 Matrizes de dissimilaridade/semelhança	28
2.3.5 Visualização e caracterização dos <i>clusters</i>	29

2.4	Modelação do dióxido de carbono em função das características habitacionais	30
2.4.1	Pré-processamento	30
2.4.2	Regressão função-sobre-escalar	32
2.4.2.1	Representação por bases e suavização	33
2.4.2.2	Estimação via modelos generalizados aditivos	35
2.4.3	Modelação simples	36
2.4.4	Modelação múltipla	37
2.4.5	Análise de sensibilidade	40
3	Análise exploratória	42
3.1	Características habitacionais	42
3.2	Dióxido de carbono	45
4	Resultados	48
4.1	Identificação de perfis habitacionais	48
4.2	Modelação do dióxido de carbono em função das características habitacionais	53
5	Discussão e conclusão	65
	Referências Bibliográficas	70
A	Pré-processamento	77
B	Caracterização dos dados e seleção de variáveis	78
C	Análise exploratória	79
D	Identificação de perfis habitacionais	82
E	Modelação do dióxido de carbono em função das características habitacionais	85

Lista de Tabelas

4.1	Resumo das variáveis do estudo por país e respetivas medidas descritivas, incluindo valores-p de testes estatísticos ajustados pelo método de Holm.	48
4.2	Média dos valores do vetor de semelhança de Jaccard obtidos pelos métodos <i>bootstrap</i> e <i>subsetting</i> para cada método de <i>clustering</i>	50
4.3	Tamanho dos <i>clusters</i> e índices de validação interna para comparação dos métodos de <i>clustering</i> e BIC para o método FMM.	51
4.4	ARI entre o resultado do método de <i>ensemble</i> e dos métodos de <i>clustering</i> individuais.	52
4.5	Número de indivíduos por país em cada <i>cluster</i> obtido pelo método de <i>ensemble</i>	52
4.6	Medidas descritivas das variáveis com diferenças estatisticamente significativas entre os <i>clusters</i> obtidos pelo método de <i>ensemble</i>	53
4.7	Efeito das características habitacionais nos níveis de CO ₂ e respetiva interpretação.	55
4.8	Estimativas do termo constante do modelo múltiplo para os diferentes conjuntos de dados, expressas em ppm.	57
4.9	RMSE, MAE e R^2_{adj} do modelo múltiplo para os diferentes conjunto de teste.	63
D.1	Valores-p com correção de Holm dos testes estatísticos para as variáveis contínuas.	82
D.2	Coeficiente médio de silhueta e BIC para diferentes números de <i>clusters</i>	83
E.1	Sumário dos termos suaves estimados pelo modelo múltiplo para os três conjuntos de dados.	85

Lista de Figuras

1.1	Percentagem de crianças expostas a pelo menos um indicador de qualidade do ambiente interior inadequado em toda a UE-28.	5
2.2	Ilustração dos quatro métodos de <i>clustering</i>	17
3.3	Caixa de bigodes por país para as variáveis contínuas.	42
3.4	Mapa de calor para as variáveis binárias por país.	43
3.5	Gráfico de barras por país para as variáveis categóricas.	44
3.6	Caixa de bigodes do período de construção por renovação ou reabilitação do edifício e da densidade de ocupação por tipologia da habitação, por país.	45
3.7	Evolução temporal da concentração de CO ₂ por habitação, ao longo de todo o período de monitorização, antes e após a suavização por MM.	46
3.8	Evolução temporal dos níveis de CO ₂ por dia e por habitação, considerando os diferentes conjuntos de dados.	46
4.9	Mapas de calor das matrizes de dissemelhança e semelhança dos métodos de <i>clustering</i>	51
4.10	Mapas de calor das matrizes de pertença obtidas pelos métodos <i>soft euclidean</i> e <i>soft Manhattan</i> , representando a probabilidade de cada indivíduo pertencer a cada <i>cluster</i>	52
4.11	Efeito suave do tempo sobre a concentração de CO ₂ para cada <i>cluster</i>	54
4.12	Efeito suave do tempo sobre a concentração de CO ₂ estimado pelo modelo múltiplo para os diferentes conjuntos de dados.	58
4.13	Efeito suave do tempo sobre a concentração de CO ₂ estimado pelo modelo múltiplo para os diferentes conjuntos de dados para cada país em relação à categoria de referência.	59
4.14	Efeito suave do tempo sobre a concentração de CO ₂ estimado pelo modelo múltiplo para os diferentes conjuntos de dados para cada variável categórica em relação à categoria de referência.	61

4.15 Efeito suave da densidade de ocupação sobre a concentração de CO ₂ estimado pelo modelo múltiplo para os diferentes conjuntos de dados. . . .	62
4.16 QQ-plot dos efeitos aleatórios estimados pelo modelo múltiplo para os diferentes conjuntos de dados.	63
4.17 Gráficos de dispersão entre os valores observados e os valores previstos pelo modelo múltiplo, com base nos conjuntos de teste, por habitação. . .	64
A.1 Evolução temporal da concentração de CO ₂ por habitação, ao longo de todo o período de monitorização, antes da remoção das habitações com sensores descalibrados.	77
B.2 Número de observações com valores em falta por variável e para cada combinação de ausência em múltiplas variáveis.	78
C.1 Evolução temporal dos níveis de CO ₂ para cada dia de cada habitação, num conjunto de 20 habitações selecionadas aleatoriamente, considerando os diferentes conjuntos de dados.	80
C.2 ACF da concentração de CO ₂ para 20 habitações selecionadas aleatoriamente, considerando os diferentes conjuntos de dados.	81
D.1 Dendrogramas resultantes do <i>clustering</i> hierárquico com método de Ward e com ligação completa.	83
D.2 Dendrograma resultante do <i>clustering</i> hierárquico com ligação completa aplicado aos neurónios e mapa de calor da rede de neurónios, ambos referentes ao SOM.	84
D.3 <i>Clusters</i> obtidos pelos métodos PAM, hierárquico e SOM, representados nas duas primeiras dimensões após redução da dimensionalidade.	84
E.1 Avaliação dos pressupostos do modelo múltiplo: normalidade dos resíduos, homocedasticidade dos resíduos, linearidade e independência das observações.	86

Lista de Abreviaturas

ACF	FUNÇÃO DE AUTOCORRELAÇÃO
AIC	CRITÉRIO DE INFORMAÇÃO DE AKAIKE
AR	AUTORREGRESSIVO
ARI	ÍNDICE DE RAND AJUSTADO
BC	<i>BASE CLUSTERING</i>
BIC	CRITÉRIO DE INFORMAÇÃO BAYESIANO
CO	MONÓXIDO DE CARBONO
CO ₂	DIÓXIDO DE CARBONO
COVs	COMPOSTOS ORGÂNICOS VOLÁTEIS
EDF	GRAUS DE LIBERDADE EFETIVOS
EM	<i>EXPECTATION-MAXIMIZATION</i>
FMM	MODELOS DE MISTURA FINITA
FoSR	REGRESSÃO FUNÇÃO-SOBRE-ESCALAR
GAM	MODELOS GENERALIZADOS ADITIVOS
MAE	ERRO ABSOLUTO MÉDIO
MDS	<i>MULTIDIMENSIONAL SCALING</i>
MM	MÉDIAS MÓVEIS
MV	MÁXIMA VEROSIMILHANÇA
MVRE	MÁXIMA VEROSIMILHANÇA RESTRITA
NO ₂	DIÓXIDO DE AZOTO
O ₃	OZONO
OMS	ORGANIZAÇÃO MUNDIAL DE SAÚDE
PACF	FUNÇÃO DE AUTOCORRELAÇÃO PARCIAL
PAM	<i>PARTITIONING AROUND MEDOIDS</i>
PM	PARTÍCULAS EM SUSPENSÃO
QAI	QUALIDADE DO AR INTERIOR
RI	ÍNDICE DE RAND
RMSE	ERRO QUADRÁTICO MÉDIO
SO ₂	DIÓXIDO DE ENXOFRE
SOM	<i>SELF ORGANIZING MAP</i>
UE	UNIÃO EUROPEIA

1 Introdução

A poluição do ar constitui uma das maiores ameaças ambientais à saúde pública global. A gravidade deste problema é reconhecida pela Organização Mundial de Saúde (OMS¹), que reporta a poluição do ar como responsável por cerca de 6,6 milhões de mortes em 2021 [1]. Apesar dos avanços tecnológicos e das políticas ambientais implementadas, em 2024, aproximadamente 2,0 mil milhões de pessoas ainda dependiam de combustíveis e tecnologias poluentes para cozinhar [1]. Além disso, 99% da população mundial residia em zonas onde os níveis de poluição do ar ultrapassavam os limites de exposição recomendados pela OMS [1]. No contexto deste desafio ambiental, torna-se crucial fomentar a realização de estudos que aumentem o conhecimento sobre os fatores que influenciam a qualidade do ar, de modo a suportar o desenvolvimento de estratégias baseadas na evidência, capazes de promover ambientes mais saudáveis e que reduzam os impactos adversos na saúde e no bem-estar.

1.1 Contexto geral

A poluição atmosférica pode ser classificada como poluição do ar ambiente (ou exterior) e poluição do ar interior. Ambas podem ter origem numa ampla variedade de fontes emissoras de agentes agressores, as quais se dividem em naturais e antropogénicas. As fontes naturais incluem, por exemplo, pólen, poeiras e incêndios florestais. Por sua vez, as fontes antropogénicas decorrem de atividades humanas, como a produção de eletricidade ou calor (nomeadamente através da combustão de combustíveis sólidos), o transporte rodoviário, as atividades industriais ou agrícolas, entre outras [2].

Um poluente corresponde a uma substância que, quando libertada para a atmosfera, pode provocar impactos negativos na saúde ou no equilíbrio do ecossistema, até que a mesma seja eliminada [3]. Os poluentes atmosféricos podem ser classificados em três grupos: gases inorgânicos ou orgânicos, partículas em suspensão (PM) e agentes microbiológicos [4]. Os poluentes orgânicos incluem, por exemplo, os compostos orgânicos voláteis (COVs), enquanto os poluentes inorgânicos abrangem gases como o monóxido de carbono (CO), o ozono (O₃), o dióxido de azoto (NO₂) e o dióxido de enxofre (SO₂), bem como metais. Entre os diversos poluentes atmosféricos prejudiciais à saúde, destacam-se as PM, o CO, o O₃, o NO₂ e o SO₂, considerados poluentes prioritários em

¹As siglas são apresentadas em português, exceto quando são amplamente estabelecidas na literatura científica em inglês (por exemplo, para o critério de informação Bayesiano utilizou-se BIC em vez de CIB).

termos de risco para a saúde humana [5].

Os efeitos dos poluentes na saúde variam consoante o tempo de exposição, o que pode resultar em impactos distintos quando esta ocorre por curtos períodos (horas, dias ou semanas) ou de forma prolongada (meses ou anos) [2]. A exposição de curto prazo pode provocar doenças agudas, como agravamento da asma, dificuldades respiratórias e alterações do sistema cardiovascular [2]. Em situações de elevada exposição a curto prazo, podem ainda surgir sintomas temporários, como irritação nos olhos, nariz e garganta ou desconforto torácico, bem como doenças respiratórias e cardíacas mais graves [2]. Por outro lado, a exposição prolongada aumenta o risco de desenvolver doenças crónicas e incapacitantes, podendo, em casos extremos, contribuir para a morte prematura [2].

Além do tempo de exposição, o risco associado aos poluentes atmosféricos depende igualmente de outros fatores, como a concentração dos contaminantes e as propriedades toxicológicas, que influenciam a capacidade de penetração no organismo [5]. A vulnerabilidade individual também desempenha um papel crucial, sendo crianças, idosos, grávidas e pessoas com doenças pré-existentes mais suscetíveis aos efeitos adversos da poluição do ar [5].

1.2 Qualidade do ar interior

Embora a qualidade do ar exterior tenha um impacto significativo na saúde e tenha merecido particular atenção em termos de número de estudos e implementação de políticas, a qualidade do ar interior (QAI) tem vindo a assumir crescente importância, permanecendo, no entanto, relativamente pouco estudada e legislada.

De facto, além da população passar a maior parte do tempo em espaços fechados (no interior de edifícios), grande parte das substâncias prejudiciais à saúde que comprometem a QAI provém de fontes existentes nesses mesmos ambientes, o que pode levar a níveis de poluentes superiores no interior face aos existentes no exterior [6]. Diversos estudos indicam que, dependendo das emissões com contributo no local e das condições de ventilação, as concentrações no interior podem ser entre 2 a 5 vezes superiores às concentrações do exterior, podendo atingir valores até 100 vezes mais elevados [7]. Neste contexto, a QAI depende de quatro grupos principais de fatores [6]:

1. Ambiente exterior: Engloba poluição atmosférica, condições meteorológicas e características do solo, na qual a localização do edifício (por exemplo, em áreas urba-

nas ou próximas de vias rodoviárias com tráfego intenso) influencia a composição do ar que entra no interior dos edifícios;

2. Infraestrutura do edifício: Abrange materiais de construção, sistemas de climatização e/ou ventilação e danos estruturais, que podem favorecer a libertação de poluentes e a infiltração de humidade;
3. Mobiliário e equipamentos: Inclui móveis, têxteis, eletrodomésticos e outros equipamentos eletrónicos, que podem emitir PM, COVs e outros contaminantes;
4. Atividades e comportamentos dos ocupantes e produtos de consumo: Engloba atividades domésticas, como cozinhar, limpar e abertura de janelas, e o uso de produtos emissores, como ambientadores, incensos, velas, tabaco e produtos de higiene.

Tendo em conta os inúmeros fatores influenciadores e a crescente preocupação com os efeitos da poluição em espaços interiores, torna-se necessário compreender a evolução dos estudos relacionados com a QAI.

1.2.1 Evolução da evidência científica

De acordo com Zhang *et al.* [8], desde a década de 1970 têm surgido diversos estudos sobre a QAI, nos quais foram definidos os primeiros padrões internacionais de ventilação (ASHRAE Standard 62) [9]. Durante este período, foram identificados contributos significativos dos materiais de construção para a emissão de compostos como o formaldeído [10], bem como a presença de compostos orgânicos semivoláteis em partículas do ar interior, frequentemente em concentrações superiores às do exterior [11]. Paralelamente, a OMS publicou os primeiros relatórios sobre QAI, introduzindo, em 1980, o conceito de síndrome do edifício doente, que corresponde a sintomas como irritações respiratórias, dores de cabeça, fadiga e erupções cutâneas associados à permanência prolongada num determinado edifício [12]. Na década de 1990, destacou-se o primeiro estudo sobre exposições pessoais em ambientes interiores em sete cidades da Europa, o estudo European EXPOLIS [13].

Mais recentemente, a OMS tem reforçado a evidência científica através da publicação de diretrizes globais de qualidade do ar (2005 e 2021) [14, 15] e de recomendações específicas para ambientes habitacionais (2018), estabelecendo valores limite para poluentes prioritários como PM_{2,5}, PM₁₀, O₃, NO₂, SO₂ e CO [16]. Apesar destes avanços, estas diretrizes estabelecem apenas limites para cada poluente individual, com base no impacto na saúde, não considerando a exposição simultânea a vários poluentes.

Em particular, a nível nacional, existe regulamentação que estabelece a necessidade de avaliar a QAI, incluindo o cumprimento de valores limites e a realização de avaliações periódicas, contudo estas exigências aplicam-se sobretudo a edifícios de comércio, serviços e bens. Assim, não existe uma exigência equivalente de avaliação sistemática de QAI em edifícios de habitação, que permanecem relativamente negligenciados na regulamentação em vigor.

Deste modo, a revisão da literatura demonstra o crescimento do interesse pela QAI ao longo das últimas décadas e, consequentemente, evidencia a necessidade de estudos contínuos e atualizados que acompanhem a evolução dos materiais e sistemas dos edifícios, dos hábitos dos ocupantes e os diferentes perfis de consumo de energia e produtos.

1.2.2 Exposição em habitações e grupos vulneráveis

A exposição aos poluentes depende não apenas das concentrações presentes no ar, mas também do tempo de permanência nos diferentes ambientes. Deste modo, torna-se relevante identificar os locais onde os diferentes grupos populacionais passam a maior parte do tempo. Estudos realizados indicam que, na Europa, a maioria das pessoas passa cerca de 90% do tempo em espaços interiores, sendo a habitação o ambiente interior predominante [6]. Esta realidade torna as habitações particularmente relevantes para a avaliação da QAI, uma vez que fatores como materiais de construção, ventilação e hábitos domésticos influenciam diretamente as concentrações de poluentes.

Entre os grupos mais vulneráveis à poluição do ar interior destacam-se crianças e recém-nascidos, devido à imaturidade dos sistemas respiratório e imunológico, à maior taxa de inalação e ao tempo prolongado que permanecem em casa [17]. A combinação destes fatores, nessa fase precoce da vida, pode contribuir para uma maior suscetibilidade aos efeitos adversos da exposição a uma QAI inadequada, que pode aumentar o risco de desenvolvimento de doenças respiratórias, alergias e outros efeitos adversos na saúde a longo prazo. De acordo com a OMS, a inalação de partículas provenientes da poluição do ar doméstico é responsável por quase metade das mortes por infeções respiratórias inferiores em crianças com menos de cinco anos [18]. As crianças mais novas são particularmente afetadas, uma vez que passam mais tempo em casa do que as mais velhas, que dividem parte do dia entre a escola e outros espaços interiores. Além disso, nas últimas décadas, tem-se verificado uma diminuição do tempo diário que as crianças passam ao ar livre, reforçando a importância de melhorar a QAI, de forma a prevenir problemas de saúde neste grupo etário.

Em particular, segundo um estudo realizado em 2019 nos 28 países da União Europeia (UE), cerca de uma em cada três crianças, aproximadamente 26 milhões, vivia em habitações que apresentam pelo menos um dos quatro indicadores de clima interior inadequado: humidade, iluminação deficitária, ruído ou incapacidade de manter o lar adequadamente aquecido [19]. Essas crianças têm 1,7 vezes mais probabilidade de reportar problemas de saúde. A Figura 1.1 apresenta a percentagem de crianças expostas a pelo menos um indicador de clima interior inadequado para toda a UE-28 e para cada país. Portugal apresenta a maior percentagem (50,6%), indicando que uma em cada duas crianças vive em casas com condições de qualidade do ambiente interior inadequado, enquanto países como Finlândia (18,8%) e Estónia (21,4%) apresentam percentagens muito inferiores à percentagem média da UE.

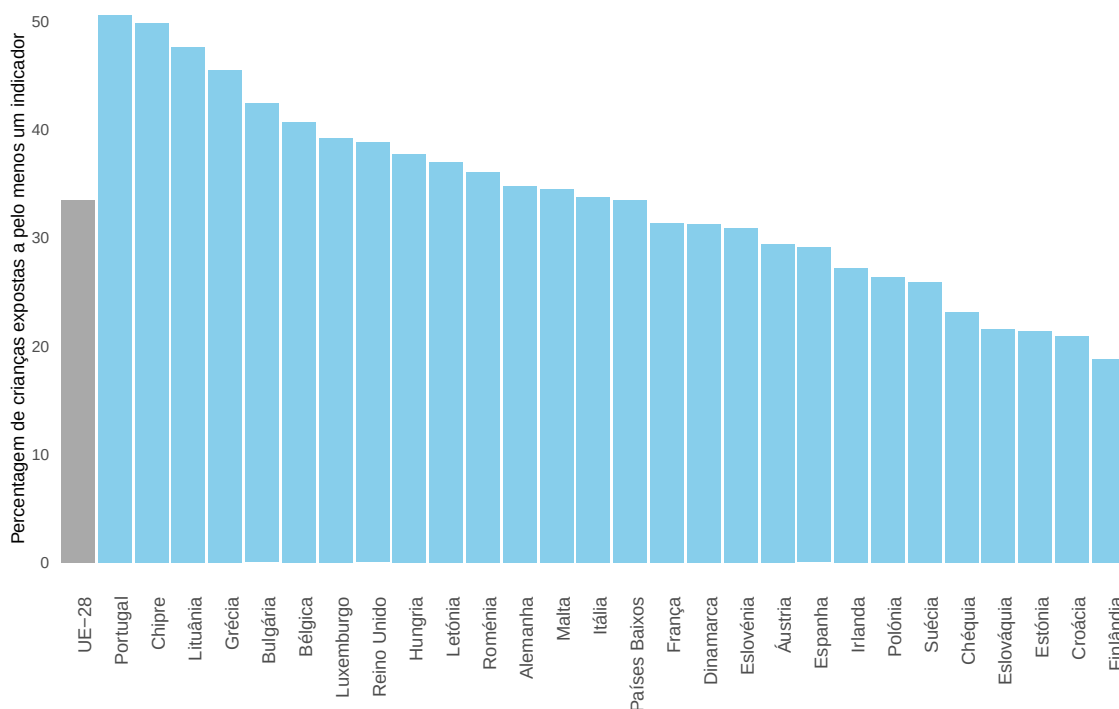


Figura 1.1: Percentagem de crianças expostas a pelo menos um indicador de qualidade do ambiente interior inadequado em toda a UE-28. Figura elaborada a partir dos dados presentes em [19].

1.2.2.1 Dióxido de carbono e fatores influenciadores

Entre os diversos parâmetros utilizados na avaliação da QAI, o dióxido de carbono (CO_2) assume particular relevância. Apesar de não ser considerado um poluente pela maioria dos autores, é amplamente utilizado como indicador da qualidade das condições de ventilação no interior de edifícios ocupados. O CO_2 é produzido principalmente pela respiração dos ocupantes e removido através da ventilação, pelo que as concentrações refletem tipicamente a relação entre a taxa de troca de ar com o exterior e a ocupação

do espaço [21]. Adicionalmente, concentrações elevadas de CO_2 podem estar associadas à acumulação de outros poluentes produzidos no interior, tornando este parâmetro particularmente útil na avaliação da QAI. Deste modo, torna-se necessário compreender os principais fatores que influenciam as concentrações de CO_2 em habitações.

A contribuição do metabolismo humano para as concentrações de CO_2 aumenta com o número de ocupantes, o nível de atividade física e o tempo de permanência no espaço. Assim, os níveis de CO_2 tendem a estar mais elevados nos quartos durante o período noturno, quando a ocupação é prolongada e a ventilação é mais limitada [20].

Além disso, a abertura e fecho de janelas torna-se um fator determinante na concentração de CO_2 , principalmente em habitações sem sistemas de ventilação mecânica. Segundo a literatura [10], a abertura de janelas tende a ser influenciada por condições térmicas interiores e exteriores, bem como pela concentração de CO_2 . Por sua vez, o fecho das janelas ocorre frequentemente em situações de desconforto térmico, especialmente durante períodos frios ou de elevada humidade. Em condições de ventilação insuficiente, as concentrações elevadas de CO_2 tendem a estar associadas à acumulação de poluentes emitidos no interior, como os COVs e PM. No entanto, quando as concentrações de PM no exterior são elevadas, estas partículas podem ser transportadas para o interior através da ventilação, mesmo quando os níveis de CO_2 permanecem reduzidos [21].

A combustão no interior das habitações também constitui uma fonte relevante de CO_2 . Em particular, o uso de gás engarrafado para cozinhar ou aquecer, bem como a utilização de lareiras e outros equipamentos de combustão, pode aumentar significativamente as concentrações de CO_2 , sobretudo em espaços com ventilação insuficiente [22]. Estudos demonstram que, durante a confeção de alimentos, ocorrem picos de curta duração nas concentrações de CO_2 , que dependem da quantidade de combustível consumido e da dimensão do espaço, podendo atingir valores próximos de 1440 ppm [22]. De forma consistente, os níveis de CO_2 em cozinhas são geralmente mais elevados quando se utilizam fogões a gás em comparação com aparelhos elétricos [20].

Para além dos padrões de ocupação, as características estruturais das habitações, como a tipologia, o ano de construção e a presença de remodelações, podem também influenciar as concentrações de CO_2 . Estudos indicam que as moradias, devido à maior dimensão e maior potencial de ventilação e infiltração natural, tendem a apresentar menor concentração de CO_2 [23]. Adicionalmente, habitações mais antigas ou não sujeitas a renovação tendem a apresentar maior permeabilidade da envolvente, o que se traduz

em menores níveis de CO₂ quando comparadas com edifícios mais recentes ou remodelados [23].

Deste modo, a definição da estratégia de ventilação adequada a cada espaço deve considerar simultaneamente a taxa de ocupação, a qualidade do ar exterior, as condições térmicas exteriores e a presença de fontes de poluentes ativas no interior. O principal desafio consiste em assegurar a diluição eficaz dos poluentes gerados no interior, minimizando, em simultâneo, a entrada de poluentes provenientes do exterior durante períodos de pico, como as horas de maior intensidade do tráfego, sem comprometer o conforto térmico nem aumentar de forma significativa as necessidades energéticas associadas ao aquecimento ou arrefecimento. Por essa razão, a monitorização do CO₂ tem vindo a ser cada vez mais integrada nos sistemas de ventilação dos edifícios.

1.2.3 Monitorização e valores limite

A definição e utilização de valores limite permitem estabelecer concentrações de poluentes acima das quais poderá existir um risco acrescido de efeitos adversos na saúde. Deste modo, a monitorização da QAI constitui uma ferramenta essencial de apoio à gestão da qualidade do ar em edifícios, à identificação da necessidade de medidas corretivas e ao estabelecimento de políticas ajustadas, contribuindo para a redução da carga global de doença associada à exposição a poluentes em ambientes interiores [15].

A avaliação da QAI requer cuidados metodológicos específicos, dado que a poluição em ambientes interiores resulta de múltiplas fontes, fortemente condicionadas pelo tipo de espaço, pelo nível de ocupação e padrões de utilização. Os métodos de monitorização podem ser passivos, recolhendo amostras que requerem análise posterior em laboratório, ou ativos, que permitem medições diretas no local. Dispositivos de medição em tempo real possibilitam o acompanhamento contínuo de parâmetros de qualidade do ar específicos, embora, em algumas situações, possam exigir equipamento fixo e operadores qualificados. Independentemente do método escolhido, fatores como a altura de instalação, a duração da medição e o registo adequado dos dados são determinantes para assegurar a fiabilidade das medições e a correta interpretação dos resultados [24].

Neste contexto, os sensores *low-cost* surgem como uma ferramenta promissora, que se caracteriza pela facilidade de utilização e monitorização em tempo real, fornecendo, geralmente, informações quantitativas sobre as concentrações de poluentes associadas a diferentes parâmetros da QAI [25]. No entanto, estes dispositivos são dependentes de

calibração específica ao local de aplicação e das condições ambientais. Além disso, a ausência de metodologias de calibração normalizadas e a variabilidade de desempenho entre dispositivos reforçam a importância de uma interpretação cautelosa dos resultados, sobretudo quando utilizados para análises quantitativas e comparação com valores de referência. Apesar de alguns estudos reportarem falhas de funcionamento e limitações na precisão destes sensores para determinados poluentes, nomeadamente PMs, as soluções atualmente disponíveis para a medição de CO₂ têm demonstrado melhor desempenho [26]. Quando assegurados critérios adequados de manutenção e calibração, os sensores *low-cost* de CO₂ tendem a evidenciar maior estabilidade e fiabilidade ao longo do tempo, sendo por isso frequentemente utilizados em estudos de QAI [26].

Para permitir a interpretação dos níveis de CO₂ nas habitações estudadas, é fundamental conhecer os valores limite estabelecidos. Apesar de existirem diversas opiniões, a maioria sugere que espaços interiores com boa qualidade do ar e adequadamente ventilados, devem apresentar concentrações inferiores a 1000 ppm [27]. Contudo, também é importante considerar a relação entre o nível de ocupação e concentração de CO₂, de modo a ajustar a taxa de ventilação ou limitar a ocupação [21].

1.3 Enquadramento metodológico

Apesar do crescente reconhecimento do papel crítico da QAI, a maioria dos estudos tem-se concentrado sobretudo na saúde e bem-estar dos ocupantes e no desempenho energético dos edifícios, frequentemente numa perspetiva centrada no conforto térmico. Contudo, à medida que os ambientes interiores se tornam mais complexos, influenciados por medidas de eficiência energética, pela introdução de novos materiais e sistemas e por padrões de ocupação em evolução, torna-se necessário avaliar os efeitos combinados de múltiplas variáveis, em vez de analisar os parâmetros de forma isolada. Ainda assim, observa-se uma predominância de análises estatísticas tradicionais (>50%), sendo menos frequente a aplicação de abordagens analíticas mais robustas, como modelos preditivos, modelos adaptativos e modelos mistos, capazes de captar relações mais complexas entre variáveis. Por outro lado, a maioria dos estudos envolve dados recolhidos apenas num país, sendo ainda escassas as análises comparativas entre múltiplos países. Consequentemente, muitos estudos apresentam aplicabilidade limitada noutros cenários, sobretudo quando não consideram as diferenças nas características dos edifícios, nas condições climáticas ou nos hábitos de utilização dos espaços interiores [28].

No contexto específico das habitações, esta lacuna é ainda mais evidente. Apenas 13,64% dos estudos foram realizados em edifícios residenciais, enquanto a maioria incidiram sobre edifícios comerciais (69,57%) [28]. Além disso, muitos estudos desenvolvidos em habitações recorrem a abordagens estatísticas univariadas ou bivariadas, testando a associação entre uma determinada característica comportamental ou habitacional e um parâmetro específico da QAI, sendo a caracterização das habitações geralmente realizada com base em medidas descritivas [17]. Embora a literatura reconheça que as características construtivas e funcionais das habitações desempenham um papel determinante na concentração de poluentes interiores, permanece reduzida a aplicação de análises comparativas das características habitacionais entre diferentes contextos geográficos antes da modelação dos parâmetros da QAI.

Alguns estudos recentes na área da QAI têm recorrido a métodos de *clustering* para identificar perfis latentes de ocupantes com base em múltiplos indicadores de saúde, conforto e preferências ambientais. Esta abordagem foi utilizada, por exemplo, por Bluysen [29] para identificar diferentes perfis de indivíduos em distintos espaços interiores e por Kim *et al.* [30] para identificar perfis de trabalhadores em escritórios com base nas suas respostas relativas à saúde e ao conforto. Em contexto escolar, Zhang *et al.* [31] aplicaram a análise de *clusters* para identificar tipologias de alunos com diferentes preferências e necessidades relativamente à qualidade do ar em salas de aula. Estes estudos demonstram o potencial do *clustering* para captar estruturas multidimensionais e padrões emergentes que dificilmente seriam detetados através de análises univariadas ou bivariadas tradicionais. Neste sentido, optou-se por aplicar métodos de *clustering* com base nas características das habitações, com o objetivo de identificar perfis habitacionais distintos. Em particular, pretendeu-se avaliar se os grupos identificados correspondem a padrões estruturais comuns, independentemente da localização geográfica, ou se são predominantemente condicionados pela distribuição das observações por país.

Após a identificação de padrões estruturais nas habitações, torna-se relevante analisar de que forma as características habitacionais influenciam os parâmetros da QAI, em particular, a concentração de CO₂. A literatura tem demonstrado que os fenómenos ambientais apresentam frequentemente relações não lineares entre variáveis explicativas e variáveis resposta, o que limita a capacidade dos modelos lineares convencionais para descrever adequadamente estes processos [32]. Neste contexto, os modelos generalizados aditivos (GAM) têm sido amplamente utilizados em estudos ambientais por permitirem representar associações não lineares através de funções suaves [32]. No entanto, estes

modelos assumem uma variável resposta escalar e estimam efeitos independentes do tempo. Quando o interesse recai sobre a evolução temporal contínua de um parâmetro e sobre respostas que tomam valores num espaço de funções, torna-se mais adequado recorrer a abordagens funcionais, como a regressão função-sobre-escalar (FoSR), que permite modelar a variável resposta como uma função no tempo. Deste modo, recorreu-se a essa abordagem para avaliar a influência das características habitacionais na evolução temporal da concentração de CO_2 .

1.4 Objetivo

O presente trabalho tem como objetivo principal identificar as características habitacionais que influenciam os níveis de CO_2 , em habitações de famílias com crianças na Europa, bem como compreender a relação entre essas características e as concentrações medidas. Os dados analisados são provenientes do projeto de investigação INQUIRE (*"Identification of chemical and biological determinants, their sources, and strategies to promote healthier homes in Europe"*) e foram estudados durante o estágio do Mestrado em Estatística Computacional e Análise de Dados, da Faculdade de Ciências da Universidade do Porto, realizado no Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial. Foram também definidos os seguintes subobjetivos:

- Descrever e comparar as características habitacionais entre os países participantes;
- Identificar padrões de agrupamento das características habitacionais através de métodos de *clustering*, permitindo a definição de perfis habitacionais distintos;
- Explorar diferenças na evolução temporal da concentração de CO_2 entre os perfis habitacionais;
- Analisar a relação entre as características habitacionais e a evolução temporal da concentração de CO_2 através de modelos FoSR;
- Avaliar a robustez dos resultados obtidos através de análises de sensibilidade, considerando diferentes níveis de agregação temporal dos dados e a aplicação de um método de suavização.

1.5 Organização do relatório

Este trabalho encontra-se organizado em seis capítulos:

- Capítulo 1 - Introdução: Introduz e enquadra o tema da qualidade do ar, com particular foco na exposição ao CO₂ em habitações com crianças, define os objetivos do estudo e apresenta a estrutura do relatório.
- Capítulo 2 - Métodos: Descreve a metodologia adotada no estudo, incluindo a caracterização do estudo e dos dados. Apresenta os fundamentos teóricos e matemáticos das abordagens utilizadas para a identificação de perfis habitacionais e para a modelação do CO₂ em função das características habitacionais, nomeadamente métodos de *clustering*, modelos funcionais e a respetiva avaliação da qualidade.
- Capítulo 3 - Análise exploratória: Apresenta a análise exploratória dos dados, com destaque para as características habitacionais e os níveis de CO₂.
- Capítulo 4 - Resultados: Apresenta os resultados obtidos na identificação de perfis habitacionais e na modelação do CO₂ em função das características habitacionais.
- Capítulo 5 - Discussão: Discute os principais resultados obtidos, confrontando-os com a literatura existente referida na introdução, bem como as limitações do estudo.
- Capítulo 6 - Conclusão: Apresenta as principais conclusões do estudo, de forma a responder aos objetivos definidos, sintetiza os principais resultados e destaca os contributos do trabalho desenvolvido.

2 Métodos

2.1 Design do estudo

Este trabalho enquadra-se no âmbito das atividades do projeto de investigação INQUIRE [33], cujo objetivo consiste em proteger a saúde dos cidadãos, através da recolha de dados sobre poluentes interiores e do desenvolvimento de ferramentas e medidas para melhorar a QAI. O projeto concentra-se particularmente em famílias com bebés e crianças pequenas (<5 anos), que passam grande parte do tempo em ambientes domésticos e, consequentemente, são mais vulneráveis à exposição a poluentes interiores. Deste modo, espera-se que o INQUIRE amplie a lista de poluentes interiores relevantes, contribua para normas regulatórias mais robustas, analise poluentes emergentes e estabeleça relações entre poluentes, produtos de consumo e materiais de construção.

O estudo baseia-se em dados recolhidos através de questionários e da monitorização da QAI em 205 habitações distribuídas por oito países europeus: República Checa, Estónia, Itália, Países Baixos, Portugal, Suécia, Eslovénia e Reino Unido. Para garantir uma comparação equilibrada entre os diferentes países, a amostra incluiu 25 habitações por país, com exceção da Eslovénia, onde foram analisadas 30 habitações. O questionário era composto por três partes: (1) questões relacionadas com o edifício, (2) questões relativas à divisão da casa em estudo e (3) questões sobre fontes exteriores. Paralelamente, foram instalados sensores nas habitações para medir as concentrações de parâmetros da QAI. A análise incidia sobre a sala de estar, divisão onde foram realizadas as medições.

2.2 Caracterização dos dados e seleção de variáveis

O questionário encontrava-se estruturado em três secções, das quais resultaram três ficheiros de dados distintos:

- DWELLING: 140 variáveis relativas às características do edifício;
- ROOM: 180 variáveis relativas às características da divisão da habitação em estudo;
- OUTDOOR: 32 variáveis relativas a fontes exteriores de poluição na envolvente do edifício.

Dado o elevado número de variáveis disponíveis, foi necessário selecionar um subcon-

junto com as mais relevantes para a análise. Durante este processo, algumas variáveis foram transformadas com o objetivo de melhorar a qualidade das mesmas. Essas transformações incluíram:

- Agrupamento de categorias para aumentar a representatividade e/ou reduzir o número de categorias;
- Conversão de variáveis aparentemente numéricas em variáveis categóricas, dado que não apresentavam um comportamento verdadeiramente contínuo;
- Criação de novas variáveis a partir da agregação de variáveis individuais com baixa representatividade ou cuja análise isolada não seria informativa;

Além disso, foram removidas variáveis preenchidas incorretamente, redundantes ou com baixa representatividade. Assim, para a análise subsequente, foram selecionadas as 31 variáveis seguintes:

- Variáveis contínuas:
 - Período de construção;
 - Densidade de ocupação:

Rácio entre a área da divisão e o número de habitantes;
 - Número de plantas/vasos;
- Variáveis binárias (Sim/Não):
 - Renovação ou reabilitação do edifício/divisão;
 - Aquecimento: Madeira/*Pellets*;
 - Aquecimento: Outros;
 - Arrefecimento: Eletricidade;
 - Cozinhar: Eletricidade
 - Cozinhar: Gás natural;
 - Ventilação/climatização: Equipamentos de combustão:

Engloba habitações que contenham pelo menos um dos seguintes equipamentos de climatização baseados em combustão: fogão a lenha, aquecedor portátil a gás e lareira;
 - Ventilação/climatização: Equipamentos elétricos:

Engloba habitações que contenham pelo menos um dos seguintes equipamentos elétricos de climatização: piso radiante/aquecido, aquecedor elé-

trico, aquecimento central e radiadores;

- Ventilação/climatização: Ar condicionado/ventilação:

Engloba habitações que contenham pelo menos um dos seguintes equipamentos de climatização e ventilação: ar condicionado, desumidificadores, humidificadores, ventilador e purificadores de ar;

- Aquecimento de água: Gás engarrafado;

- Aquecimento de água: Outros;

- Ambientadores e fragrâncias;

- Pesticidas/inseticidas;

- Animais de estimação no interior;

- Prática de fumar dentro de casa;

- Sinais físicos de patologias internas:

Engloba habitações que contenham pelo menos um dos seguintes sinais físicos de patologias nas paredes ou tetos: rachaduras, fissuras, manchas e descascamento;

- Sinais de patologias relacionadas com humidade:

Engloba habitações que contenham pelo menos um dos seguintes sinais de patologias relacionadas com humidade: mofo ou bolor, danos causados pela água, cheiro de mofo e tinta descascada nas paredes ou nos peitoris das janelas;

- Ventilação: Apenas janela(s) aberta(s);

- Ventilação: Natural:

Engloba habitações que contenham pelo menos uma das seguintes ventilações naturais: aberturas de entrada de ar na parede exterior, na janela ou em fendas no caixilho da janela;

- Exposição a tráfego;

- Exposição a indústrias;

- Exposição a atividades agrícolas;

- Exposição a atividades comerciais/recreativas;

- Outras fontes externas;

- Variáveis categóricas:

- Tipologia da habitação (Apartamento/Moradia);

- Tipo de sistema de ventilação mecânica (Controlo central/Controlo local/Nenhum);

- Localização da divisão (AB/CD/EF/GH):

Localização da divisão em estudo em relação ao número de andares da moradia/prédio. AB: moradia com um piso; CD: prédio ou moradia com vários andares e divisão junto ao telhado; EF: prédio ou moradia com vários andares e divisão junto ao rés do chão; GH: prédio ou moradia com vários andares e divisão num piso intermedio;

- Número de janelas ou portas abríveis (0/1/2/3/>4).

Importa ainda referir que o conjunto de dados resultante continha apenas 0,9% de valores omissos, sem evidência de padrões, sendo consistente com um mecanismo de valores omissos completamente aleatórios. Assim, não se justificou a aplicação de métodos de imputação (Figura B.2).

Relativamente aos dados provenientes dos sensores instalados nas habitações, estes incluíram medidas de conforto térmico, nomeadamente temperatura e humidade relativa, bem como concentrações de CO₂, PM_{2,5}, PM₁₀ e PM₁, e ainda COVs totais e taxa de renovação do ar por hora. As medições foram realizadas com um intervalo de 5 minutos, durante períodos de duração variável entre habitações, compreendidos entre 15 e 35 dias. Assim, para cada habitação, o ficheiro de dados continha os instantes de medição e as respetivas concentrações, bem como a data de início e de fim do período de monitorização. No presente trabalho, a análise centrou-se exclusivamente nas concentrações de CO₂, por se tratar de um dos principais indicadores da ventilação e, consequentemente, um parâmetro relevante para a QAI.

O conjunto de dados incluía ainda a variável país, constituída pelas categorias correspondentes aos oito países participantes no projeto (CZ: República Checa, EE: Estónia, IT: Itália, NL: Países Baixos, PT: Portugal, SE: Suécia, SI: Eslovénia, UK: Reino Unido).

2.3 Identificação de perfis habitacionais

Todas as análises estatísticas subsequentes foram realizadas no R versão 4.4.2. Numa primeira fase, a análise teve como objetivo identificar perfis habitacionais através de métodos de *clustering*, bem como avaliar de que forma esses perfis se distinguem entre os diferentes países. Os detalhes de cada método adotado são descritos nas seções seguintes.

2.3.1 Testes estatísticos

Antes da aplicação dos métodos de *clustering*, foram realizados testes estatísticos para identificar as variáveis do questionário que apresentavam diferenças estatisticamente significativas entre países, as quais constituíram o conjunto de variáveis utilizado na análise de *clustering*. Como cada observação corresponde a uma habitação distinta e estes dados foram apenas recolhidos uma vez, foram utilizados testes para amostras independentes.

Para as variáveis categóricas, aplicou-se o teste do qui-quadrado, cuja hipótese nula assume independência entre o país e a variável em análise, enquanto a hipótese alternativa pressupõe a existência de uma associação entre ambos [34]. Quando pelo menos uma das frequências esperadas era reduzida, a estatística de teste poderia não ser confiável, por isso, nessas situações, recorreu-se ao teste de Fisher [34].

Para as variáveis contínuas, utilizou-se o teste de Kruskal-Wallis, que não assume normalidade e considera como hipótese nula a igualdade das distribuições entre os grupos. Quando foram observadas diferenças estatisticamente significativas, aplicou-se o teste de Wilcoxon-Mann-Whitney para identificar especificamente os grupos que diferiam entre si (Tabela D.1).

Todos os valores-p obtidos foram ajustados com a correção de Holm para controlar o erro do tipo I decorrente das múltiplas comparações [35]. Este método consiste numa abordagem sequencial que ajusta progressivamente o nível de significância, reduzindo a probabilidade de rejeitar incorretamente a hipótese nula. Seja m o número de testes e α o nível de significância global, o algoritmo é definido da seguinte forma:

1. Ordenam-se os valores-p de forma crescente $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$, com hipóteses nulas associadas $H_{(1)}, H_{(2)}, \dots, H_{(m)}$;
2. Para $l = 1, \dots, m$, compara-se $p_{(l)}$ com $\alpha/(m - l + 1)$;
3. Regra de decisão:
 - Se $p_{(l)} \leq \alpha/(m - l + 1)$, rejeita-se $H_{(l)}$ e o procedimento continua;
 - Caso contrário, o procedimento termina e todas as hipóteses $H_{(s)}$, com $s \geq l$, não são rejeitadas.

2.3.2 Clustering

Para agrupar as habitações com base nas suas características e posteriormente analisar a composição de cada grupo por país, aplicaram-se métodos de *clustering*. Trata-se de uma classificação não supervisionada, uma vez que os agrupamentos não eram conhecidos a priori, em que o objetivo consiste na identificação de grupos de observações semelhantes entre si.

Os dados recolhidos eram de tipo misto, dado que as variáveis selecionadas incluíam tanto variáveis contínuas quanto categóricas. Assim, aplicaram-se quatro métodos de *clustering*: *partitioning around medoids* (PAM), hierárquico, modelos de mistura finita (FMM) e *self organizing map* (SOM), ilustrados na Figura 2.2.

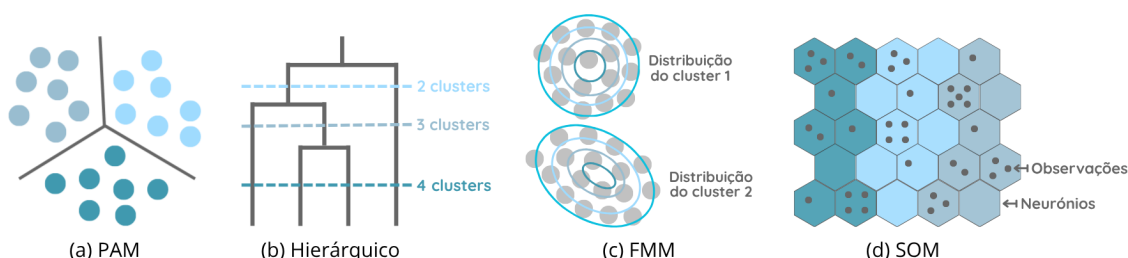


Figura 2.2: Ilustração dos quatro métodos de *clustering*. Figura elaborada com base na figura do artigo [36].

2.3.2.1 Partitioning around medoids

O algoritmo PAM baseia-se na divisão de um conjunto de observações em k grupos distintos, através da identificação de k observações representativas, designadas por medóides, de forma a minimizar a soma das dissemelhanças entre cada observação e o respetivo medóide mais próximo. Este algoritmo é composto por duas fases [37]:

- Fase BUILD: Seleciona um conjunto inicial de k medóides.
 1. Seleciona a observação mais central do conjunto de dados, cuja soma das dissemelhanças relativamente a todas as restantes observações é mínima.
 2. Seleciona sucessivamente novos medóides, escolhendo a observação que mais reduz a soma das dissemelhanças, até serem selecionados os k medóides.
 3. Cada observação é atribuída ao medóide mais próximo, formando os *clusters* iniciais.
- Fase SWAP: Procura melhorar a qualidade do agrupamento obtido na fase anterior.

1. Considera todas as trocas possíveis entre um medóide selecionado e uma observação não selecionada.
2. Para cada troca, calcula a soma das dissemelhanças de cada observação ao medóide mais próximo após a substituição, realizando essa troca se esse valor for inferior ao da solução atual.
3. O algoritmo termina quando não é possível obter uma redução adicional da soma das dissemelhanças através da substituição de medóides.

Este algoritmo assemelha-se ao *K-means* por ambos serem métodos de partição, no qual agrupam-se as observações minimizando a distância entre os pontos de um *cluster* e o respetivo centro. Contudo, diferem na forma como definem esse centro e na função objetivo associada. No *K-means*, os centros dos *clusters* (centróides) correspondem à média das observações, sendo a função objetivo baseada na minimização da soma dos quadrados das distâncias das observações ao centróide mais próximo (distância euclidiana) [38]. Em contraste, no PAM os centros (medóides) correspondem a observações reais do conjunto de dados, o que o torna mais robusto à presença de valores aberrantes [39]. A função objetivo minimiza a soma das dissemelhanças em relação ao medóide mais próximo, o que permite a utilização de diferentes medidas de dissemelhança, tornando o método mais flexível [38].

Para aplicar a metodologia, recorreu-se à função `pam()` do pacote `cluster` [40]. Esta função tem como argumentos a matriz de distâncias, o número de *clusters* e `diss = TRUE`, que indica que o objeto fornecido é a matriz de dissemelhança e não o ficheiro de dados. Para determinar o número ideal de *clusters*, o algoritmo foi aplicado para diferentes valores (2 a 8 *clusters*) e a qualidade da separação dos *clusters* foi avaliada através do coeficiente médio de silhueta, obtido a partir da função `silhouette()`. O número de *clusters* foi escolhido com base no valor máximo do coeficiente, garantindo uma melhor separação dos grupos. Uma descrição mais detalhada do coeficiente de silhueta é apresentada na Secção 2.3.4.1.

2.3.2.2 *Clustering* hierárquico

O *clustering* hierárquico divide ou agrupa os *clusters* com base numa medida de dissemelhança, construindo uma hierarquia de agrupamentos representada através de um dendrograma [39]. A construção dessa hierarquia depende do tipo de *clustering* hierárquico adotado, nomeadamente:

- *Clustering* divisivo: Considera inicialmente um único *cluster*, que contém todas as observações, e divide-o sucessivamente até que cada observação forme um *cluster* individual.
- *Clustering* aglomerativo: Considera inicialmente que cada observação tem o próprio *cluster* e agrupa sucessivamente os *clusters* mais semelhantes até existir apenas um único *cluster* com todas as observações.

Entre as duas abordagens, o *clustering* hierárquico aglomerativo é o mais utilizado na prática, tendo sido a abordagem selecionada para o presente estudo. Um dos principais desafios desse método consiste em determinar as distâncias entre *clusters*, à medida que estes passam a ser constituídos por mais do que uma observação. Nesse sentido, surgem as funções de ligação, como a ligação completa, simples, média, centróide e o método de Ward, que definem o cálculo dessa distância [39]. Neste estudo, para permitir a comparação de resultados obtidos com diferentes funções de ligação, utilizou-se a ligação completa e o método de Ward, devido à preferência das mesmas em diversos artigos [39]. Na ligação completa, a distância entre dois *clusters* corresponde ao valor máximo das distâncias entre os seus elementos, o que geralmente resulta em *clusters* mais compactos [39]. O método de Ward minimiza a variância total dentro dos *clusters*, escolhendo em cada passo os dois *clusters* cujo agrupamento aumenta o erro quadrático total o mínimo possível [41]. Matematicamente, as funções de ligação podem ser descritas pela fórmula de Lance-Williams (Equação (1)), que define a distância entre um novo *cluster* $A \cup B$ e um *cluster* C [42].

$$\text{dist}(A \cup B, C) = \alpha_A \text{dist}(A, C) + \alpha_B \text{dist}(B, C) + \beta \text{dist}(A, B) + \gamma |\text{dist}(A, C) - \text{dist}(B, C)|, \quad (1)$$

onde dist representa a dissimilaridade entre *clusters* e os parâmetros $\alpha_A, \alpha_B, \beta, \gamma$ dependem do método de ligação escolhido. Para os métodos utilizados neste estudo, os parâmetros são:

- Ligação completa: $\alpha_A = \alpha_B = 0.5, \beta = 0, \gamma = 0.5$
- Método de Ward: $\alpha_A = \frac{n_A + n_C}{n_A + n_B + n_C}, \alpha_B = \frac{n_B + n_C}{n_A + n_B + n_C}, \beta = -\frac{n_C}{n_A + n_B + n_C}, \gamma = 0,$

onde n_A, n_B e n_C representam o número de observações nos *clusters* A, B e C , respetivamente.

De forma sucinta, o *clustering* hierárquico aglomerativo começa por calcular a matriz de dissimilaridade entre todas as observações, tratando cada uma como um *cluster* individual. Em seguida, identifica os dois *clusters* mais semelhantes, agrupando-os num único *cluster*. Após cada agrupamento, atualiza as distâncias entre o novo *cluster* e os restantes com base na função de ligação. O processo repete-se até que todas as observações formem um único *cluster*.

Este método foi aplicado através da função `hclust()` do pacote `stats`, que requer como argumentos a medida de dissimilaridade e a função de ligação [43]. Para a ligação completa, utilizou-se `hclust(dist, method = "complete")`, onde `dist` representa a matriz de dissimilaridades. Para o método de Ward, aplicou-se `hclust(dist, method = "ward.D2")`, que não exige elevar a matriz de dissimilaridades ao quadrado, sendo equivalente a `hclust(dist^2, method = "ward.D")` [44]. A semelhança entre os dendrogramas obtidos pelas duas funções de ligação foi avaliada através do coeficiente de correlação cogenética, utilizando a função `cor_cophenetic()` do pacote `dendextend` [39]. Este valor varia entre -1 e 1, sendo que valores próximos de 0 refletem nenhuma semelhança estrutural.

Ao contrário do método anterior, o número de *clusters* não precisa de ser definido previamente, dado que o dendrograma resultante pode ser cortado em diferentes níveis para obter o número adequado de *clusters*. Normalmente, o corte corresponde à linha horizontal que interseja as maiores alturas do dendrograma. Cada altura representa a dissimilaridade entre os *clusters* agrupados, logo quanto maior a altura, menos semelhantes são os objetos [39]. No entanto, a decisão do número ótimo de *clusters* foi apoiada pelo coeficiente médio de silhueta.

2.3.2.3 Modelos de mistura finita

Os FMM assumem que os dados podem ser representados como uma combinação linear de um número finito de distribuições de probabilidade [45]. Matematicamente, a função de densidade de probabilidade é definida pela Equação (2):

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x} | \theta_k), \quad (2)$$

onde K representa o número de componentes da mistura, π_k as proporções de mistura, tais que $\pi_k > 0$ e $\sum_{k=1}^K \pi_k = 1$, e $f_k(\mathbf{x} | \theta_k)$ a densidade da componente k parametrizada por θ_k [45]. Cada observação provém de uma única componente da mistura e, dentro

de cada componente, as variáveis contínuas são geralmente modeladas por distribuições normais, enquanto as variáveis categóricas são modeladas por distribuições multinomiais independentes [45].

Os parâmetros do modelo (π_k e θ_k) são estimados por máxima verosimilhança (MV) através do algoritmo *Expectation-Maximization* (EM), que consiste em duas etapas iterativas [45]. Na etapa E, são estimadas as probabilidades à *posteriori* de cada observação pertencer a cada componente da mistura. Na etapa M, os parâmetros são atualizados pela maximização da verosimilhança esperada, utilizando as probabilidades obtidas na etapa anterior como pesos. Essas duas etapas são repetidas alternadamente até que seja atingido um critério de convergência, nomeadamente quando o aumento da verosimilhança entre duas iterações consecutivas for inferior a um limiar pré-definido ou quando o número máximo de iterações for alcançado.

A seleção do número adequado de componentes é realizada com base em critérios de informação, como o critério de informação Bayesiano (BIC), que permite comparar modelos com diferentes valores de K através do equilíbrio entre o ajuste aos dados e uma penalização pela complexidade do modelo [45]. Formalmente, o BIC é definido pela Equação (3), em que L representa a função de verosimilhança do modelo ajustado, p o número de parâmetros estimados e n o número de observações.

$$\text{BIC} = -2 \ln(L) + p \ln(n) \quad (3)$$

Para implementar este método, utilizou-se a função `flexmixedruns()` do pacote `fpc` [46]. Esta função tem como parâmetros as colunas correspondentes às variáveis contínuas e às variáveis categóricas, o intervalo de valores do número de componentes testado (`n.cluster = 2:8`) e o número de inicializações aleatórias do algoritmo EM (`simruns = 20`), garantindo a estabilidade da solução final. Os indivíduos com valores em falta foram excluídos da análise, dado que a função não permite incluir valores omissos e por corresponderem a um número reduzido de observações. Adicionalmente, as variáveis contínuas foram estandardizadas para evitar o impacto de diferentes escalas na estimação das componentes da mistura.

2.3.2.4 Self organizing map

No algoritmo SOM os dados são projetados num mapa, geralmente bidimensional. Esse mapa possui uma estrutura fixa, normalmente organizada como uma grelha, composta

por várias unidades (neurónios). Cada unidade (u) está associada a um vetor, designado por protótipo (p_u) ou *codebook vector*, que representa a posição dessa unidade no espaço original dos dados [47].

Na versão estocástica do algoritmo, o treino decorre de forma iterativa, onde em cada iteração é selecionada uma única observação (x_i). Em cada iteração, o algoritmo é composto por dois passos fundamentais: o *assignment step* e o *representation step*. No *assignment step*, procura-se o protótipo mais próximo dessa observação, através da minimização da distância euclidiana (Equação (4)) [47].

$$f(x_i) = \arg \min_{u=1,\dots,U} \|x_i - p_u\| \quad (4)$$

No *representation step*, os protótipos são atualizados de forma a aproximarem-se da observação selecionada, afetando também os neurónios vizinhos [47]. Esta etapa é definida pela Equação (5):

$$p_u^{\text{new}} = p_u^{\text{old}} + \mu H(d(f(x_i), u)) (x_i - p_u^{\text{old}}), \quad (5)$$

onde μ representa a taxa de aprendizagem e $H(\cdot)$ a função de vizinhança, que satisfaz as propriedades $H(0) = 1$ e $\lim_{d \rightarrow \infty} H(d) = 0$ [47].

O SOM relacional surge como uma extensão do SOM estocástico para situações em que os dados não podem ser representados como vetores num espaço Euclidiano, utilizando apenas a matriz de dissimilaridade (D). Nesta abordagem, os protótipos deixam de ser representados explicitamente no espaço dos dados e passam a ser definidos como combinações convexas das observações, descritas pelos coeficientes γ_{ui} , com $\gamma_{ui} \geq 0$ e $\sum_i \gamma_{ui} = 1$. Deste modo, o *assignment step* passa a ser expresso pela Equação (6) [47].

$$f(x_i) = \arg \min_{u=1,\dots,U} (D\gamma_u)_i - \frac{1}{2} \gamma_u^T D \gamma_u \quad (6)$$

Para auxiliar na escolha da dimensão da grelha, são frequentemente utilizados dois critérios principais: o erro de quantização e o erro topológico. O erro de quantização mede a capacidade do mapa em representar os dados originais. Formalmente, corresponde à distância média entre cada observação e o protótipo associado, onde valores mais baixos indicam uma melhor representação dos dados [47]. O erro topológico avalia a preservação da estrutura topológica dos dados no mapa, medindo a frequência com que o segundo melhor neurónio associado a uma observação não pertence à vizinhança do

neurónio atribuído [47]. Valores baixos indicam uma boa preservação da vizinhança dos dados no mapa.

Apesar de cada neurónio representar um conjunto de observações semelhantes, o número de neurónios pode ser elevado, o que dificulta a interpretação. Por este motivo, é comum agrupar os neurónios em *super-clusters* através do método de *clustering* hierárquico [47]. Neste trabalho, considerou-se o *clustering* hierárquico com ligação completa e o método de Ward. Uma vez que ambos produziram resultados iguais, na secção 4.1 serão apenas apresentados os resultados correspondentes à ligação completa.

Este método foi ajustado através da função `trainSOM()` do pacote `SOMbrero` [48], utilizando como entrada a matriz de dissemelhança, com o argumento `type="relational"`. Posteriormente, aplicou-se o *clustering* hierárquico aos neurónios recorrendo à função `superClass()`, sendo a escolha do número de *clusters* baseada na análise do dendrograma. A correspondência final entre as observações e os *clusters* foi obtida associando cada observação ao neurónio a que ficou atribuída no mapa, permitindo identificar o *cluster* correspondente a cada elemento do conjunto de dados.

2.3.2.5 Distância DAFI-Gower

Dada a presença simultânea de variáveis contínuas e categóricas, recorreu-se à distância DAFI-Gower como medida de dissemelhança nos métodos de *clustering* PAM, hierárquico e SOM. Esta distância constitui uma adaptação da distância de Gower, proposta por D’Orazio, desenvolvida para equilibrar as contribuições de variáveis contínuas e categóricas [49]. De facto, na distância de Gower, as variáveis categóricas podem influenciar desproporcionalmente a distância, uma vez que qualquer diferença numa variável categórica resulta na atribuição da distância máxima de 1, o que pode reduzir a relevância de diferenças mais subtis nas variáveis contínuas.

Com o objetivo de mitigar este problema, na distância DAFI-Gower, a dissemelhança associada à variável j entre duas observações x_i e x_k é definida pela Equação (7). Em particular, para as variáveis contínuas ($j \in \{1, 2, \dots, h\}$), a dissemelhança é calculada através da distância de Manhattan normalizada pelo intervalo interquartil. Para cada variável categórica ($j \in \{h + 1, \dots, p\}$), as categorias são convertidas em *dummies*, sendo a dissemelhança obtida pela soma das distâncias entre observações para cada *dummy* dividida pelo número total de *dummies* (L).

$$d_j(x_{ji}, x_{jk}) = \begin{cases} \frac{|x_{ji} - x_{jk}|}{\text{IQR}_j}, & \text{se } j \in \{1, 2, \dots, h\} \\ \frac{\sum_{m=1}^L I(x_{(j[m])i} \neq x_{(j[m])k})}{L}, & \text{se } j \in \{h+1, h+2, \dots, p\} \end{cases} \quad (7)$$

Para garantir uma contribuição equilibrada de variáveis contínuas e categóricas, é introduzido um fator de ajustamento (α_j), que corresponde à razão entre a média das dissimilaridades das variáveis contínuas e a dissimilaridade total da variável categórica. Quando a dissimilaridade total da variável categórica é maior que a média das dissimilaridades das variáveis contínuas ($\alpha_j < 1$), a influência da variável categórica na distância final é reduzida. Além disso, a importância de cada variável é incorporada através de pesos (w_j) derivados da informação mútua normalizada, para garantir que variáveis mais informativas tenham maior influência no cálculo da distância. Deste modo, a distância DAFI-Gower entre duas observações x_i e x_k é dada pela Equação (8).

$$\text{DAFI}(x_i, x_k) = \frac{1}{p} \sum_{j=1}^p \alpha_j \cdot w_j \cdot d_j(x_{ji}, x_{jk}), \quad (8)$$

A distância DAFI-Gower foi implementada através do código em R disponibilizado pelos autores em [50]. No entanto, as funções não são compatíveis com a presença de valores omissos, pelo que os indivíduos com dados em falta foram excluídos da análise.

É relevante mencionar que, numa fase preliminar, considerou-se a distância de Gower, contudo, a distância DAFI-Gower revelou-se mais adequada à estrutura dos dados, conforme indicado pelas métricas internas de validação.

2.3.3 Métodos de *ensemble* para *clustering*

Diferentes métodos de *clustering* podem produzir soluções distintas para o mesmo conjunto de dados, devido a diferenças na definição de semelhança, nos critérios de otimização e na escolha de parâmetros. Consequentemente, a seleção de um único método pode não ser suficiente para capturar de forma robusta a estrutura dos dados. Os métodos de *ensemble* para *clustering* tentam ultrapassar esta limitação, combinando diferentes soluções de *clustering*. Esta abordagem permite identificar os padrões de agrupamento que são consistentes entre os métodos e reduzir a influência de particularidades de cada um. Deste modo, os métodos de *ensemble* apresentam diversas vantagens [51]:

- Robustez: desempenho superior em comparação com algoritmos de *clustering* in-

dividuais;

- Novidade: soluções combinadas que não seriam acessíveis a qualquer algoritmo individual;
- Estabilidade: menor sensibilidade a ruído e a valores extremos.

Formalmente, cada solução de *clustering* obtida a partir de um método ou parametrização específica é designada por *base clustering* (BC). Cada BC (b) pode ser representada por uma matriz de pertença (M_b), onde cada linha corresponde a uma observação e cada coluna a um *cluster*. No caso de *hard clustering*, estas matrizes são binárias, ou seja, cada observação pertence exclusivamente a um único *cluster* [52]. No caso de *soft clustering*, cada observação pode pertencer simultaneamente a vários *clusters*, sendo representada por graus de pertença no intervalo $[0, 1]$, com soma unitária por linha [52].

Para combinar os diferentes BCs, utiliza-se uma função de consenso, geralmente formulada como um problema de otimização, que procura a solução que maximize a semelhança ou minimize a dissemelhança em relação aos BCs [51]. Embora existam outras abordagens, como a construtiva e a axiomática, são menos comuns [52]. No presente trabalho, para calcular a dissemelhança entre os BCs, utilizou-se a distância euclidiana e a distância de Manhattan. Em termos matemáticos, pode ser expresso pela Equação (9):

$$\min_{M, \Pi_1, \dots, \Pi_B} \sum_b w_b \|M - M_b \Pi_b\|^p, \quad (9)$$

onde M é a matriz de consenso, M_b a matriz de pertença de cada BC, w_b o peso de cada BC e Π_b as permutações que alinham os *clusters* entre os BCs [52]. O expoente p define a métrica utilizada: $p = 2$ corresponde à distância euclidiana, enquanto $p = 1$ corresponde à distância de Manhattan.

Isto foi realizado com a função `cl_consensus()` do pacote *clue* [52]. Como os métodos de *clustering* utilizados produzem soluções *hard*, a função foi configurada para gerar também uma solução *hard*, através dos parâmetros `method = "HE"` e `method = "HM"`, de forma a tornar o resultado diretamente comparável às soluções originais. Os métodos *soft*, correspondentes a `method = "SE"` e `method = "SM"`, foram utilizados apenas para avaliar a incerteza na atribuição das observações. No entanto, os métodos *soft* também podem ser convertidos em soluções *hard*. Adicionalmente, para avaliar o grau de incerteza na atribuição das observações aos *clusters*, calculou-se o coeficiente de partição

(função `cl_fuzziness()`). Esse coeficiente varia entre 0 e 1, onde valores mais baixos indicam maior incerteza.

Importante referir que todos os métodos de *clustering* foram considerados com o mesmo peso, uma vez que não existe uma métrica única que permita comparar diretamente os métodos, especialmente devido à utilização do FMM, que não permite o cálculo de métricas internas. Diferentes critérios de qualidade (índices internos, estabilidade ou concordância) avaliam aspetos distintos da estrutura dos dados, pelo que seria necessário um método capaz de atribuir pesos considerando vários critérios em simultâneo.

2.3.4 Avaliação da qualidade dos *clusters*

2.3.4.1 Índices de validação interna

Nos métodos de *clustering* PAM, hierárquico e SOM, a qualidade do agrupamento foi avaliada com base em índices de validação interna. Estas métricas não requerem rótulos externos, pelo que permitem avaliar a qualidade de uma solução de *clustering* com base na coesão e separação dos *clusters* [53].

Entre os índices utilizados, destaca-se a silhueta, que mede o grau de semelhança de uma observação ao *cluster* ao qual foi atribuída, em comparação com o *cluster* alternativo mais próximo. Assim, a largura de silhueta associada à observação i é dada pela Equação (10):

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}, \quad (10)$$

onde a_i representa a dissemelhança média entre a observação i e as restantes observações do *cluster* ao qual pertence e b_i corresponde à menor dissemelhança média entre i e as observações de qualquer outro *cluster*. Os valores variam entre -1 e 1, sendo que valores próximos de 1 indicam boa separação, valores próximos de 0 sugerem que a observação se encontra na fronteira entre *clusters* e valores negativos apontam para uma possível má classificação. Este coeficiente foi calculado através da função `silhouette()`.

Além da silhueta, foi utilizado o índice de Dunn, que avalia simultaneamente a separação inter-*clusters* e a coesão intra-*clusters*. Este índice é definido como a razão entre a menor dissemelhança observada entre dois *clusters* e a maior dissemelhança observada dentro de um *cluster*. Assim, valores mais elevados indicam melhor qualidade do agrupamento, correspondendo a *clusters* bem separados e internamente compactos.

Adicionalmente, recorreu-se ao índice de Davies-Bouldin, que mede a qualidade do agrupamento através da relação entre a dispersão interna dos *clusters* e a sua separação relativamente aos restantes *clusters*. Este índice é definido pela Equação (11):

$$DB = \frac{1}{k} \sum_{c=1}^k \max_{p \neq c} \frac{S_c + S_p}{d(C_c, C_p)}, \quad (11)$$

onde k representa o número de *clusters*, c e p dois *clusters* distintos, S_c e S_p a disseminação média das observações desses *clusters* ao respetivo centróide e $d(C_c, C_p)$ a disseminação entre os centróides desses *clusters*. Desta forma, valores baixos indicam *clusters* compactos e bem separados, enquanto valores elevados sugerem sobreposição ou dispersão interna elevada. Os dois índices mencionados foram obtidos pela função `cluster.stats()` do pacote `fpc`.

2.3.4.2 Estabilidade

A estabilidade avalia a robustez dos *clusters* face a pequenas alterações nos dados. *Clusters* estáveis permanecem inalterados quando se aplicam perturbações, como a seleção de subconjuntos, a adição de valores aberrantes ou pequenas alterações aleatórias nos valores das observações [54].

Para avaliar a estabilidade das soluções de *clustering* encontradas, recorreu-se à função `clusterboot()` do pacote `fpc`. Como o *clustering* foi realizado com base numa matriz de disseminação externa (distância DAFI-Gower), a matriz foi fornecida diretamente à função em vez dos dados originais. Nestes casos, a função encontra-se limitada aos métodos de reamostragem *bootstrap* e *subsetting*, dado que não é possível reconstruir ou perturbar os dados originais. Ambos seguem o mesmo procedimento geral [54]:

1. Seleção de índices aleatórios:
 - *Bootstrap*: escolhe n índices com reposição (n = número de observações);
 - *Subsetting*: escolhe m índices sem reposição (normalmente $m = n/2$);
2. Extração da submatriz de disseminação correspondente aos índices selecionados;
3. Aplicação do método de *clustering* à submatriz;
4. Comparação dos *clusters* obtidos com os *clusters* do conjunto completo, utilizando o coeficiente de Jaccard;
5. Repetição do processo B vezes.

Este processo resulta num vetor de coeficientes de semelhança de Jaccard, na qual cada valor representa a proporção de observações que pertencem a ambos os conjuntos, relativamente a todas as observações que pertencem a pelo menos um deles. Para permitir a comparação da estabilidade entre os diferentes métodos de *clustering*, calculou-se a média dos coeficientes. Valores próximos de 1 indicam conjuntos quase idênticos, enquanto valores próximos de 0 revelam pouca ou nenhuma sobreposição entre os conjuntos. O número de repetições não precisa de ser elevado, sendo que, no caso do coeficiente de Jaccard, uma estimativa bastante precisa pode ser obtida com apenas 50 repetições [54].

2.3.4.3 Matrizes de dissemelhança/semelhança

As matrizes de semelhança e dissemelhança permitem avaliar a semelhança e a diferença entre os *clusters* obtidos por diferentes métodos, sendo a matriz de dissemelhança o complemento unitário da matriz de semelhança. Neste sentido, recorreu-se à função `cl_dissimilarity()` do pacote `clue` para calcular as matrizes de dissemelhança associadas à distância euclidiana, à distância de Manhattan e ao índice de Rand (RI) [55]. Adicionalmente, a matriz de semelhança baseada no índice de Rand ajustado (ARI) foi obtida através da função `adjustedRandIndex()` do pacote `mclust`. As medidas utilizadas podem ser agrupadas em dois tipos: distâncias baseadas em matrizes de pertença, que incluem as distâncias euclidiana e de Manhattan, e medidas baseadas em pares de observações, que incluem o RI e o ARI, na qual cada uma se caracteriza da seguinte forma:

- Distância euclidiana: Raiz da soma dos quadrados das diferenças entre as matrizes de pertença;
- Distância de Manhattan: Soma dos valores absolutos das diferenças entre as matrizes de pertença;
- RI: Proporção de pares de observações classificadas de forma igual nas duas soluções. Considerando um conjunto de n observações, sendo a o número de pares de observações que pertencem ao mesmo *cluster* e b o número de pares que pertencem a *clusters* diferentes em ambas as soluções, define-se $RI = (a + b) / \binom{n}{2}$. Em particular na função utilizada, o índice corresponde a $1 - RI/2$, assumindo valores entre 0 (soluções idênticas) e 0,5 (discordância máxima).
- ARI: Corrige a concordância esperada ao acaso, sendo definido por $ARI = (RI -$

$\mathbb{E}[\text{RI}]/(\max(\text{RI}) - \mathbb{E}[\text{RI}])$ [56]. Assume o valor 1 quando as soluções são idênticas, valores próximos de 0 quando a concordância é semelhante à esperada ao acaso e valores negativos quando a concordância é inferior ao acaso.

No cálculo de algumas matrizes, nomeadamente as distâncias euclidiana e de Manhattan, é necessário considerar que os rótulos dos *clusters* podem ser permutados sem alterar o agrupamento subjacente, ou seja, procura a melhor correspondência de *clusters* entre dois métodos e calcula a diferença mínima [55].

2.3.5 Visualização e caracterização dos *clusters*

Para os métodos PAM, hierárquico e SOM, a representação gráfica dos *clusters* foi realizada utilizando a função `fviz_cluster()` do pacote `factoextra`. Quando aplicada diretamente aos dados originais, esta função recorre à análise em componentes principais para reduzir a dimensionalidade e representar os dados em duas dimensões [39]. No entanto, como temos dados mistos, a análise em componentes principais não é aplicável, uma vez que pressupõe variáveis numéricas e uma estrutura euclidiana. Além disso, os resultados do *clustering* baseiam-se numa matriz de dissemelhanças, não sendo utilizados diretamente os dados originais. Por esse motivo, recorreu-se ao *Multidimensional Scaling* (MDS) clássico, através da função `cmdscale()` do pacote `stats`, que reduz a dimensionalidade enquanto preserva as dissemelhanças entre as observações [57, 58]. A matriz de dissemelhanças é transformada numa matriz de produtos internos centrados, cuja decomposição espectral fornece as coordenadas das observações. As dimensões do MDS correspondem aos vetores próprios associados aos maiores valores próprios positivos dessa matriz [59]. No método FMM esta representação não foi realizada, uma vez que se trata de um modelo probabilístico que não se baseia numa matriz de dissemelhanças.

Para caracterizar os *clusters* formados pelos métodos de *ensemble*, foi construída uma tabela de contingência entre os *clusters* e os países, que permitiu a análise da distribuição dos países por *cluster*. Além disso, elaborou-se uma tabela contendo, para cada *cluster*, a categoria mais frequente de cada variável categórica, e a respetiva frequência relativa, bem como a média (desvio-padrão) e a mediana (intervalo interquartil) das variáveis numéricas, permitindo compreender que características predominam em cada *cluster*. Adicionalmente, a tabela inclui os valores-p ajustados pela correção de Holm, resultantes dos testes estatísticos qui-quadrado, de Fisher ou de Kruskal-Wallis, para avaliar a

existência de diferenças estatisticamente significativas na distribuição das variáveis entre os *clusters*.

2.4 Modelação do dióxido de carbono em função das características habitacionais

Numa segunda parte, o objetivo consistiu em avaliar a relação entre as características habitacionais e as concentrações de CO₂. Para modelar esta relação, foram utilizados modelos de regressão funcional do tipo FoSR. Esta abordagem permitiu quantificar o efeito das características habitacionais na evolução temporal das concentrações de CO₂ e identificar os fatores mais relevantes para explicar a variabilidade observada. Os detalhes metodológicos são apresentados nas seções seguintes.

2.4.1 Pré-processamento

Relativamente aos dados do CO₂, que incluíam medições de 5 em 5 minutos ao longo de vários dias, foram excluídos os registos correspondentes ao período pós-intervenção, uma vez que o presente estudo centra-se exclusivamente na caracterização da QAI antes de qualquer intervenção realizada pela equipa de investigação. Além disso, nas habitações com múltiplos valores para o mesmo instante temporal, manteve-se apenas a primeira medição registada.

De seguida, realizou-se um alinhamento temporal, para permitir a comparação do comportamento do CO₂ ao longo do tempo entre as diferentes habitações, ainda que tenham iniciado as medições em datas e horários distintos. Para cada habitação, definiu-se o instante zero como as 00:00 horas do dia da primeira medição e calculou-se o tempo decorrido de forma contínua, em minutos, sem reiniciar ao final de cada dia.

Através da representação gráfica das curvas de CO₂ ao longo do tempo, identificaram-se duas habitações com valores muito superiores às restantes, cuja causa mais provável será o descalibramento dos sensores (Figura A.1). Assim, estas habitações foram removidas da análise.

Além disso, observaram-se habitações em que os valores de CO₂ permaneciam muito baixos e com variabilidade reduzida durante períodos superiores a um dia, o que poderá indicar ausência dos habitantes. Para evitar enviesamentos decorrentes dessas situações, as respetivas medições foram substituídas pela média da habitação correspon-

dente, excluindo as medidas que serão imputadas. Esses períodos foram identificados através da seleção de períodos com mais do que um dia com variabilidade inferior a 1% da variabilidade da habitação. Adicionalmente, para garantir uma série temporal completa dentro de cada habitação, os instantes sem medições de CO₂ foram preenchidos com NAs e, posteriormente, substituídos pela média da respetiva habitação.

Verificaram-se ainda diferenças na duração dos períodos de monitorização entre habitações, o que poderia comprometer a comparabilidade das curvas. Assim, optou-se por restringir a análise ao intervalo temporal comum a todas as habitações. O menor número de observações registado foi de 4735. No entanto, para garantir um período totalmente sobreposto entre todas as curvas, foi definido um subconjunto de 4032 observações, correspondente a 14 dias de monitorização. Desta forma, para todas as habitações foi considerado o intervalo entre $t = 2880$ minutos e $t = 23040$ minutos.

Para reduzir o ruído de curto prazo presente nas curvas de CO₂, aplicou-se uma suavização através de médias móveis (MM) com janelas de 12 períodos de 5 minutos (1 hora) e de 96 períodos de 5 minutos (8 horas), substituindo cada observação pela média das observações anteriores dentro da respetiva janela temporal. A escolha da janela de 96 observações foi motivada pela análise da função de autocorrelação (ACF), que evidenciou autocorrelação significativa em torno do lag 96 em grande parte das habitações, sugerindo a existência de padrões intra-diários com periodicidade aproximada de 8 horas.

Posteriormente, utilizou-se a decomposição sazonal e de tendência por *Loess*, definida por $y_t = T_t + S_t + R_t$, para separar as curvas de CO₂ (y_t) nas componentes de tendência (T_t), sazonalidade (S_t) e resíduo (R_t). Considerando que as medições foram registadas a cada 5 minutos, definiu-se uma frequência temporal de 288 observações por dia, correspondente a um ciclo sazonal diário de 24 horas. De forma complementar, calculou-se a força da componente sazonal, definida como $1 - \frac{Var(R_t)}{Var(S_t + R_t)}$, na qual valores próximos de 1 indicam uma maior predominância da componente sazonal face ao ruído residual.

Observou-se que a aplicação de MM aumentou progressivamente a definição da componente sazonal, sendo a janela de 96 a que melhor preservou o padrão sazonal diário, ao mesmo tempo que reduziu o ruído e os picos extremos. Deste modo, o conjunto de dados obtido através de MM com janela 96 foi o utilizado para as análises subsequentes.

Dado que as curvas de CO₂ apresentaram forte dependência temporal intra-diária, optou-

se por uma agregação temporal dos dados que preservasse a variabilidade diária, permitindo simultaneamente reduzir a influência de oscilações de curta duração. Nesse sentido, os valores foram agrupados por hora, calculando-se a média horária para cada habitação em cada dia. Desta forma, cada dia passou a ser representado por uma curva diária de CO₂ composta pelas respetivas médias horárias.

2.4.2 Regressão função-sobre-escalar

A análise de dados funcionais constitui um ramo da estatística dedicado ao estudo de dados funcionais, em que cada observação é representada por uma função definida num domínio contínuo, como, por exemplo, uma curva observada ao longo do tempo, idade, ou qualquer outra variável contínua que corresponda ao índice da característica em estudo.

Dependendo da natureza da variável resposta e das variáveis explicativas, podem distinguir-se três modelos de regressão funcional:

- Regressão escalar-sobre-função: A resposta é escalar e os preditores são funcionais;
- Regressão função-sobre-escalar: A resposta é funcional e os preditores escalares;
- Regressão função-sobre-função: A resposta e os preditores são funções.

No presente trabalho, o principal interesse recai sobre os modelos FoSR. Neste caso, o modelo pode ser definido pela Equação (12):

$$y_i(t) = \beta_0(t) + \sum_{j=1}^p \beta_j(t)x_{ij} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad (12)$$

onde $y_i(t)$ corresponde à função resposta do indivíduo i , x_{ij} ao valor da j -ésima covariável escalar para o indivíduo i , $\beta_j(t)$ ao coeficiente funcional associado à covariável x_{ij} e ε_i corresponde ao erro aleatório. Deste modo, ao contrário da regressão linear clássica, em que os coeficientes são constantes, nos modelos FoSR os coeficientes são funções definidas ao longo do domínio considerado (t). Assim, cada coeficiente funcional descreve a forma como o efeito da respetiva covariável altera a resposta para cada t .

Apesar de ser possível ajustar modelos de regressão linear de forma independente em cada instante do domínio, tal abordagem não explora diretamente a estrutura temporal na estimação dos efeitos, não impõe condições de suavidade ao longo do domínio e ignora

potenciais correlações entre os resíduos observados dentro do mesmo indivíduo [60]. Estas limitações motivam a utilização de abordagens funcionais, nas quais os efeitos das covariáveis são modelados como funções suaves do domínio de interesse.

2.4.2.1 Representação por bases e suavização

A estimação dos coeficientes funcionais é geralmente realizada através da combinação linear de elementos de uma base de funções, tais como polinómios, funções de Fourier ou *splines* [61]. Deste modo, o coeficiente $\beta_j(t)$ é definido pela Equação (13), onde $b_{jl}(t)$ representa os elementos da base de funções e β_{jl} os coeficientes a estimar [61].

$$\beta_j(t) = \sum_{l=1}^{L_j} \beta_{jl} b_{jl}(t) \quad (13)$$

Entre as bases de funções mais utilizadas encontram-se as *splines*, que se destacam pela flexibilidade e boas propriedades de aproximação [62]. Uma *spline* corresponde a uma função definida por segmentos polinomiais ligados por pontos (nós), na qual são impostas condições de continuidade entre segmentos adjacentes [62]. Mais especificamente, os polinómios de ordem m são unidos de forma a garantir continuidade da função e das derivadas até à ordem $m - 1$ [61].

Apesar de existirem vários tipos de *splines*, como *B-splines*, *splines* cúbicas ou *thin plate splines*, neste trabalho foram utilizadas *splines* cúbicas cíclicas penalizadas. Esta escolha impõe condições de periodicidade ao longo do domínio, garantindo continuidade da função e das primeiras duas derivadas entre o início e o fim do intervalo [63]. Esta restrição é particularmente apropriada para variáveis com comportamento intrinsecamente cíclico, como a concentração de CO₂ ao longo do dia, uma vez que o fim de um dia e o início do seguinte correspondem a instantes temporalmente consecutivos. Desta forma, evitam-se descontinuidades artificiais nas extremidades do domínio e obtêm-se curvas suavizadas mais coerentes com a natureza periódica do processo.

A flexibilidade do modelo depende tanto da ordem dos polinómios como do número dos nós, sendo que um maior número de nós aumenta a capacidade de adaptação do modelo às variações dos dados [61]. Contudo, uma base demasiado flexível pode originar ajustamentos excessivamente irregulares e, consequentemente, conduzir ao sobreajuste do modelo [61].

Para evitar este problema, introduz-se um termo de penalização que controla a suavidade

das funções estimadas. Assim, no caso gaussiano, a estimação dos coeficientes resulta na minimização da Equação (14).

$$\min_{\{\beta_{jl}\}} \left\{ \sum_{i=1}^n \left\| y_i(t) - \beta_0(t) - \sum_{j=1}^p x_{ij} \sum_{l=1}^{L_j} \beta_{jl} b_{jl}(t) \right\|^2 + \sum_{j=1}^p \lambda_j \int (\beta_j''(t))^2 dt \right\} \quad (14)$$

O primeiro termo corresponde à qualidade do ajustamento do modelo aos dados observados e o segundo termo impõe uma penalização de suavidade aos coeficientes funcionais [61]. O termo $\int (\beta_j''(t))^2 dt$ mede a curvatura das funções estimadas e o parâmetro de suavização $\lambda_j \geq 0$ controla o peso atribuído a essa penalização. Valores elevados de λ_j resultam em funções mais suaves, enquanto valores reduzidos permitem maior flexibilidade [61].

O parâmetro de suavização λ_j pode ser estimado através da minimização do erro de previsão ou da maximização da verosimilhança [64]. Neste trabalho, utilizou-se a segunda abordagem, na qual os termos suaves são interpretados como efeitos aleatórios, o que permite estimar o parâmetro de suavização por MV ou máxima verosimilhança restrita (MVRE).

O problema de estimação pode ser escrito na forma matricial através da Equação (15), onde S_j representa a matriz de penalização associada ao j -ésimo termo suave, construída a partir dos elementos da base de funções, e β o vetor que contém todos os coeficientes β_{jl} [64, 65].

$$\min_{\beta} \left\{ \|y - X\beta\|^2 + \sum_{j=1}^p \lambda_j \beta^T S_j \beta \right\} \quad (15)$$

Desta forma, assumindo que os coeficientes associados às componentes suaves são variáveis aleatórias, tem-se que $\beta \sim \mathcal{N}(0, \theta S^-)$, em que $S = \sum_{j=1}^p \lambda_j S_j$, e $\lambda_j = \sigma^2 / \theta_j$, ou seja, o parâmetro de suavização é inversamente proporcional à variância dos efeitos aleatórios [64]. Consequentemente, a seleção dos parâmetros de suavização pode ser reformulada como um problema de estimação da variância.

A maximização da verosimilhança é conseguida por integração dos coeficientes aleatórios, resultando na verosimilhança marginal, descrita na Equação (16), onde $p(y | \beta)$ corresponde à distribuição dos dados condicionada aos coeficientes e $p(\beta | \lambda)$ à distribuição normal assumida para os efeitos aleatórios [64].

$$p(y | \lambda) = \int p(y | \beta) p(\beta | \lambda) d\beta, \quad (16)$$

Esta integração resulta numa verosimilhança marginal que depende apenas dos parâmetros de suavização λ . Assim, esses parâmetros são estimados por maximização da função resultante (Equação (17)), sendo necessário reestimar os coeficientes do modelo para cada valor de λ [64].

$$\hat{\lambda} = \arg \max_{\lambda} \log p(y | \lambda) \quad (17)$$

Por sua vez, a MVRE baseia-se numa verosimilhança modificada, obtida após eliminar a contribuição dos efeitos fixos do modelo antes da estimação das componentes de variância. Este procedimento corrige o viés associado à estimação por MV, conduzindo geralmente a estimativas mais estáveis dos parâmetros de suavização [64]. Por esta razão, o método MVRE é frequentemente recomendado e foi o utilizado neste trabalho.

2.4.2.2 Estimação via modelos generalizados aditivos

Os modelos FoSR podem ser interpretados no contexto dos GAM para o caso em que a variável resposta é funcional. Enquanto os GAM modelam a relação entre uma resposta escalar e covariáveis através de efeitos lineares e funções suaves aplicadas às covariáveis, no FoSR os efeitos das covariáveis são representados por coeficientes funcionais definidos ao longo de um domínio contínuo.

Os GAM constituem uma extensão dos modelos lineares generalizados (GLM), que permitem modelar relações não lineares entre a variável resposta e os preditores através de funções suaves [62]. De forma geral, um modelo GAM pode ser representado pela Equação (18):

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \sum_j f_j(x_{ij}) + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad (18)$$

onde y_i representa a variável resposta para a observação i , β_0 é o termo independente e β_k são os coeficientes associados às covariáveis lineares x_{ik} , na qual estes dois últimos correspondem à componente paramétrica do modelo. O terceiro termo representa a componente não paramétrica, constituída por funções suaves $f_j(\cdot)$ aplicadas às covariáveis x_{ij} . O termo ε_i corresponde ao erro aleatório, assumidos independentes e identicamente distribuídos com distribuição normal de média zero e variância constante σ^2 .

Em particular, os modelos GAM permitem estimar modelos penalizados como os descritos na secção anterior, nos quais os efeitos funcionais são representados através de elementos de uma base de funções e sujeitos a uma penalização da suavidade [62].

A estimação dos parâmetros é realizada através de um método iterativo que combina a estimação dos coeficientes e dos parâmetros de suavização [64]. Para um conjunto fixo de parâmetros de suavização, os coeficientes dos modelos são estimados através do algoritmo *penalized iteratively reweighted least squares*, que consiste na resolução iterativa de um sistema de equações penalizadas equivalente ao problema de mínimos quadrados da Equação (14) [64]. Quando há convergência, os parâmetros de suavização são atualizados através da MVRE, calculados com base nos coeficientes estimados [64]. No entanto, uma vez que a verosimilhança marginal depende dos coeficientes estimados, este processo implica a reestimação dos coeficientes sempre que os parâmetros de suavização são atualizados. Este processo é repetido iterativamente até à convergência simultânea dos coeficientes do modelo e dos parâmetros de suavização.

2.4.3 Modelação simples

Considerando o conjunto de dados relativos à média horária, para uma análise exploratória inicial das tendências temporais médias da concentração de CO_2 em função das diferentes características habitacionais, foram ajustados modelos FoSR simples, com uma única variável explicativa (categórica). Para cada categoria de uma dada variável, considerou-se um modelo da forma $\text{CO}_2 = f(t) + \epsilon$, onde $f(t)$ é uma função suave do tempo. As curvas obtidas representam a evolução média suavizada da concentração de CO_2 ao longo do dia para cada categoria, permitindo identificar padrões médios de variação temporal.

De forma complementar, procedeu-se à análise das tendências temporais por cada um dos perfis habitacionais previamente identificados através dos métodos de *clustering*. Assim, para cada *cluster*, foi ajustado o modelo simples referido anteriormente, o que permitiu comparar a evolução média diária da concentração de CO_2 entre perfis habitacionais distintos.

Para a implementação destes modelos, utilizou-se a função `gam()` do pacote `mgcv` [63]. A função suave $f(t)$ foi estimada através de uma *spline* cúbica cíclica penalizada, especificada por `s(t, bs="cc", k=24)`, que impõe periodicidade ao longo de um ciclo de 24 horas. O parâmetro $k = 24$ define o número máximo de elementos da base de fun-

ções utilizados para representar a suavização (um nó por hora), permitindo flexibilidade suficiente para identificar padrões horários sem ocorrer um sobreajuste excessivo. As curvas estimadas desta forma tiveram um caráter exclusivamente descritivo e representam tendências agregadas em todas as observações pertencentes a cada categoria. Estes modelos não incorporam medições repetidas dentro das habitações, autocorrelação temporal ou potenciais interações entre covariáveis. Consequentemente, as diferenças observadas entre curvas devem ser interpretadas com cautela, constituindo sobretudo evidência exploratória que motiva uma modelação múltipla mais robusta.

2.4.4 Modelação múltipla

Após a análise exploratória, procedeu-se à modelação múltipla com o objetivo de avaliar a influência conjunta das características habitacionais nos níveis de CO₂. Neste caso, foi utilizada a função `bam()`, uma extensão da função `gam()` desenvolvida para conjuntos de dados de elevada dimensão. A estimação dos parâmetros foi realizada através de uma adaptação mais rápida do método MVRE. Adicionalmente, definiu-se `discrete = TRUE`, que discretiza as covariáveis contínuas, reduzindo significativamente o tempo de estimação sem comprometer a qualidade do ajustamento [63].

O processo de modelação foi realizado de forma sequencial, sendo que, em cada etapa, os modelos encaixados foram comparados através do critério de informação de Akaike (AIC) definido pela Equação (19), onde L corresponde à verosimilhança do modelo e k ao número de parâmetros. Os graus de liberdade efetivos (EDF) substituem o número de parâmetros devido à presença de termos suaves penalizados [63].

$$\text{AIC} = -2\log(L) + 2k \quad (19)$$

Adicionalmente, a qualidade do ajustamento foi avaliada através do R^2 ajustado, obtido pela função `summary()`. Esta medida quantifica a proporção de variabilidade explicada pelo modelo, dada pela Equação (20) [66]:

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}}{\text{TSS}} \cdot \frac{n-1}{n - \text{EDF}_{\text{total}}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \cdot \frac{n-1}{n - \text{EDF}_{\text{total}}}, \quad (20)$$

sendo que RSS representa a soma dos quadrados dos resíduos, TSS a soma total dos quadrados (variabilidade total da variável resposta), y_i os valores observados da variável resposta na unidade i , $\hat{y}_i$ os valores ajustados pelo modelo para a unidade i , $\bar{y}$ a média

da variável resposta e n o número de observações.

Além disso, a inclusão de termos suaves foi também apoiada pela análise dos EDF, uma vez que estes quantificam a complexidade da função suavizada estimada. Em particular, valores de EDF próximos de 1 indicam uma relação aproximadamente linear entre a covariável e a variável resposta, enquanto valores superiores a 1 refletem desvios dessa linearidade, correspondendo a relações não lineares progressivamente mais flexíveis. No mesmo contexto, a significância estatística desses termos foi avaliada com base nos valores-p provenientes dos testes aproximados sob uma abordagem bayesiana. Nestes testes, a hipótese nula consiste em $f(x) = 0$ para todo o domínio da covariável, ou seja, assume-se que o termo suavizado não contribui significativamente para o modelo [62]. Tanto os EDF como os valores-p provêm da função `summary()`.

O modelo base incluiu apenas o efeito suave do tempo, $s(t, bs="cc", k=24)$, assumindo uma família Gaussiana e utilizando o logaritmo da concentração de CO_2 como variável resposta, permitindo assim descrever a dinâmica temporal média da concentração.

A partir desse modelo, avaliou-se a inclusão de efeitos aleatórios com o objetivo de captar diferentes níveis de variabilidade na estrutura dos dados. Num primeiro passo, foi adicionado um efeito aleatório por habitação, $s(habitação, bs="re")$, de forma a modelar a variabilidade entre unidades habitacionais. Posteriormente, foi incluído um efeito aleatório adicional associado ao dia, $s(dia, bs="re")$, para incorporar a variabilidade entre dias. Por fim, considerou-se uma estrutura com efeitos aleatórios encaixados, combinando a variabilidade entre habitações e entre dias dentro de cada habitação, $s(habitação, bs="re") + s(habitação, dia, bs="re")$. Este último modelo apresentou uma melhor qualidade do ajustamento, evidenciado por uma evidente melhoria do AIC e do R^2 , bem como por uma redução da variância residual. Observou-se ainda que o efeito aleatório associado à habitação foi o principal responsável pela explicação da variabilidade.

Considerando o modelo anterior, foi introduzida a variável país, a qual foi inicialmente considerada como efeito fixo e, posteriormente, em interação com um efeito suave do tempo, $s(t, bs="cc", k=24, by=país)$, dada a melhoria da qualidade de ajustamento do modelo. Esta inclusão, realizada antes das restantes covariáveis, deve-se ao facto de se tratar de uma variável de interesse no estudo, permitindo avaliar diferenças na dinâmica temporal da concentração de CO_2 entre países.

De seguida, foram incluídas no modelo apenas as características habitacionais selecionadas com base nos resultados da modelação simples. Esta seleção teve por base a presença de padrões diferentes entre categorias, evidenciados pelas curvas suavizadas da concentração de CO_2 . As restantes variáveis, cujas curvas se encontravam substancialmente sobrepostas, foram excluídas por não apresentarem padrões distintivos de variação temporal. Neste caso, as covariáveis foram adicionadas em simultâneo ao modelo, sendo posteriormente refinadas através de um procedimento iterativo de simplificação do modelo, essencialmente por eliminação regressiva (*backwards*). Neste processo, os termos com menor evidência estatística (maior valor-p) foram removidos progressivamente até se obter um conjunto de preditores estatisticamente relevante.

Além disso, foi adicionada a covariável densidade de ocupação, dada a relevância teórica para a concentração de CO_2 e por apresentar uma maior correlação com o CO_2 do que as restantes covariáveis contínuas. Por ser uma variável contínua, não pode ser adicionada como um efeito suave do tempo, dado que implicaria estimar uma curva temporal distinta para cada valor diferente de densidade. Por isso, foi testada como efeito fixo e como termo suave, $s(\text{densidade})$, o que permite relações não lineares com a concentração de CO_2 . Este último modelo apresentou melhor qualidade de ajustamento.

Para o refinamento da estrutura funcional do modelo, substituíram-se os efeitos fixos das covariáveis habitacionais por efeitos suaves do tempo. Esta abordagem permite captar possíveis diferenças nas tendências temporais entre categorias, o que se traduziu numa melhoria da qualidade de ajustamento do modelo. No entanto, verificou-se que uma das covariáveis incluídas (sinais físicos de patologias internas) não era estatisticamente significativa. Quando testada novamente como efeito fixo, manteve a ausência de significância, razão pela qual foi excluída do modelo.

Por fim, procedeu-se à modelação da correlação dos resíduos. Essa decisão foi fundamentada pela análise da ACF e da função de autocorrelação parcial (PACF), que permitem identificar a estrutura de dependência temporal dos dados e distinguir entre modelos autorregressivos (AR) e de MM. De forma geral, num processo AR de ordem p ($\text{AR}(p)$), a ACF tende a decair gradualmente e a PACF apresenta um corte abrupto após o lag p , enquanto, num processo MM de ordem q ($\text{MM}(q)$), observa-se o comportamento inverso, com uma ACF que corta após o lag q e uma PACF que decai progressivamente [67]. Com base nos padrões observados, foi testada a estrutura de autocorrelação $\text{AR}(1)$. Para a implementação dessa estrutura, foram utilizados os argumentos `rho` e `AR.start`.

O argumento `rho` corresponde ao parâmetro de autocorrelação de primeira ordem, cuja estimativa inicial foi obtida a partir do valor da ACF no primeiro lag [63, 68]. Por sua vez, o argumento `AR.start` foi utilizado para identificar o início de cada série temporal [63, 68]. Inicialmente, assumiu-se cada habitação como uma série temporal independente e, posteriormente, considerou-se cada dia dentro de cada habitação como uma série temporal distinta. Após o ajustamento dos modelos, a autocorrelação residual foi reavaliada através da função `acf_resid()`, tanto para o conjunto total das observações como separadamente por habitação e por dias dentro de habitações, permitindo verificar a adequação da estrutura de correlação introduzida [68]. Apesar de uma ligeira diminuição do R^2 , observou-se uma redução substancial do AIC e uma diminuição da autocorrelação residual, tendo o modelo com estrutura AR(1) considerando uma série temporal independente para cada dia dentro de cada habitação apresentado melhor qualidade de ajustamento.

Deste modo, o modelo final encontra-se descrito na Equação (21), onde o índice h corresponde ao identificador de cada habitação, d ao dia e t ao tempo. As categorias de referência consideradas para as variáveis categóricas foram Portugal para o país, ausência de renovação, apartamento para a tipologia e ausência de patologias de humidade.

$$\log(\text{CO}_2_{h,d,t}) = \beta_0 + f_1(t) + f_2(t, \text{País}_h) + f_3(t, \text{Renovação}_h) + f_4(t, \text{Tipologia}_h) + f_5(t, \text{Patologias Humidade}_h) + f_6(\text{Densidade}_h) + b_h + b_{d(h)} + \varepsilon_{h,d,t}, \quad (21)$$

$$b_h \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_h^2), \quad b_{d(h)} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_d^2), \quad \varepsilon_{h,d,t} = \rho \varepsilon_{h,d,t-1} + u_{h,d,t}, \quad u_{h,d,t} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

sendo os efeitos aleatórios independentes entre si e independentes dos erros.

2.4.5 Análise de sensibilidade

Após a definição do modelo final com base nos dados agregados por hora e suavizados por MM, procedeu-se a uma análise de sensibilidade com o objetivo de avaliar a robustez dos resultados face à discretização do intervalo temporal e ao efeito de suavização dos dados. Nesse contexto, o modelo previamente selecionado foi reajustado aos dados de 5 em 5 minutos previamente suavizados por MM e aos mesmos dados de 5 em 5 minutos mas sem qualquer suavização prévia. A utilização dos dados de 5 em 5 minutos permite captar variações de curta duração na concentração de CO_2 que podem ser atenuadas ou eliminadas na agregação horária. Por outro lado, os dados sem MM pre-

servam a variabilidade original (oscilações momentâneas, ruído, alterações abruptas), embora possam gerar maior instabilidade nos resultados. Assim, esta análise permite avaliar se as relações identificadas no modelo principal são consistentes ou se derivam de padrões induzidos pela agregação temporal e/ou pela suavização aplicada aos dados. A qualidade do ajustamento dos modelos foi avaliada através do R^2 ajustado.

Adicionalmente, procedeu-se a uma validação preditiva para avaliar a qualidade dos modelos ajustados. Os dados foram divididos em subconjuntos de treino e teste, considerando, respetivamente, aproximadamente 70% e 30% da dimensão temporal. Para o conjunto de treino, o modelo foi reajustado, enquanto para o conjunto de teste foram obtidas previsões da concentração de CO_2 para cada observação, com base no modelo estimado (função `predict()`). O desempenho preditivo foi avaliado através do erro quadrático médio (RMSE) e do erro absoluto médio (MAE), definidos pela Equação (22) [69]. Estas medidas quantificam a diferença entre os valores observados e os valores previstos, sendo que valores mais baixos indicam uma melhor capacidade preditiva do modelo.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (22)$$

3 Análise exploratória

3.1 Características habitacionais

Para caracterizar a distribuição das variáveis selecionadas e identificar padrões preliminares, procedeu-se a uma análise exploratória utilizando representações gráficas, de forma a orientar as etapas seguintes da análise estatística.

A descrição iniciou-se pelas variáveis estáticas, recolhidas num único instante. As distribuições por país das variáveis contínuas foram representadas através de caixas de bigodes (Figura 3.3). No período de construção, observa-se que Estónia, Portugal e Países Baixos apresentam as medianas mais elevadas, sugerindo edifícios mais recentes, enquanto o Reino Unido apresenta a mediana mais baixa. Também se observa uma maior dispersão na República Checa, Países Baixos e Reino Unido, o que indica a coexistência de edifícios antigos e recentes. Relativamente à densidade de ocupação, Portugal apresenta a mediana mais elevada, enquanto o Reino Unido evidencia a mediana mais baixa. Alguns valores aberrantes indicam casos específicos de subocupação. Quanto ao número de plantas, Suécia e Países Baixos apresentam medianas mais altas e maior dispersão. Pelo contrário, Itália e Reino Unido apresentam as medianas mais baixas, próximas de 0.

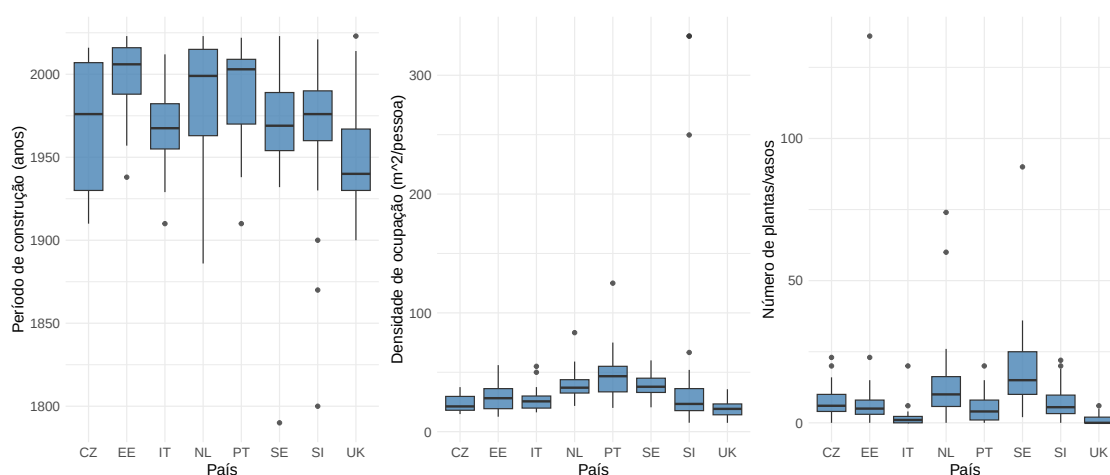


Figura 3.3: Caixa de bigodes por país para as variáveis contínuas.

Para as variáveis categóricas binárias e para concisão, foi construído um mapa de calor, que indica, através do gradiente de cores, a percentagem da categoria "Sim" por variável e por país (Figura 3.4). Relativamente à climatização, o aquecimento a madeira/*pellets* é mais frequente em Portugal, Suécia e Eslovénia, enquanto o arrefecimento elétrico pre-

domina na Itália, Eslovénia e Reino Unido. Observa-se ainda uma maior utilização de gás na Itália e Reino Unido. De forma geral, os sistemas de climatização e ventilação nas habitações são maioritariamente elétricos. A ventilação exclusivamente por abertura de janelas é mais comum na República Checa, Itália, Portugal e Reino Unido, enquanto os restantes países recorrem sobretudo a ventilação natural. Em relação às patologias internas, todos os países apresentam sinais, com exceção da República Checa, que regista valores muito reduzidos, e da Suécia e Eslovénia, onde quase não se observam habitações com sinais físicos. Itália apresenta o maior número de habitações com patologias associadas à humidade. No entanto, não se observa uma relação direta entre o tipo de ventilação e a presença de humidade. Quanto aos produtos domésticos e hábitos de ocupação, o uso de ambientadores e fragrâncias é mais frequente na Estónia, Portugal e Reino Unido, enquanto pesticidas/inseticidas apresentam baixa utilização em todos os países. A presença de animais de estimação é particularmente notável na Suécia e a prática de fumar dentro de casa é praticamente inexistente em todos os países, com uma ligeira incidência nas habitações de Itália. Por fim, a exposição a fontes externas mostra que o tráfego apresenta uma maior percentagem do que as restantes fontes, embora na Estónia e Suécia outras fontes sejam mais relevantes.

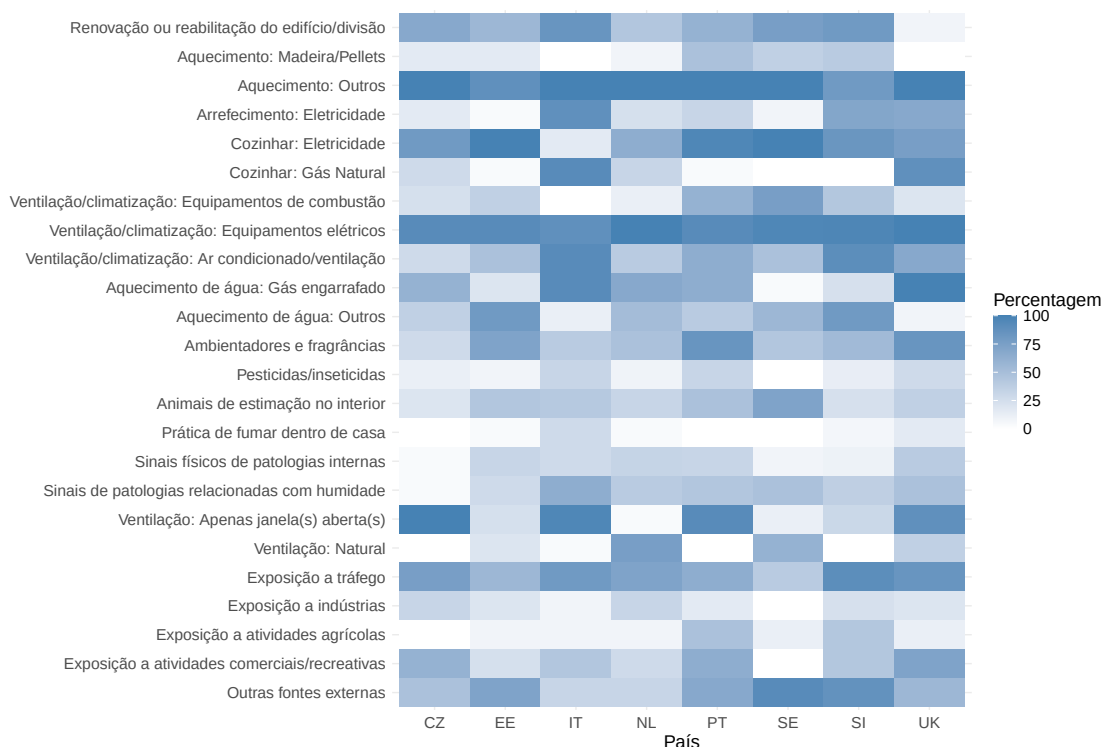


Figura 3.4: Mapa de calor para as variáveis binárias por país.

As variáveis categóricas com mais de duas categorias foram representadas através de gráficos de barras por país (Figura 3.5). Observa-se uma predominância de apartamen-

tos na República Checa, Estónia e Itália, sendo que, no Reino Unido, todas as habitações correspondem a moradias, o que limita a comparação entre tipologias. Quanto aos sistemas de ventilação mecânica, verifica-se que a maioria das habitações não dispõe deste tipo de sistemas. Nos Países Baixos, observa-se a maior frequência de sistemas de controlo central, enquanto na Suécia predominam os sistemas de controlo local. Na localização das habitações, verifica-se uma predominância da categoria GH nos países República Checa, Estónia, Itália, Portugal e Eslovénia, enquanto nos restantes países predomina a categoria EF. Relativamente ao número de janelas ou portas abríveis, predominam uma ou duas aberturas, sendo que a categoria >4 é mais comum na Eslovénia. Não se identifica uma relação direta evidente entre esta variável e a tipologia da habitação.

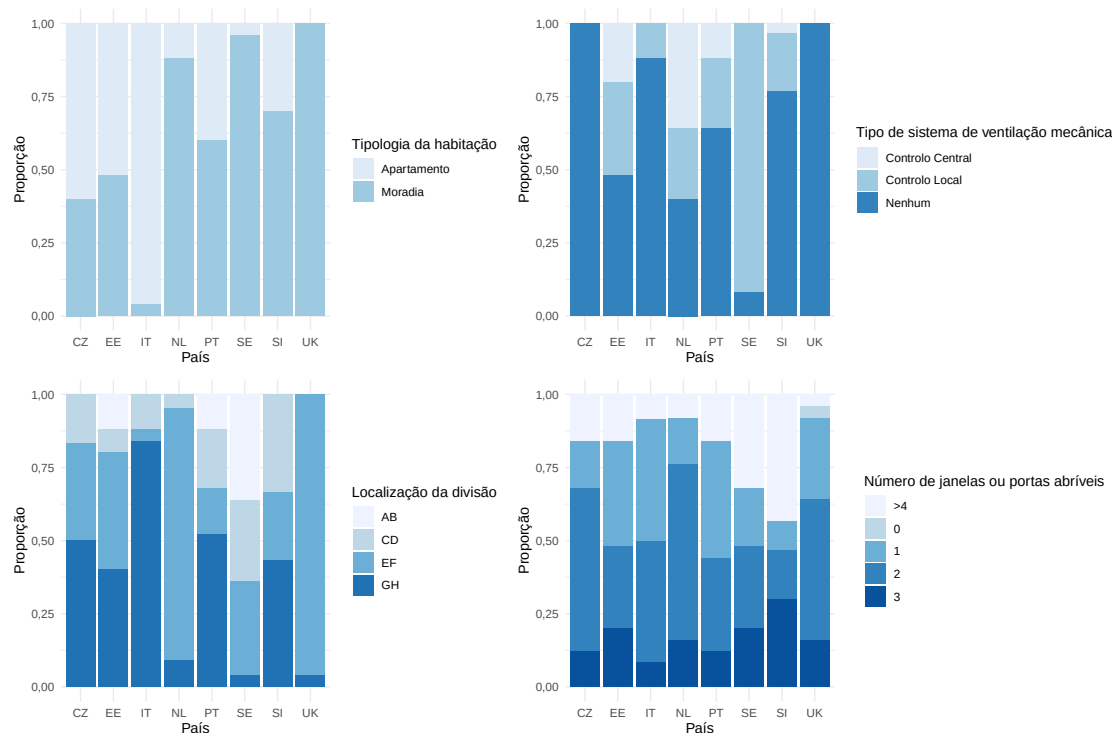


Figura 3.5: Gráfico de barras por país para as variáveis categóricas.

Por fim, foram elaboradas caixas de bigodes para explorar a relação entre variáveis contínuas e categóricas de interesse. A Figura 3.6 apresenta, à esquerda, a distribuição do período de construção em função da renovação ou reabilitação do edifício e, à direita, a distribuição da densidade de ocupação segundo a tipologia da habitação, por país. Em geral, as habitações renovadas tendem a ser mais antigas, com exceção de Itália e Reino Unido, onde as não renovadas apresentam períodos de construção mais antigos. Relativamente às habitações não renovadas, o Reino Unido apresenta edifícios mais antigos do que os restantes países. Por outro lado, a Estónia tem habitações

mais recentes, independentemente da existência de renovação. Assim, a relação entre o período de construção e a renovação varia consideravelmente entre países, sobretudo quando não há renovação. Quanto à densidade de ocupação, observa-se que os apartamentos tendem a apresentar valores inferiores aos das moradias. Em particular, Portugal destaca-se por apresentar uma densidade de ocupação relativamente mais elevada em ambas as tipologias, especialmente nas moradias. Estes resultados sugerem que a tipologia e a densidade de ocupação constituem fatores relevantes na distinção de padrões habitacionais entre países.

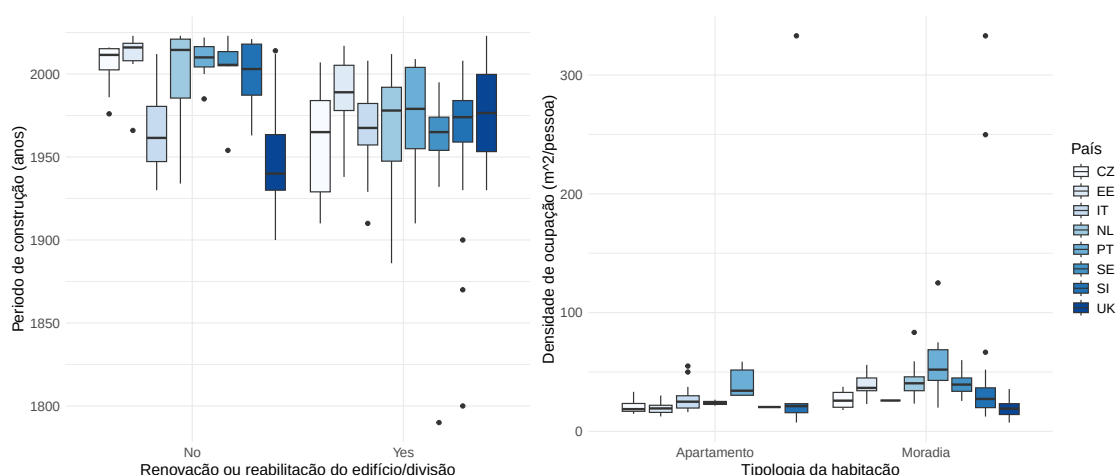


Figura 3.6: Caixa de bigodes do período de construção por renovação ou reabilitação do edifício (lado esquerdo) e caixa de bigodes da densidade de ocupação por tipologia da habitação (lado direito) por país.

3.2 Dióxido de carbono

Após o pré-processamento descrito na secção anterior, procedeu-se à análise exploratória das curvas de CO₂ obtidas, em diferentes cenários. A Figura 3.7 apresenta a evolução temporal da concentração de CO₂ por habitação, ao longo de todo o período de medição, antes e após a suavização por MM. Tal como previsto, a suavização reduz a variabilidade das curvas e atenua os valores extremos, mas mantém a tendência geral dos dados. Adicionalmente, verifica-se que, em ambos os conjuntos de dados, ocorrem concentrações de CO₂ que excedem o valor de referência de 1000 ppm. Isto pode estar associados a diferentes padrões de exposição, nomeadamente habitações com concentrações persistentemente elevadas ou episódios pontuais de aumento ao longo do dia.

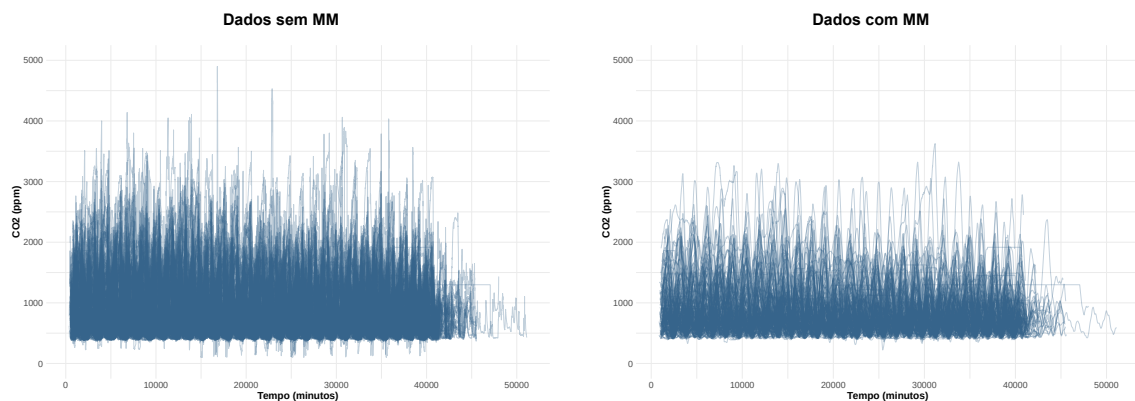


Figura 3.7: Evolução temporal da concentração de CO_2 por habitação, ao longo de todo o período de monitorização, antes e após a suavização por MM.

Para uma análise mais detalhada da variabilidade intra-diária, a Figura 3.8 apresenta a evolução dos níveis do CO_2 para cada dia de cada habitação, considerando os diferentes conjuntos de dados. Observa-se uma maior dispersão das curvas no início e no final do dia, com algumas habitações a apresentarem valores mais elevados de CO_2 nestes períodos. Isto pode estar associado à menor ocupação das habitações durante o período central do dia, quando os residentes se encontram, por exemplo, no local de trabalho.

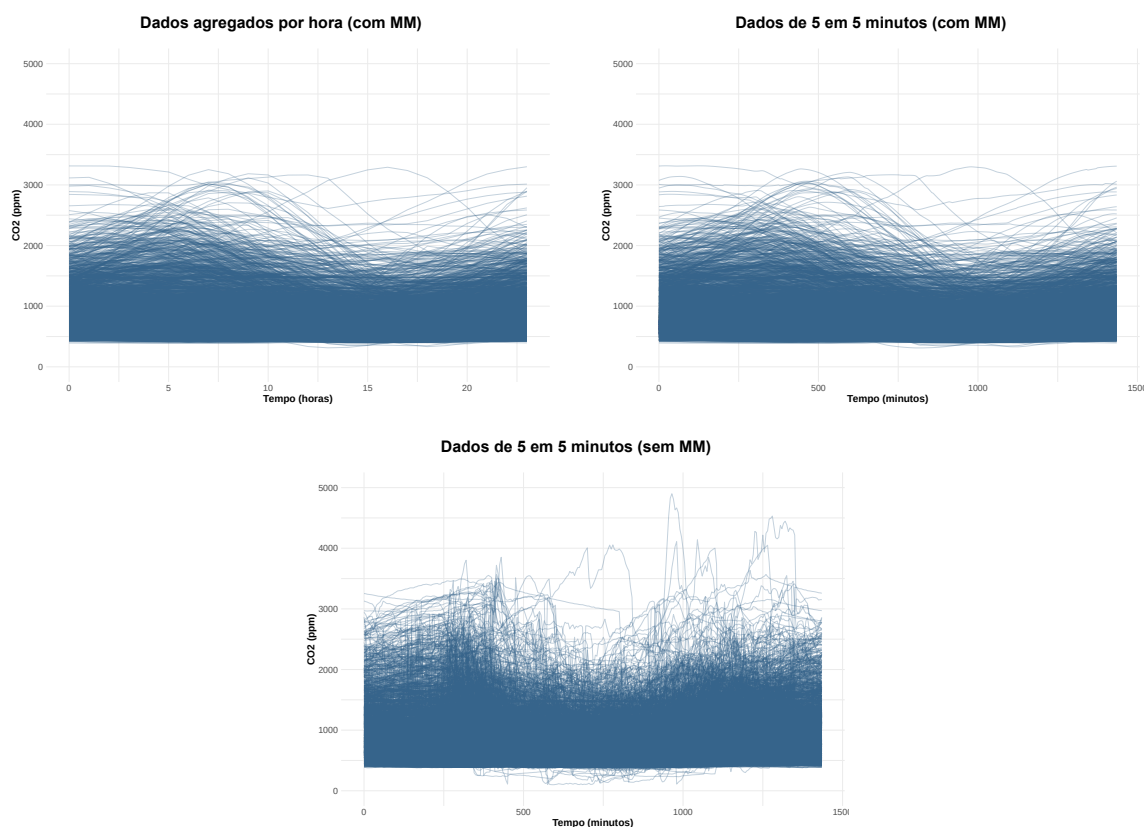


Figura 3.8: Evolução temporal dos níveis de CO_2 por dia e por habitação, considerando os diferentes conjuntos de dados.

Adicionalmente, a Figura C.1 mostra a evolução temporal do CO_2 para 20 habitações selecionadas aleatoriamente, o que permite uma melhor visualização dos padrões intra-

diários existentes e do impacto da suavização. Além disso, a Figura C.2 representa as ACFs para as mesmas habitações. No caso dos dados horários, verifica-se uma autocorrelação positiva elevada nos primeiros lags, seguida de um decaimento progressivo, o que indica uma evolução suave do CO_2 ao longo do dia. Na maioria das habitações observa-se ainda autocorrelação negativa na parte central do dia o que corrobora as conclusões obtidas a partir do gráfico anterior. Nos dados de 5 em 5 minutos, observa-se um comportamento mais oscilatório, com alternância entre valores positivos e negativos ao longo dos lags, resultante do menor intervalo entre medições. Ainda assim, mantém-se evidente um decaimento progressivo, semelhante ao observado nos dados horários. Além disso, o padrão de autocorrelação entre os dados com e sem suavização é bastante semelhante, com um ligeiro aumento da autocorrelação nos dados suavizados. Logo, conclui-se que as diferentes escalas de agregação e a suavização não introduzem alterações estruturais relevantes na ACF.

4 Resultados

4.1 Identificação de perfis habitacionais

Após a seleção das variáveis de interesse, foram mantidas 31 variáveis, apresentadas na Tabela 4.1, juntamente com as respetivas medidas descritivas, tanto para a totalidade dos dados como para cada país. Com base nos resultados dos testes estatísticos, apenas as variáveis relativas à presença de equipamentos elétricos para ventilação/climatização e à proximidade a fontes de poluição industriais na envolvente não foram utilizadas nos métodos de *clustering*, por não apresentaram diferenças estatisticamente significativas entre os grupos.

Tabela 4.1: Resumo das variáveis do estudo por país e respetivas medidas descritivas, incluindo valores-p de testes estatísticos ajustados pelo método de Holm.

Variável	CZ N = 25	EE N = 25	IT N = 25	NL N = 25	PT N = 25	SE N = 25	SI N = 30	UK N = 25	Total	valor-p (Holm)
Período de construção	1971 (36); 1976 (77)	1998 (22); 2006 (28)	1967 (25); 1968 (27)	1986 (36); 1999 (52)	1989 (30); 2003 (39)	1967 (44); 1969 (35)	1968 (45); 1976 (30)	1951 (36); 1940 (37)	1974 (38); 1977 (50)	0,0001***3
Renovação ou reabilitação do edifício/divisão	17 (68%)	14 (56%)	21 (84%)	11 (44%)	15 (60%)	19 (76%)	24 (80%)	2 (8.0%)	123 (60%)	<0,0001***1
Tipologia da habitação										<0,0001***1
Apartamento	15 (60%)	13 (52%)	24 (96%)	3 (12%)	10 (40%)	1 (4.0%)	9 (30%)	0 (0%)	75 (37%)	
Moradia	10 (40%)	12 (48%)	1 (4.0%)	22 (88%)	15 (60%)	24 (96%)	21 (70%)	25 (100%)	130 (63%)	
Densidade de ocupação	23 (7); 21 (12)	29 (12); 28 (17)	28 (10); 26 (10)	39 (13); 37 (11)	49 (22); 47 (22)	39 (10); 38 (12)	54 (87); 23 (19)	19 (7); 19 (9)	36 (37); 30 (20)	<0,0001***2
Aquecimento: Madeira/Pellets	4 (16%)	4 (16%)	0 (0%)	2 (8.0%)	12 (48%)	9 (36%)	12 (40%)	0 (0%)	43 (21%)	<0,0001***1
Aquecimento: Outros	25 (100%)	22 (88%)	25 (100%)	25 (100%)	25 (100%)	25 (100%)	24 (80%)	25 (100%)	196 (96%)	0,0022**2
Arrefecimento: Eletricidade	4 (16%)	1 (4.0%)	22 (88%)	6 (24%)	8 (32%)	2 (8.0%)	21 (70%)	17 (68%)	81 (40%)	<0,0001***1
Cozinhar: Eletricidade	20 (80%)	25 (100%)	4 (16%)	16 (64%)	24 (96%)	25 (100%)	25 (83%)	19 (76%)	158 (77%)	<0,0001***1
Cozinhar: Gás natural	7 (28%)	1 (4.0%)	23 (92%)	8 (32%)	1 (4.0%)	0 (0%)	0 (0%)	22 (88%)	62 (30%)	<0,0001***1
Ventilação/climatização: Equipamentos de combustão	6 (24%)	9 (36%)	0 (0%)	3 (12%)	15 (60%)	19 (76%)	13 (43%)	5 (20%)	70 (34%)	<0,0001***1
Ventilação/climatização: Equipamentos elétricos	23 (92%)	23 (92%)	22 (88%)	25 (100%)	23 (92%)	24 (96%)	29 (97%)	25 (100%)	194 (95%)	0,5184 ²
Ventilação/climatização: Ar condicionado/ventilação	7 (28%)	12 (48%)	23 (92%)	10 (40%)	16 (64%)	12 (48%)	27 (90%)	17 (68%)	124 (60%)	<0,0001***1
Aquecimento de água: Gás engarrafado	15 (60%)	5 (20%)	23 (92%)	17 (68%)	16 (64%)	1 (4.0%)	7 (23%)	25 (100%)	109 (53%)	<0,0001***1
Aquecimento de água: Outros	9 (36%)	20 (80%)	3 (12%)	13 (52%)	10 (40%)	14 (56%)	24 (80%)	2 (8.0%)	95 (46%)	<0,0001***1
Ambientadores e fragrâncias	7 (28%)	18 (72%)	10 (40%)	12 (48%)	21 (84%)	11 (44%)	16 (53%)	21 (84%)	116 (57%)	0,0006***1
Pesticidas/inseticidas	3 (12%)	2 (8.0%)	8 (32%)	2 (9.1%)	8 (32%)	0 (0%)	4 (13%)	7 (28%)	34 (17%)	0,0270*2
Animais de estimação no interior	5 (20%)	11 (44%)	10 (42%)	8 (32%)	12 (48%)	18 (72%)	7 (23%)	9 (36%)	80 (39%)	0,0270*1
Número de plantas/vasos	8 (6); 6 (6)	11 (26); 5 (5)	2 (4); 1 (2)	15 (17); 10 (11)	5 (5); 4 (7)	20 (17); 15 (15)	7 (6); 6 (7)	1 (2); 0 (2)	9 (14); 5 (8)	<0,0001***2
Prática de fumar dentro de casa	0 (0%)	1 (4.0%)	7 (28%)	1 (4.0%)	0 (0%)	0 (0%)	2 (6.7%)	4 (16%)	15 (7.3%)	0,0077**2

Tipo de sistema de ventilação mecânica										0,0001*** ²
Controlo Central	0 (0%)	5 (20%)	0 (0%)	9 (36%)	3 (12%)	0 (0%)	1 (3.3%)	0 (0%)	18 (8.8%)	
Controlo Local	0 (0%)	8 (32%)	3 (12%)	6 (24%)	6 (24%)	23 (92%)	6 (20%)	0 (0%)	52 (25%)	
Nenhum	25 (100%)	12 (48%)	22 (88%)	10 (40%)	16 (64%)	2 (8.0%)	23 (77%)	25 (100%)	135 (66%)	
Sinais físicos de patologias internas	1 (4.0%)	8 (32%)	7 (28%)	8 (33%)	8 (32%)	2 (8.0%)	3 (10%)	10 (40%)	47 (23%)	0,0270* ¹
Sinais de patologias relacionadas com humidade	1 (4.0%)	7 (28%)	16 (64%)	10 (40%)	11 (44%)	12 (48%)	11 (37%)	12 (48%)	80 (39%)	0,0132* ¹
Localização da divisão										0,0001*** ²
AB	0 (0%)	3 (12%)	0 (0%)	0 (0%)	3 (12%)	9 (36%)	0 (0%)	0 (0%)	15 (7.5%)	
CD	4 (17%)	2 (8.0%)	3 (12%)	1 (4.5%)	5 (20%)	7 (28%)	10 (33%)	0 (0%)	32 (16%)	
EF	8 (33%)	10 (40%)	1 (4.0%)	19 (86%)	4 (16%)	8 (32%)	7 (23%)	24 (96%)	81 (40%)	
GH	12 (50%)	10 (40%)	21 (84%)	2 (9.1%)	13 (52%)	1 (4.0%)	13 (43%)	1 (4.0%)	73 (36%)	
Número de janelas ou portas abríveis										0,0184* ²
>4	4 (16%)	4 (16%)	2 (8.3%)	2 (8.0%)	4 (16%)	8 (32%)	13 (43%)	1 (4.0%)	38 (19%)	
0	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (4.0%)	1 (0.5%)	
1	4 (16%)	9 (36%)	10 (42%)	4 (16%)	10 (40%)	5 (20%)	3 (10%)	7 (28%)	52 (25%)	
2	14 (56%)	7 (28%)	10 (42%)	15 (60%)	8 (32%)	7 (28%)	5 (17%)	12 (48%)	78 (38%)	
3	3 (12%)	5 (20%)	2 (8.3%)	4 (16%)	3 (12%)	5 (20%)	9 (30%)	4 (16%)	35 (17%)	
Ventilação: Apenas janela(s) aberta(s)	25 (100%)	6 (24%)	24 (96%)	1 (4.0%)	23 (92%)	3 (12%)	9 (30%)	22 (88%)	113 (55%)	<0,0001*** ¹
Ventilação: Natural	0 (0%)	5 (20%)	1 (4.0%)	19 (76%)	0 (0%)	15 (60%)	0 (0%)	9 (36%)	49 (24%)	<0,0001*** ¹
Exposição a tráfego	19 (76%)	14 (56%)	20 (80%)	18 (72%)	16 (64%)	10 (40%)	27 (90%)	21 (84%)	145 (71%)	0,0113* ¹
Exposição a indústrias	8 (32%)	5 (20%)	2 (8.0%)	8 (32%)	4 (16%)	0 (0%)	7 (23%)	5 (20%)	39 (19%)	0,0523 ²
Exposição a atividades agrícolas	0 (0%)	2 (8.0%)	2 (8.0%)	2 (8.0%)	12 (48%)	3 (12%)	13 (43%)	3 (12%)	37 (18%)	0,0001*** ²
Exposição a atividades comerciais/recreativas	15 (60%)	6 (24%)	11 (44%)	7 (28%)	16 (64%)	0 (0%)	13 (43%)	18 (72%)	86 (42%)	<0,0001*** ¹
Outras fontes externas	12 (48%)	18 (72%)	8 (32%)	8 (32%)	17 (68%)	23 (92%)	26 (87%)	14 (56%)	126 (61%)	<0,0001*** ¹

Variáveis contínuas: média (desvio-padrão); mediana (intervalo interquartil)
 Variáveis categóricas: frequência absoluta (frequência relativa)
 Testes estatísticos: ¹ Qui-quadrado; ² Fisher; ³ Kruskal-Wallis
 Diferença estatisticamente significativa: * valor-p < 0,05; ** valor-p < 0,01; *** valor-p < 0,001

Relativamente à seleção do número ótimo de *clusters* (Tabela D.2), os métodos PAM, hierárquico e SOM convergiram de forma consistente para uma solução de dois *clusters*. Pelo contrário, o método FMM sugeriu quatro *clusters* com base no BIC. No entanto, essa solução revelou baixa estabilidade, uma vez que um dos *clusters* era composto por um número muito reduzido de indivíduos (9 indivíduos), com valores aberrantes em algumas variáveis. A exclusão destas observações foi considerada para testar a estabilidade da solução, contudo não foi adotada por estas poderem representar habitações com pior QAI e, conseqüentemente, por implicar a redução da variabilidade da amostra. Deste modo, consideraram-se dois *clusters* em todos os métodos.

No caso dos métodos hierárquicos, os dendrogramas obtidos pelas diferentes funções de ligação encontram-se representados na Figura D.1. A correlação cofenética entre as estruturas foi de 0,8788 (próximo de 1), indicando uma forte semelhança entre os dois dendrogramas, apesar das diferenças na organização interna da hierarquia.

Quanto ao SOM, a Figura D.2 apresenta o dendrograma resultante da aplicação do método hierárquico com ligação completa aos neurónios, bem como o respetivo mapa da rede de neurónios com dimensão 8×8 , que permitem visualizar a estrutura dos neurónios e o agrupamento obtido.

Para aferir a estabilidade dos *clusters* obtidos por cada método, a Tabela 4.2 apresenta a média dos valores do vetor de semelhança de Jaccard, calculada a partir dos métodos *bootstrap* e *subsetting*. De um modo geral, todos os métodos apresentaram valores elevados (próximos de 1), o que indica boa estabilidade. No entanto, o SOM apresentou os valores mais baixos no método *subsetting*, sugerindo uma menor robustez face a alterações na amostra. Além disso, entre os métodos hierárquicos, a ligação completa registou valores inferiores aos do método de Ward, refletindo uma maior sensibilidade a perturbações nos dados. Consequentemente, os métodos de *ensemble* aplicados posteriormente incluíram apenas as soluções dos métodos PAM, hierárquico com método de Ward e FMM. Importa ainda salientar que, no método hierárquico, a atribuição dos indivíduos aos *clusters* foi igual para ambas as funções de ligação, pelo que a utilização simultânea das duas funções de ligação introduziria redundância sem acrescentar informação adicional.

Tabela 4.2: Média dos valores do vetor de semelhança de Jaccard obtidos pelos métodos *bootstrap* e *subsetting* para cada método de *clustering*.

	PAM		Hierárquico - Método de Ward		Hierárquico - Ligação completa		FMM		SOM	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Bootstrap	0,8862	0,9109	0,9342	0,9635	0,7474	0,7785	0,9397	0,9194	0,8802	0,8883
Subsetting	0,8049	0,8430	0,8762	0,9319	0,6856	0,7312	0,8615	0,7816	0,4436	0,6275

A Tabela 4.3 apresenta o tamanho dos *clusters* e os índices de validação interna para comparação dos métodos PAM, hierárquico e SOM, bem como, o BIC para o método FMM. De modo geral, os três métodos apresentam um coeficiente de silhueta moderado, sugerindo uma separação apenas razoável entre os *clusters*. Ainda assim, o *clustering* hierárquico apresenta o valor médio de silhueta e o índice de Dunn mais elevados, o que indica *clusters* mais distintos e internamente mais coesos. Por outro lado, o índice de Davies-Bouldin indica um melhor equilíbrio entre compacidade *intra-cluster* e separação *inter-cluster* no método PAM, devido ao valor mais baixo. Além disso, os valores mínimos de silhueta sugerem que a maioria dos indivíduos está corretamente atribuída aos respetivos *clusters*. Contudo, no PAM e no SOM observam-se alguns casos com silhuetas negativas, o que indica que essas observações poderão estar mais próximas de outro

cluster do que daquele ao qual foram atribuídas.

Tabela 4.3: Tamanho dos *clusters* e índices de validação interna para comparação dos métodos de *clustering* e BIC para o método FMM.

	Tamanho cluster 1	Tamanho cluster 2	Coeficiente de silhueta								Índices		
			Média cluster 1	Média cluster 2	Mínimo	1º Quartil	Média	Mediana	3º Quartil	Máximo	Dunn	Davies-Bouldin	BIC
PAM	82	112	0,3015	0,2265	-0,0609	0,1486	0,2582	0,2607	0,3638	0,5399	0,1265	2,601	-
Hierárquico	76	118	0,3178	0,2635	0,0129	0,2120	0,2847	0,2966	0,3589	0,4668	0,2281	2,809	-
FMM	117	77	-	-	-	-	-	-	-	-	-	-	7801
SOM	92	102	0,2446	0,2308	-0,1827	0,1329	0,2373	0,2339	0,3808	0,4893	0,1452	2,999	-

De forma complementar, a Figura D.3 representam os dois *clusters* obtidos pelos métodos PAM, hierárquico e SOM, projetados nas duas primeiras dimensões após redução de dimensionalidade pelo MDS. Apesar do coeficiente médio de silhueta não ser elevado, a representação visual confirma a separação dos *clusters*. No entanto, o método hierárquico revela-se ligeiramente mais robusto na definição dos *clusters*, sendo concordante com os resultados da tabela anterior.

Para avaliar a semelhança entre os resultados dos diferentes métodos, foram representadas matrizes de dissimilaridade/semelhança através de mapas de calor (Figura 4.9). Todas as medidas indicam uma maior semelhança entre os métodos PAM e SOM, enquanto o método hierárquico e o FMM apresentam os resultados mais distintos.

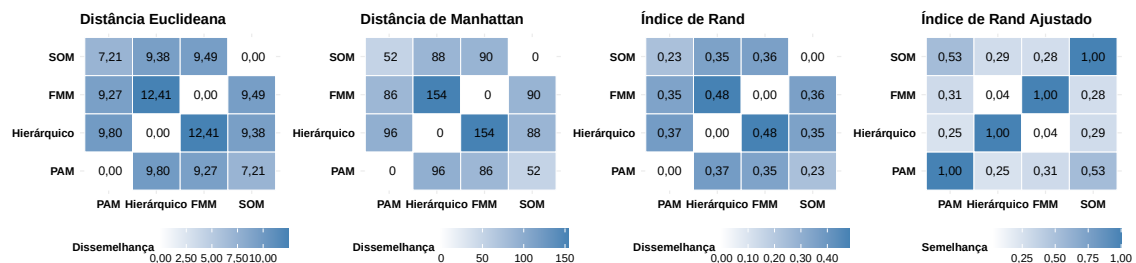


Figura 4.9: Mapas de calor das matrizes de dissimilaridade e semelhança dos métodos de *clustering*.

Relativamente aos métodos de *ensemble*, com o intuito de comparar as diferentes abordagens, representaram-se graficamente as matrizes de pertinência obtidas pelos métodos *soft euclidean* e *soft manhattan* através de mapas de calor (Figura 4.10). Observa-se que o método *soft manhattan* apresenta apenas valores 0 ou 1, enquanto o método *soft euclidean* apresenta valores intermédios (0,33 e 0,67), refletindo alguma incerteza na atribuição das observações aos *clusters*. O cálculo do coeficiente de partição corrobora esta conclusão, na qual o método *soft manhattan* apresentou um coeficiente de 1, indicando uma solução totalmente *hard*, enquanto o método *soft euclidean* apresentou um coeficiente de 0,81, refletindo uma solução levemente *soft*. Apesar desta diferença, os

métodos resultaram exatamente na mesma solução final, assim como no mesmo agrupamento obtido pelos métodos *hard euclidean* e *hard manhattan*, enfatizando a robustez da solução de consenso. Dada a equivalência entre as soluções obtidas, optou-se por apresentar os resultados subsequentes com base no método *hard Manhattan*.

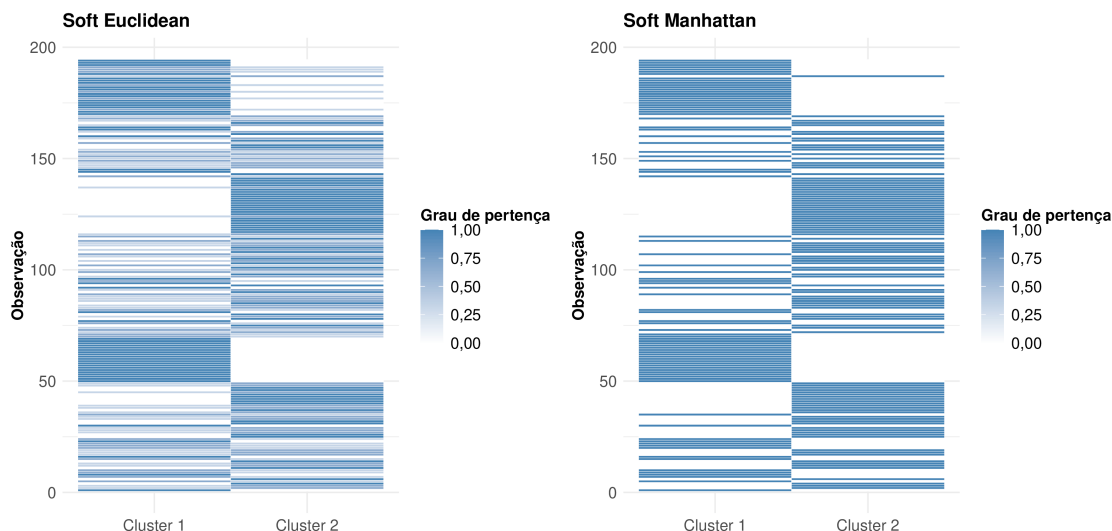


Figura 4.10: Mapas de calor das matrizes de pertença obtidas pelos métodos *soft euclidean* e *soft Manhattan*, representando a probabilidade de cada indivíduo pertencer a cada *cluster*.

Posteriormente, avaliou-se a concordância entre a solução de consenso e os resultados dos métodos de *clustering* individuais através do ARI (Tabela 4.4). Constatou-se que a solução do método de *ensemble* apresentou maior semelhança com o resultado obtido pelo PAM em comparação com os restantes métodos, sendo o método hierárquico aquele que apresentou maior divergência.

Tabela 4.4: ARI entre o resultado do método de *ensemble* e dos métodos de *clustering* individuais.

	PAM	Hierárquico	FMM
ARI	0,8601	0,3297	0,3922

A Tabela 4.5 apresenta o número de indivíduos por país em cada *cluster*. Observa-se que a maioria dos indivíduos residentes em Itália e Reino Unido encontram-se no *cluster* 1, enquanto os restantes países foram predominantemente atribuídos ao *cluster* 2. Além disso, verificou-se que Itália e Reino Unido foram consistentemente alocados ao mesmo *cluster*, independentemente da abordagem utilizada, evidenciando a estabilidade das soluções obtidas pelos diferentes métodos de *clustering*.

Tabela 4.5: Número de indivíduos por país em cada *cluster* obtido pelo método de *ensemble*.

	CZ	EE	IT	NL	PT	SE	SI	UK
Cluster 1	13	2	22	6	9	0	11	24
Cluster 2	11	23	0	13	16	25	18	1

Por fim, a Tabela 4.6 apresenta as medidas descritivas das variáveis com diferenças estatisticamente significativas entre os *clusters*. No geral, o *cluster* 1 é caracterizado por edifícios mais antigos e predominantemente apartamentos, com uma maior utilização de gás e uma maior ventilação exclusivamente baseada na abertura de janelas. Em contraste, o *cluster* 2 é constituído sobretudo por moradias, com uma maior utilização de madeira/*pellets* para aquecimento, maior prevalência de ventilação mecânica e um maior número de plantas.

Tabela 4.6: Medidas descritivas das variáveis com diferenças estatisticamente significativas entre os *clusters* obtidos pelo método de *ensemble*.

Variável	Cluster 1	Cluster 2	valor-p (Holm)
Período de construção	1966 (34), 1966 (52)	1978 (40), 1984 (44)	0,0196 ^{3*}
Tipologia da habitação	Apartamento (52%)	Moradia (76%)	0,0015 ^{2**}
Densidade de ocupação	31 (36), 23 (14)	40 (38), 33 (22)	0,0008 ^{3***}
Aquecimento: Madeira/Pellets	Não (95%)	Não (64%)	<0,0001 ^{2***}
Arrefecimento: Eletricidade	Sim (70%)	Não (86%)	<0,0001 ^{2***}
Cozinhar: Eletricidade	Sim (60%)	Sim (95%)	<0,0001 ^{2***}
Cozinhar: Gás natural	Sim (62%)	Não (97%)	<0,0001 ^{2***}
Ventilação/climatização: Equipamentos de combustão	Não (89%)	Sim (55%)	<0,0001 ^{2***}
Ventilação/climatização: Ar condicionado/ventilação	Sim (74%)	Sim (50%)	0,0162 ^{2*}
Aquecimento de água: Gás engarrafado	Sim (85%)	Não (74%)	<0,0001 ^{2***}
Aquecimento de água: Outros	Não (84%)	Sim (70%)	<0,0001 ^{2***}
Número de plantas/vasos	4 (5), 2 (6)	13 (17), 8 (10)	<0,0001 ^{3***}
Tipo de sistema de ventilação mecânica	Nenhum (90%)	Nenhum (48%)	0,0002 ^{3***}
Localização da divisão	GH (47%)	EF (41%)	0,0018 ^{2**}
Ventilação: Apenas janela(s) aberta(s)	Sim (78%)	Não (62%)	<0,0001 ^{2***}
Exposição a atividades comerciais/recreativas	Sim (59%)	Não (71%)	0,0008 ^{2***}
Outras fontes externas	Não (54%)	Sim (75%)	0,0009 ^{2***}

Variáveis contínuas: média (desvio-padrão), mediana (intervalo interquartil)

Variáveis categóricas: categoria mais frequente (frequência relativa)

Testes estatísticos: ¹ Qui-quadrado; ² Fisher; ³ Kruskal-Wallis

Diferença estatisticamente significativa: * valor-p < 0,05; ** valor-p < 0,01; *** valor-p < 0,001

4.2 Modelação do dióxido de carbono em função das características habitacionais

Para analisar a relação entre os perfis habitacionais e a concentração de CO₂, a Figura 4.11 apresenta o efeito suave do tempo para cada um dos *clusters* identificados, proveniente do ajustamento dos modelos simples. As curvas obtidas revelam um comportamento temporal semelhante entre *clusters*, caracterizado por concentrações mais

elevadas durante a madrugada, com um máximo em torno das 2h, e por uma redução gradual ao longo do dia, atingindo valores mínimos por volta das 16h. No entanto, o *cluster* 1 apresenta sistematicamente concentrações de CO₂ superiores às observadas no *cluster* 2.

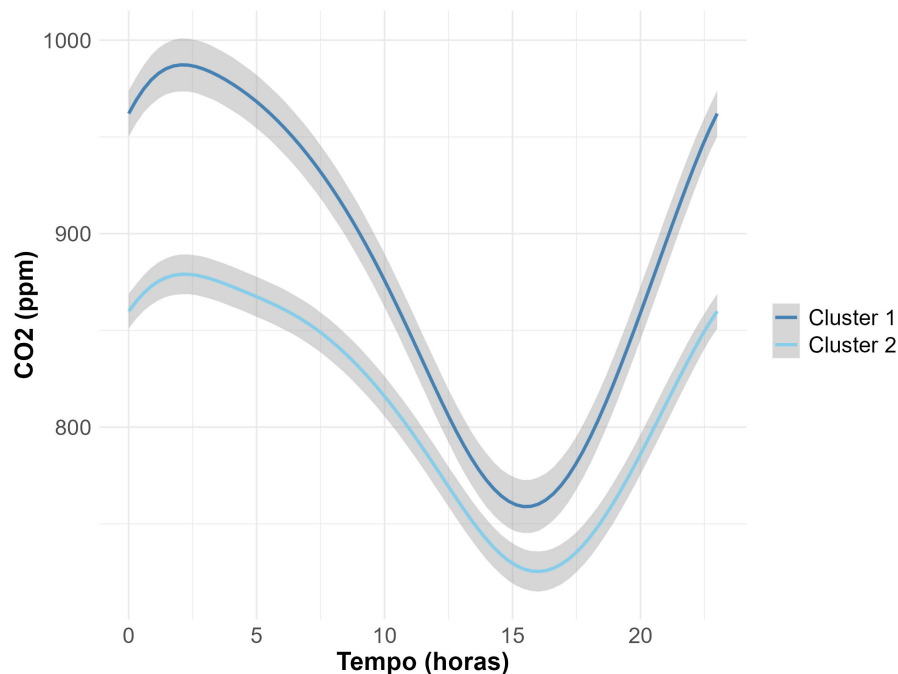
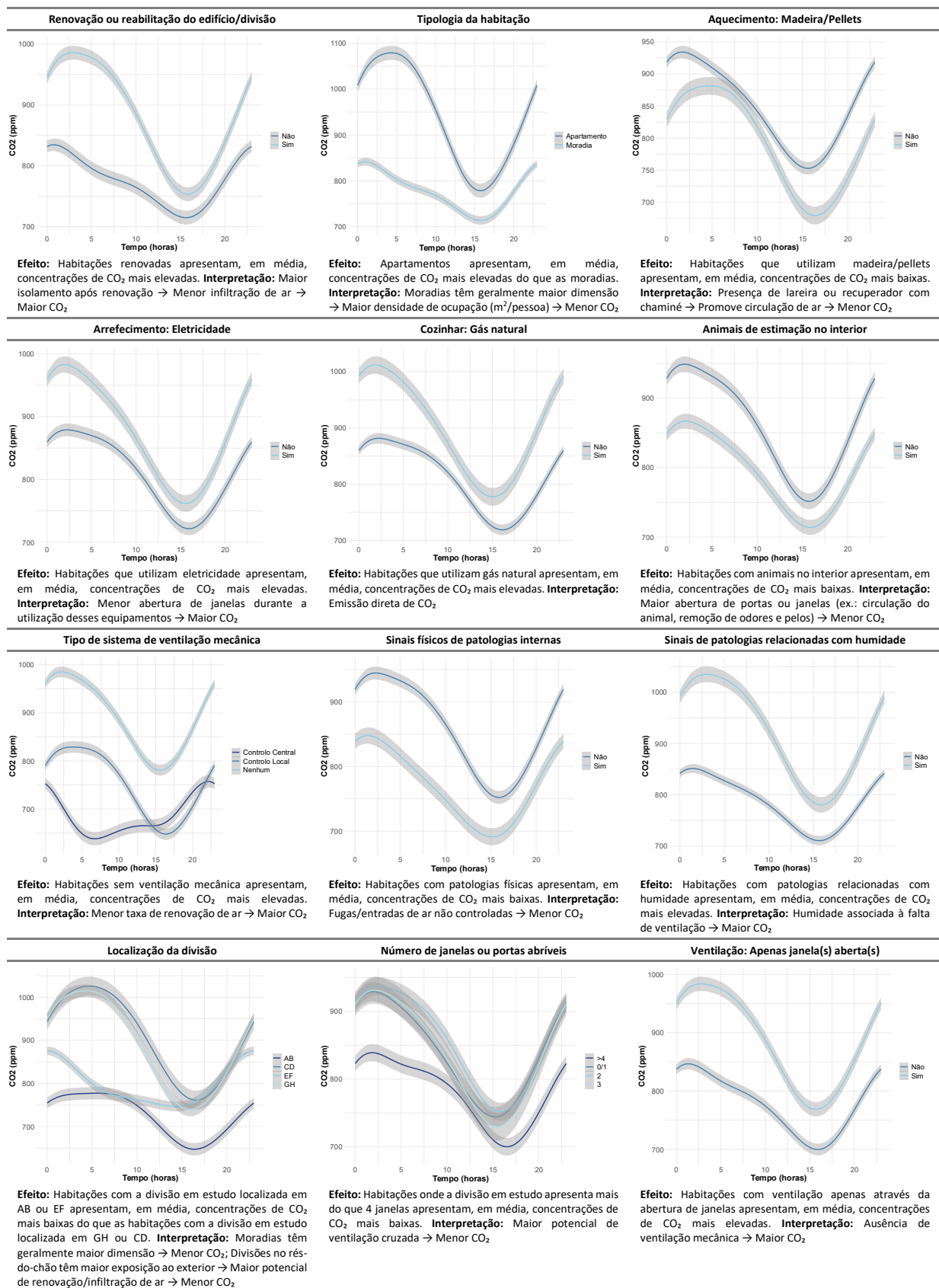


Figura 4.11: Efeito suave do tempo sobre a concentração de CO₂ para cada *cluster*.

Com base nos modelos simples ajustados por categoria das características habitacionais, procedeu-se à análise exploratória do efeito destas variáveis na concentração de CO₂. A Tabela 4.7 sintetiza os principais padrões observados nas variáveis selecionadas para a modelação subsequente, indicando, para cada característica, se os níveis de CO₂ tendem a ser mais elevados ou mais baixos ao longo do dia, bem como a sua interpretação no contexto habitacional e comportamental.

Tabela 4.7: Efeito das características habitacionais nos níveis de CO₂ e respetiva interpretação.



Relativamente ao desempenho do modelo múltiplo ajustado aos diferentes conjuntos de dados, o R^2_{adj} foi de 0,757 para os dados agregados por hora com MM, de 0,769 para os

dados de 5 em 5 minutos com MM e de 0,626 para os dados de 5 em 5 minutos sem MM. Estes valores indicam que os modelos apresentam uma boa capacidade para explicar a variabilidade observada nas concentrações de CO_2 . Em particular, verifica-se que os modelos ajustados aos dados suavizados apresentam desempenhos muito semelhantes, enquanto o modelo ajustado aos dados de 5 em 5 minutos sem suavização evidencia uma qualidade de ajustamento inferior, sugerindo que a suavização contribui para reduzir a variabilidade de curto prazo e melhorar a capacidade explicativa do modelo.

A Tabela E.1 apresenta o sumário dos termos suaves estimados, incluindo os EDF, os graus de liberdade de referência (Ref.df), a estatística de teste F e o respetivo valor-p. Globalmente, todos os termos suaves revelaram-se estatisticamente significativos para os três conjuntos de dados, com exceção do efeito aleatório do dia nos dados de 5 em 5 minutos com MM, que apresentou um valor de 0,2290, não sendo estatisticamente significativo. Os valores elevados dos EDF indicam a presença de relações não lineares com elevada complexidade.

A análise dos pressupostos dos modelos (Figura E.1) indica que, apesar de se observarem ligeiros desvios à normalidade nas caudas dos QQ-plots e alguma autocorrelação residual nos modelos com suavização, os pressupostos de homocedasticidade (resíduos vs valores ajustados) e de linearidade (resposta vs valores ajustados) encontram-se razoavelmente satisfeitos. A autocorrelação observada é substancialmente inferior à verificada no modelo sem modelação dos resíduos. Assim, considera-se que os modelos apresentam um ajustamento adequado e que as estimativas obtidas podem ser consideradas fiáveis.

A Tabela 4.8 apresenta as estimativas do termo constante, bem como o respetivo intervalo de confiança a 95% e o valor-p. Como os coeficientes foram ajustados na escala logarítmica, os valores apresentados correspondem às estimativas transformadas, através da exponencial, para a escala original da concentração de CO_2 . O termo constante representa a concentração média de CO_2 esperada para uma habitação de referência, considerando o efeito da variável contínua zero. Comparando os diferentes modelos, observa-se que as estimativas do termo constante são bastante semelhantes. Os respetivos intervalos de confiança apresentam uma amplitude reduzida e os valores-p indicam que os termos constantes são estatisticamente significativos. A partir desta concentração de referência, os termos suaves apresentados em seguida descrevem a forma como a concentração de CO_2 varia em função do tempo e das características das habitações.

Tabela 4.8: Estimativas do termo constante do modelo múltiplo para os diferentes conjuntos de dados, expressas em ppm.

Modelo	Termo constante	IC 95%	valor-p
Dados agregados por hora (com MM)	777,72	747,42 - 809,26	<0,0001***
Dados de 5 em 5 minutos (com MM)	790,38	759,38 - 822,66	<0,0001***
Dados de 5 em 5 minutos (sem MM)	766,07	736,14 - 797,21	<0,0001***

IC = Intervalo de confiança
Diferença estatisticamente significativa: * valor-p < 0,05; ** valor-p < 0,01; *** valor-p < 0,001

Dada a complexidade do modelo final, os resultados foram organizados por grupos de efeitos suaves com interpretação semelhante, de forma a facilitar a leitura e a interpretação dos padrões estimados. Importa salientar que os coeficientes apresentados nas figuras seguintes são estimados na escala logarítmica da concentração de CO₂. Assim, para interpretar os resultados na escala original, os coeficientes devem ser convertidos em variações percentuais: $(\exp(\text{coeficiente}) - 1) \times 100$. Desta forma, valores negativos correspondem a concentrações inferiores à concentração média estimada, enquanto valores positivos correspondem a concentrações superiores.

A Figura 4.12 apresenta a estimativa do termo suave do tempo sobre a concentração de CO₂, mantendo as variáveis categóricas nos níveis de referência e considerando o efeito da variável contínua igual a zero (termo constante). Deste modo, a curva representa a evolução temporal da concentração de CO₂ para uma habitação típica. Observa-se um padrão comum aos três conjuntos de dados, caracterizado por valores mais elevados no início do dia, seguidos de uma diminuição gradual até aproximadamente meio do dia e por um novo aumento no final. Os intervalos de confiança são bastante estreitos, o que indica uma elevada precisão na estimativa deste efeito. Em particular, nos dados com suavização, por volta das 17h, estima-se que a concentração de CO₂ é, em média, aproximadamente 14% inferior à concentração de CO₂. Nos dados sem suavização prévia, apesar de se observar um comportamento semelhante, a curva apresenta maior irregularidade, sugerindo que a aplicação das MM reduz o ruído sem alterar substancialmente a estrutura global do efeito temporal.

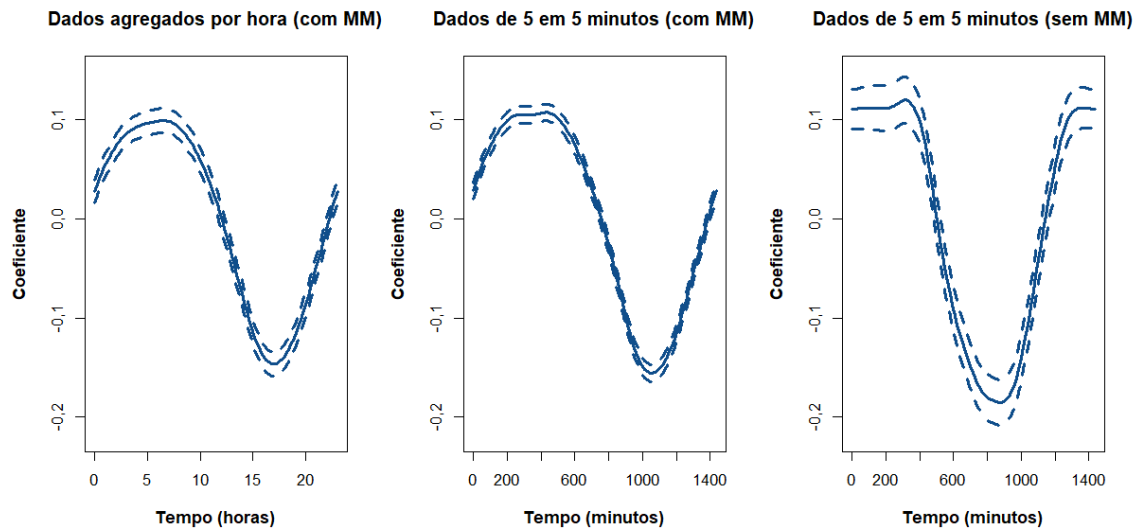


Figura 4.12: Efeito suave do tempo sobre a concentração de CO₂ estimado pelo modelo múltiplo para os diferentes conjuntos de dados. A linha representa o efeito estimado e as linhas a tracejado correspondem ao intervalo de confiança a 95%, tendo em conta a incerteza do termo constante.

A Figura 4.13 representa a estimativa do efeito suave do tempo em cada país, em relação à categoria de referência de cada variável. Por exemplo, considerando os dados com suavização, para as habitações da Eslovénia, por volta das 15h, estima-se que a concentração de CO₂ é, em média, 14% inferior à concentração de CO₂ nas habitações de Portugal. A interpretação pode ser realizada de forma semelhante para qualquer país a qualquer hora/minuto. Nos dados sem suavização observam-se curvas mais irregulares e alguns picos mais pronunciados. Contudo, a estrutura global dos efeitos mantém-se semelhante, indicando que a ausência de suavização aumenta a variabilidade local sem alterar os padrões temporais gerais associados a cada país.

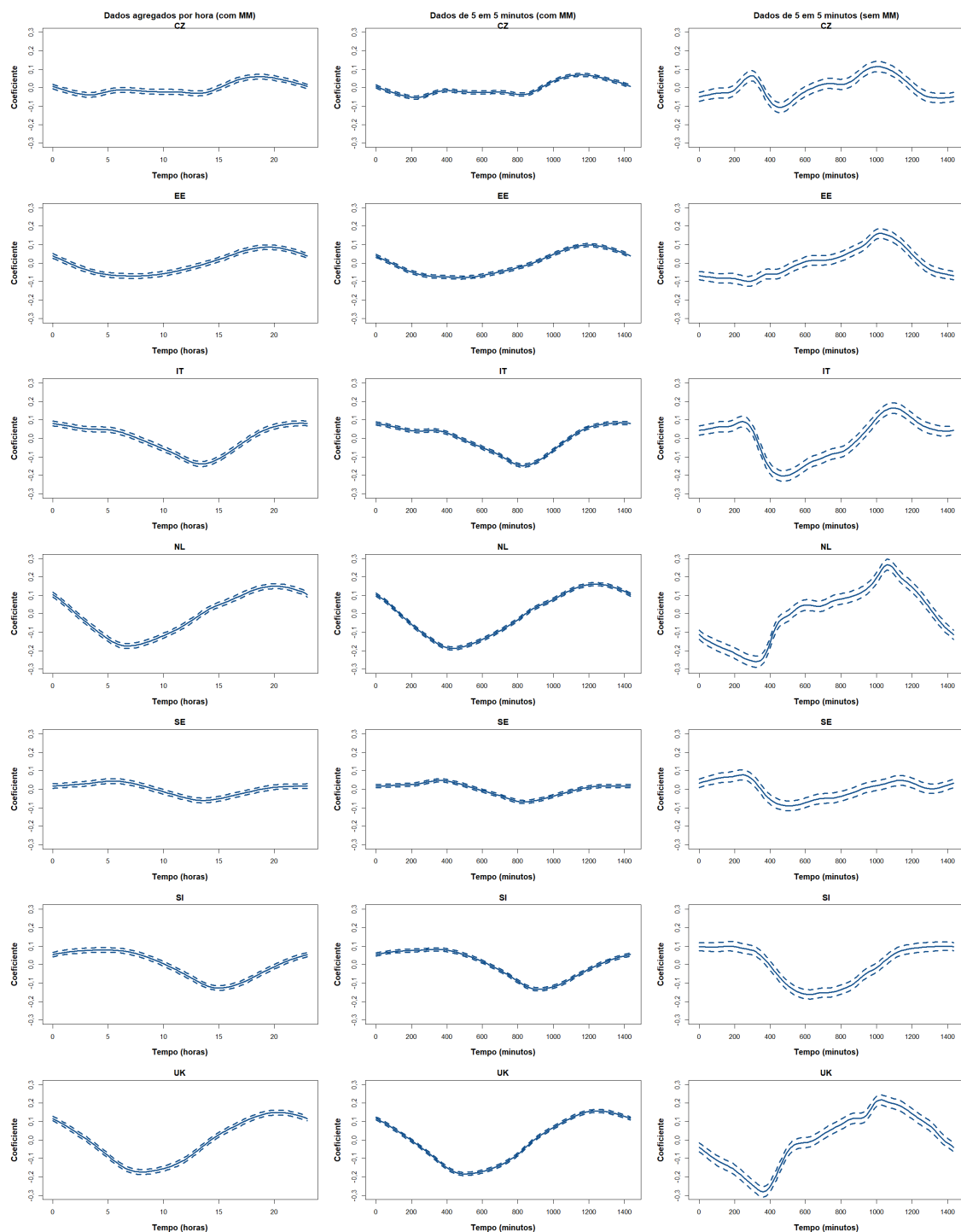


Figura 4.13: Efeito suave do tempo sobre a concentração de CO_2 estimado pelo modelo múltiplo para os diferentes conjuntos de dados para cada país em relação à categoria de referência. A linha representa o efeito estimado e as linhas a tracejado correspondem ao intervalo de confiança a 95%, tendo em conta a incerteza do termo constante.

A Figura 4.14 apresenta os efeitos suaves do tempo associados às diferentes variáveis categóricas, expressos relativamente às respetivas categorias de referência. Para as habitações com renovação, estima-se que a concentração de CO_2 é, em média, 1% a 2% superior à concentração nas habitações sem renovação durante o início do dia, entre as 0h e as 5h. A partir das 6h verifica-se uma diminuição progressiva deste efeito,

atingindo o valor mínimo por volta das 13h. Neste horário, estima-se que a concentração de CO₂ é, em média, 1% a 2% inferior à concentração nas habitações sem renovação. Após as 13h, observa-se um aumento gradual, aproximando-se novamente dos valores da categoria de referência no final do dia. Este padrão mantém-se consistente nos três conjuntos de dados, com ligeiras variações na amplitude.

Para as moradias, no caso dos dados suavizados, estima-se que a concentração de CO₂ é, em média, 7% inferior à concentração em apartamentos por volta das 6h e 8% superior por volta das 16h. Nos dados de 5 em 5 minutos sem aplicação de MM, este efeito é mais pronunciado, sendo $\approx 11\%$ inferior às 5h, e encontra-se ligeiramente deslocado no tempo, com o máximo a ocorrer por volta das 14h.

Para as habitações com patologias de humidade, no caso dos dados suavizados, estima-se que a concentração de CO₂ é, em média, 3% superior à concentração em habitações sem patologias por volta das 7h e 3% inferior por volta das 17h. Nos dados sem suavização, o efeito apresenta maior amplitude e encontra-se ligeiramente deslocado no tempo, sendo aproximadamente 3-4% superior no período das 0h-5h e 4% inferior entre as 12h-13h. No entanto, estas estimativas têm uma maior incerteza associada.

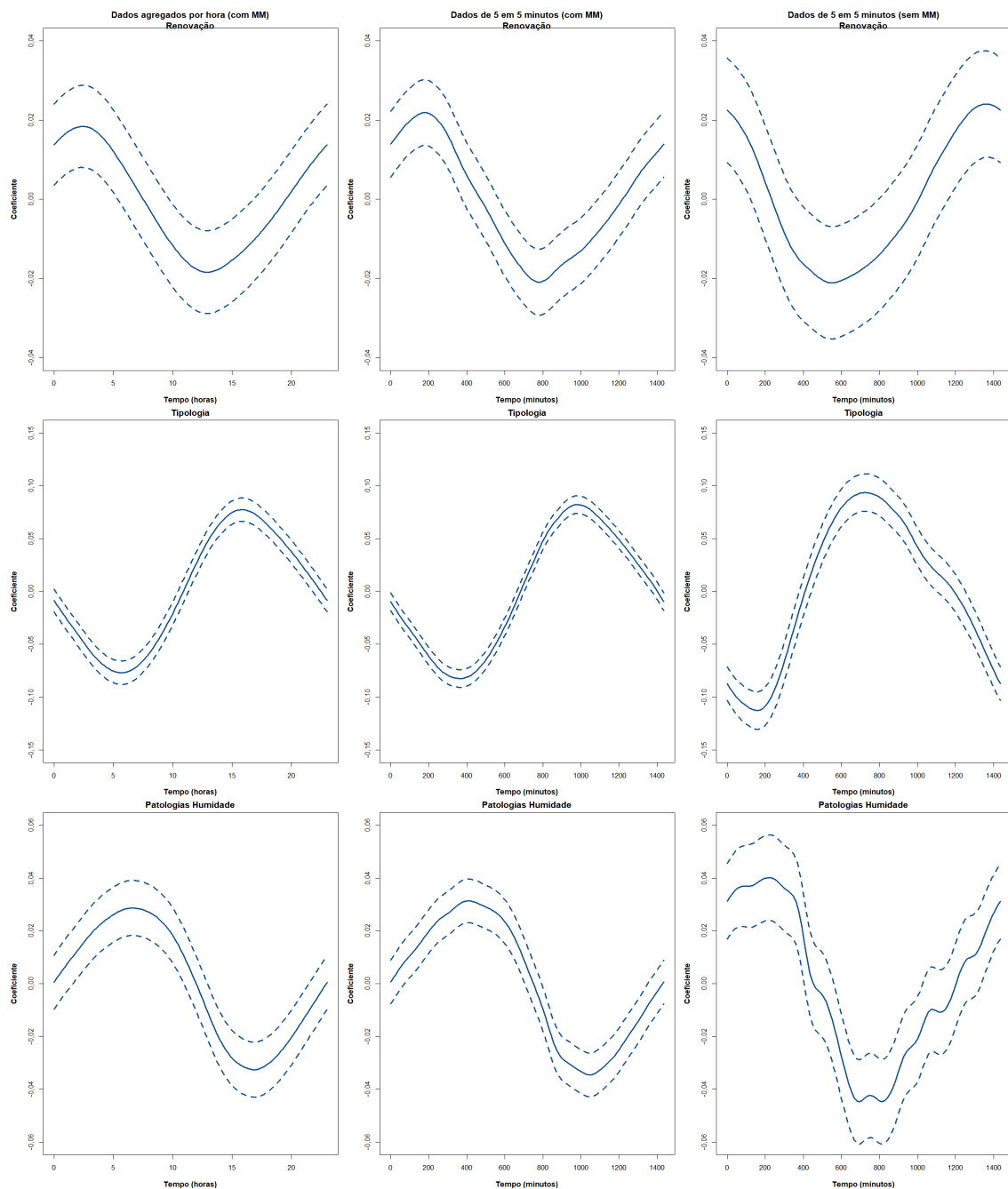


Figura 4.14: Efeito suave do tempo sobre a concentração de CO_2 estimado pelo modelo múltiplo para os diferentes conjuntos de dados para cada variável categórica em relação à categoria de referência. A linha representa o efeito estimado e as linhas a tracejado correspondem ao intervalo de confiança a 95%, tendo em conta a incerteza do termo constante.

A Figura 4.15 representa a estimativa do efeito suave da densidade de ocupação, na qual os três conjuntos de dados apresentam curvas muito semelhantes. Observa-se uma relação não linear com formato aproximadamente côncavo, caracterizada por uma diminuição da concentração de CO_2 à medida que a área disponível por ocupante aumenta até cerca de $140 \text{ m}^2/\text{pessoa}$. Para esse valor, estima-se que a concentração de CO_2 é, em média, 33% inferior à concentração de uma habitação com $f_6(\text{Densidade}) = 0$. Para densidades superiores, observa-se uma ligeira inversão da tendência, embora os

coeficientes se mantenham maioritariamente negativos. Esta região corresponde a um menor número de observações, o que se traduz numa maior incerteza das estimativas e em intervalos de confiança mais amplos.

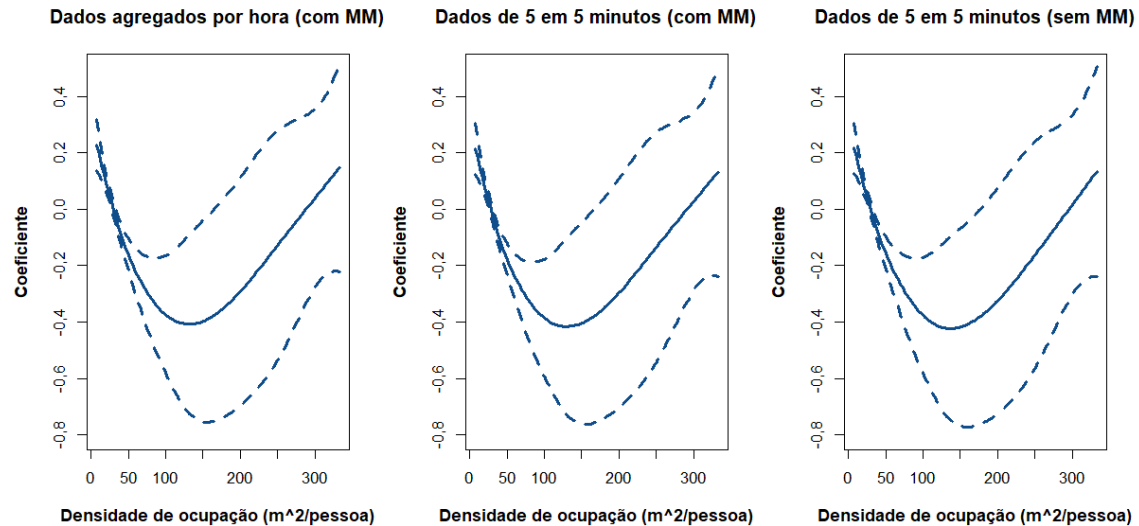


Figura 4.15: Efeito suave da densidade de ocupação sobre a concentração de CO_2 estimado pelo modelo múltiplo para os diferentes conjuntos de dados. A linha representa o efeito estimado e as linhas a tracejado correspondem ao intervalo de confiança a 95%, tendo em conta a incerteza do termo constante.

Por fim, a Figura 4.16 apresenta os QQ-plots dos efeitos aleatórios ao nível da habitação e do dia. Embora se verifiquem ligeiros desvios nas caudas, os pontos seguem, no geral, a reta de referência da distribuição normal, não existindo evidências de graves violações do pressuposto de normalidade dos efeitos aleatórios.

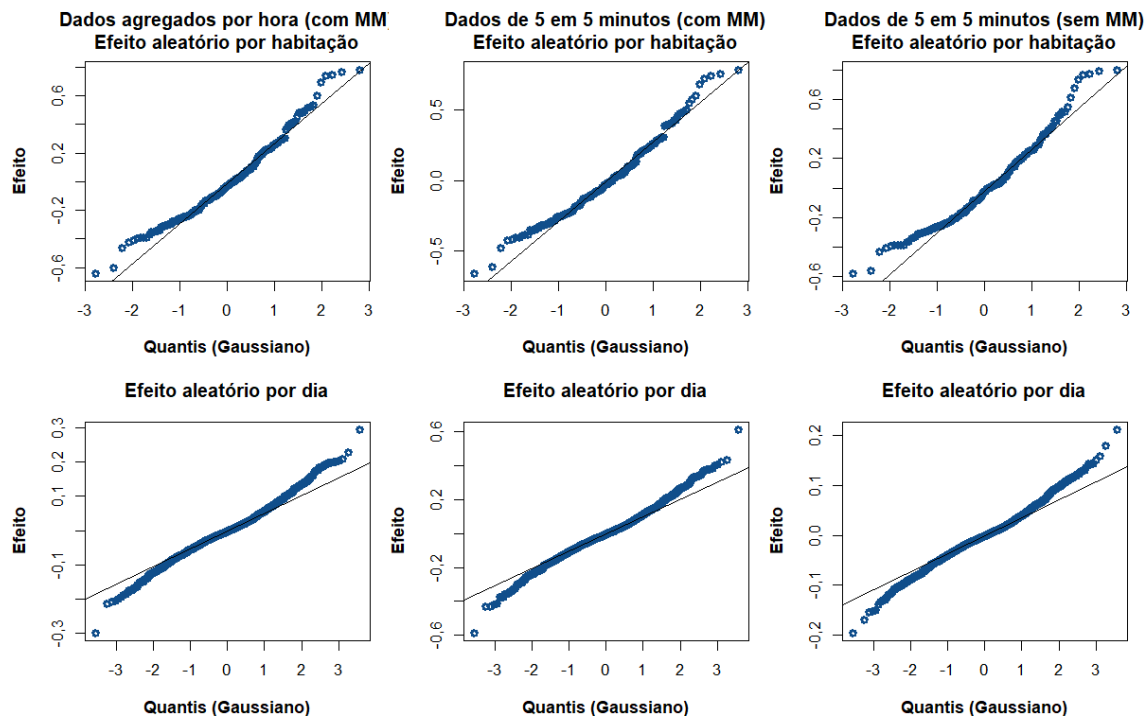


Figura 4.16: QQ-plot dos efeitos aleatórios estimados pelo modelo para os diferentes conjuntos de dados. A linha representa a reta de referência da distribuição normal.

Relativamente ao desempenho preditivo do modelo sob diferentes cenários de resolução temporal e suavização dos dados, a Tabela 4.9 apresenta os valores do RMSE, do MAE e do R^2_{adj} . Os modelos dos dados com suavização apresentam os menores erros e os maiores valores de R^2_{adj} , indicando uma melhor capacidade preditiva. Especificamente, o modelo ajustado aos dados de 5 em 5 minutos com MM apresenta o R^2_{adj} mais elevado, o que evidencia a melhor qualidade de ajustamento entre os três cenários. Por outro lado, o modelo sem suavização tem erros mais elevados e um R^2_{adj} consideravelmente inferior, o que sugere um pior desempenho, possivelmente devido à presença de ruído não filtrado.

Tabela 4.9: RMSE, MAE e R^2_{adj} do modelo múltiplo para os diferentes conjunto de teste.

Modelo	RMSE	MAE	R^2_{adj}
Dados agregados por hora (com MM)	0,2330	0,1781	0,7520
Dados de 5 em 5 minutos (com MM)	0,2362	0,1776	0,8330
Dados de 5 em 5 minutos (sem MM)	0,2910	0,2192	0,6130

A Figura 4.17 apresenta a relação entre os valores observados e os valores previstos por habitação. Nos dados suavizados, observa-se uma maior concentração de pontos em torno da linha para valores mais baixos de CO_2 . No entanto, para valores mais elevados, verifica-se uma maior dispersão dos pontos, sugerindo uma menor precisão do

modelo. Nos dados de 5 em 5 minutos sem MM, observa-se uma tendência para os valores previstos se situarem ligeiramente abaixo dos valores observados, o que evidencia uma ligeira subestimação sistemática do modelo nesse cenário.

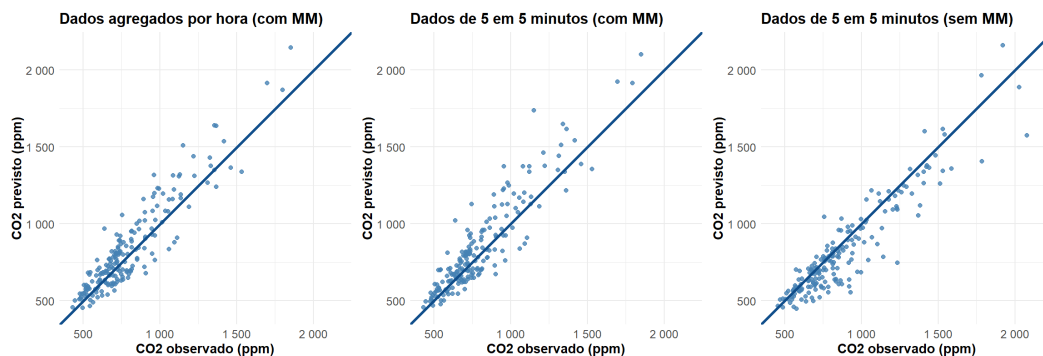


Figura 4.17: Gráficos de dispersão entre os valores observados e os valores previstos pelo modelo múltiplo, com base nos conjuntos de teste, por habitação. Cada ponto corresponde a uma habitação e a linha representa a previsão quando o valor previsto é igual ao valor observado.

O aumento da variância dos erros com o aumento dos valores previstos de CO_2 sugere a presença de heterocedasticidade. Neste sentido, a modelação da estrutura de variância poderia melhorar o ajustamento do modelo. Contudo, a função implementada não permite a inclusão desta estrutura, pelo que não foi considerada na presente modelação.

5 Discussão e conclusão

As diferenças observadas entre os dois *clusters* poderão estar relacionadas, em parte, com as características construtivas dos edifícios e com os respetivos padrões de ventilação. O *cluster* 1, caracterizado por uma maior prevalência de apartamentos e por uma menor presença de sistemas de ventilação mecânica, poderá estar associado a uma renovação do ar menos eficiente e, conseqüentemente, a uma maior concentração de CO₂. Esta interpretação é consistente com a literatura, que indica que as moradias tendem a apresentar taxas de renovação de ar superiores às dos apartamentos e que os sistemas de ventilação mecânica desempenham um papel importante na renovação do ar [21, 23].

Relativamente às características de ocupação, o *cluster* 1 apresenta uma menor densidade de ocupação, ou seja, uma menor área disponível por ocupante. Dado que, segundo a literatura, a emissão de CO₂ em espaços interiores está diretamente associada à presença humana, sendo um dos principais determinantes das concentrações de CO₂, os resultados obtidos são concordantes [20].

Além disso, o elevado número de habitações que utilizam gás natural e gás engarrafado no *cluster* 1 poderá também contribuir para o aumento das concentrações de CO₂, uma vez que a literatura associa a combustão de combustíveis no interior das habitações ao aumento das concentrações de CO₂, especialmente em ambientes pouco ventilados [22, 20]. Em contrapartida, o *cluster* 2 apresenta mais habitações que utilizam madeira/*pellets* para aquecimento. Embora esta fonte possa, isoladamente, constituir uma fonte relevante de emissão de CO₂, depende das condições de utilização e ventilação. Por outro lado, a presença de chaminés nas lareiras ou recuperadores pode favorecer a remoção de poluentes do interior, mitigando os níveis de CO₂ deste *cluster*.

O comportamento observado das concentrações de CO₂ ao longo do tempo, em ambos os *clusters*, poderá estar associado aos padrões diários de ocupação e ventilação das habitações. Durante o período noturno, a permanência prolongada dos ocupantes no interior da habitação, combinada com uma menor frequência de abertura de janelas, favorece a acumulação progressiva de CO₂ na divisão em estudo [20]. Em contrapartida, durante o período diurno, a redução da ocupação e a maior possibilidade de ventilação contribuem para a diminuição das concentrações observadas.

Adicionalmente, embora alguns países apresentem uma maior representação num dos *clusters*, observa-se uma sobreposição entre *clusters* em vários países, o que sugere que

a estrutura identificada não é determinada exclusivamente por fatores geográficos. Em vez disso, as diferenças parecem refletir sobretudo características construtivas e padrões de utilização das habitações, que variam dentro de cada país.

Os métodos de *clustering* e os modelos FoSR fornecem perspetivas complementares sobre os dados. Enquanto o *clustering* identifica perfis habitacionais globais com base em múltiplas características, a abordagem FoSR analisa o efeito de variáveis específicas na concentração de CO₂ ao longo do tempo. Assim, os perfis obtidos através do *clustering* representam uma simplificação da heterogeneidade dos dados, não permitindo uma interpretação temporal detalhada para cada característica.

Relativamente aos modelos FoSR, os resultados sugerem que as características habitacionais consideradas, em conjunto com a modelação funcional da evolução temporal do CO₂, explicam uma parte substancial da variabilidade observada nas concentrações. As características incluídas (tipologia da habitação, presença de renovação, presença de patologias de humidade e densidade de ocupação) correspondem a fatores amplamente reconhecidos na literatura como determinantes da ventilação e da QAI, o que reforça a coerência dos resultados obtidos.

A análise de sensibilidade demonstrou que a aplicação da suavização prévia por MM apresentou modelos com melhor desempenho e estimativas mais estáveis. A redução do desempenho observada nos dados não suavizados sugere a presença de variabilidade de curto prazo associada a oscilações aleatórias ou ruído de medição, o que pode dificultar a identificação dos padrões subjacentes. Apesar disso, os padrões gerais mantiveram-se consistentes entre os diferentes cenários, o que reforça a robustez dos resultados obtidos.

Um resultado particularmente relevante consiste na semelhança entre os efeitos estimados a partir dos dados horários e dos dados de 5 em 5 minutos após suavização. Este resultado sugere que a agregação horária preserva a estrutura essencial dos dados, podendo constituir uma alternativa mais simples e computacionalmente eficiente para estudos futuros.

Em termos de interpretação dos resultados, o comportamento temporal observado para o CO₂ é consistente com a literatura, que associa a maior permanência dos ocupantes e a menor frequência de ventilação durante a noite à acumulação progressiva de CO₂ no interior das habitações [20].

As diferenças na evolução temporal das concentrações entre países podem refletir variações nas características construtivas das habitações, nas práticas de ventilação, nos padrões de ocupação e nos hábitos diários dos ocupantes. Adicionalmente, fatores climáticos poderão também contribuir para as diferenças observadas. Contudo, não foi possível identificar um padrão claro que explique a semelhança entre determinados países, o que sugere que as diferenças nas concentrações de CO₂ resultam provavelmente da combinação de múltiplos fatores, para além da distribuição geográfica.

Nas habitações com renovação, o aumento da concentração de CO₂ durante o período noturno pode ocorrer devido à menor permeabilidade do edifício, uma vez que grande parte das renovações visa obter ganhos de eficiência energética através de uma maior estanquidade dos edifícios, como referido na literatura [23]. Pelo contrário, durante o dia há uma diminuição da concentração de CO₂, o que poderá estar relacionada com diferentes estratégias de ventilação adotadas nestas habitações. Assim, o impacto desta característica pode variar em função do tipo de intervenção realizada e dos comportamentos de ventilação dos ocupantes.

Nas moradias, as concentrações mais reduzidas de CO₂ observadas durante a noite são consistentes com a literatura, que reporta taxas de infiltração e renovação do ar geralmente superiores às dos apartamentos [23]. No entanto, o aumento das concentrações durante o período diurno sugere que fatores comportamentais, como os padrões de ocupação e as práticas de ventilação, poderão desempenhar um papel mais relevante do que as características construtivas durante esse período.

Relativamente às habitações com patologias de humidade, as concentrações mais elevadas de CO₂ observadas durante a noite poderão estar associadas a condições de ventilação menos favoráveis. De facto, a literatura indica que a humidade e as concentrações de CO₂ são influenciadas por fatores comuns, nomeadamente a ventilação e os padrões de ocupação, pelo que as associações observadas poderão refletir mecanismos subjacentes partilhados.

Por fim, a diminuição da concentração de CO₂ com o aumento da densidade de ocupação encontra-se em concordância com a literatura, que identifica a ocupação humana como a principal fonte de CO₂ em ambientes residenciais [20].

Apesar dos resultados obtidos contribuírem para uma melhor compreensão dos fatores associados às concentrações de CO₂ nas habitações, este trabalho também apresenta

algumas limitações. O questionário aplicado incidu essencialmente nos sistemas e produtos existentes nas habitações, bem como nas práticas habituais dos ocupantes, não garantindo que esses fatores estivessem efetivamente a ocorrer durante o período de monitorização das concentrações de CO₂. Deste modo, em estudos futuros, seria relevante considerar a aplicação de um questionário diário, preenchido pelos participantes, de forma a registar com maior precisão as práticas efetivamente realizadas e os sistemas utilizados ao longo do período de medição.

Outra limitação relevante do estudo consiste na dimensão reduzida da amostra para cada país, o que limita a representatividade dos resultados. Adicionalmente, as habitações foram selecionadas sem critérios uniformes entre os diferentes países, tendo sido apenas assegurado o recrutamento de famílias com crianças até aos 5 anos de idade, o que dificulta a comparação entre eles. Acresce ainda a seleção das habitações apenas numa única cidade por país, o que pode limitar a representatividade dos resultados a nível nacional, uma vez que não considera a variabilidade intra-país das características habitacionais e das condições ambientais. Além disso, as medições foram realizadas em diferentes períodos do ano entre países, sendo que em alguns casos um país corresponde a uma única estação e noutros a mais do que uma. Esta distribuição desequilibrada implica que a sazonalidade não é independente dos países, o que impede a separação dos efeitos do país e da estação no modelo. Ainda assim, a variabilidade sazonal pode influenciar as concentrações observadas, nomeadamente devido a alterações nas condições térmicas exteriores, na frequência da abertura de janelas e na utilização de sistemas de aquecimento ou ventilação. Assim, seria interessante realizar um estudo com medições repetidas nas mesmas habitações ao longo de diferentes estações do ano, para avaliar a variabilidade sazonal e o impacto nas concentrações de CO₂.

A modelação de fenómenos de natureza contínua através de análise funcional é uma área relativamente recente. Embora a análise de dados funcionais tenha tido os primeiros desenvolvimentos no início dos anos 90 do século XX, foi com a publicação do livro de Ramsay e Silverman (1997) [70] que esta teoria se estabeleceu. Os modelos de regressão funcional surgiram posteriormente e têm apresentado novas contribuições científicas até aos dias de hoje. Tanto quanto foi possível averiguar, existem poucos estudos que recorrem a modelos de regressão funcional para modelar concentrações de CO₂ ao longo do tempo [71, 72]. No caso presente, os modelos FoSR permitiram identificar relações temporais e efeitos não lineares que abordagens convencionais não conseguiriam representar adequadamente. Deste modo, os resultados obtidos evidenciam o potencial desta

metodologia para futuras aplicações na modelação de parâmetros de QAI.

Em suma, este trabalho permitiu analisar a relação entre as características habitacionais e as concentrações de CO₂ em habitações, através da aplicação de métodos de *clustering* e de modelos de regressão funcional. Os métodos de *clustering* permitiram identificar dois perfis habitacionais distintos, associados a diferenças estruturais e comportamentais. Adicionalmente, foi possível explorar diferenças na evolução temporal das concentrações de CO₂ entre estes perfis, onde se evidenciaram variações consistentes com as características das habitações. Por sua vez, os modelos FoSR permitiram quantificar a associação entre características habitacionais e a dinâmica temporal das concentrações de CO₂, na qual se observaram padrões temporais coerentes com o esperado com base na literatura. Além disso, a análise de sensibilidade demonstrou que os resultados são globalmente robustos a diferentes níveis de agregação temporal e à aplicação de suavização.

Conclui-se que as características habitacionais desempenham um papel relevante na variação das concentrações de CO₂, as quais poderão também influenciar as concentrações de outros poluentes interiores. Além disso, os métodos aplicados, especialmente os modelos FoSR, demonstraram ser uma ferramenta adequada para a análise desta relação em dados de parâmetros da QAI que apresentem natureza temporal contínua. Neste contexto, seria relevante estender esta abordagem a outros parâmetros de qualidade do ar, nomeadamente poluentes, para uma avaliação mais abrangente da qualidade do ar em habitações.

Referências Bibliográficas

- [1] World Health Organization. Air pollution data portal. <https://www.who.int/data/gho/data/themes/air-pollution>.
- [2] Pratap Kumar Padhy, Soumya Niyogi, Pulak Kumar Patra, and Markus Hecker. *Air Quality and Human Health*. Springer, 2024.
- [3] World Health Organization. Air quality, energy and health: Exposure & health impacts of air pollution. <https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health/health-impacts/exposure-air-pollution>.
- [4] World Health Organization et al. Air quality guidelines for europe. Technical report, World Health Organization, 2000.
- [5] World Health Organization. Air quality, energy and health: Health impacts. <https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health/health-impacts>.
- [6] Board on Population Health, Public Health Practice, Committee on the Effect of Climate Change on Indoor Air Quality, and Public Health. *Climate change, the indoor environment, and health*. National Academies Press, 2011.
- [7] Azhar Siddique, Maryam YM Al-Shamlan, Hamad E Al-Romaihi, and Haider A Khwaja. Beyond the outdoors: indoor air quality guidelines and standards—challenges, inequalities, and the path forward. *Reviews on Environmental Health*, 40(1):21–35, 2025.
- [8] Yinping Zhang, Philip K Hopke, and Corinne Mandin. *Handbook of indoor air quality*. Springer Nature, 2022.
- [9] Andrew Persily. Challenges in developing ventilation and indoor air quality standards: The story of ashrae standard 62. *Building and Environment*, 91:61–69, 2015.
- [10] Rune Vinther Andersen. Occupant behaviour with regard to control of the indoor environment. *Technical University of Denmark*, 2009.
- [11] Charles J Weschler. Characterization of selected organics in size-fractionated indoor aerosols. *Environmental Science & Technology*, 14(4):428–431, 1980.

- [12] Mohamed Sazif Mohamed Subri, Kadir Arifin, Muhamad Faiz Aiman Mohd Sohaimin, and Azlan Abas. The parameter of the sick building syndrome: A systematic literature review. *Heliyon*, 10(12), 2024.
- [13] Otto O Hänninen, Sari Alm, Klea Katsouyanni, Nino Künzli, Marco Maroni, Mark J Nieuwenhuijsen, Kristina Saarela, Radim J Srám, Denis Zmirou, and Matti J Jantunen. The expolis study: implications for exposure research and environmental policy in europe. *Journal of Exposure Science & Environmental Epidemiology*, 14(6):440–456, 2004.
- [14] World Health Organization. *Air Quality Guidelines: Global Update 2005*. WHO Regional Office for Europe, 2006.
- [15] World Health Organization et al. *WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization, 2021.
- [16] World Health Organization. Who housing and health guidelines, 2018.
- [17] Marta Fonseca Gabriel, Fátima Felgueiras, Raúl Batista, Cláudia Ribeiro, Elisabete Ramos, Zenaida Mourão, and Eduardo de Oliveira Fernandes. Indoor environmental quality in households of families with infant twins under 1 year of age living in porto. *Environmental Research*, 198:110477, 2021.
- [18] World Health Organization. Household air pollution. <https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health>.
- [19] Daniel Gehrt, Marco Hafner, Lucy Hocking, Evangelos Gkousis, Pamina Smith, and Jack Pollard. *Poor indoor climate, its impact on child health, and the wider societal costs*. RAND, 2019.
- [20] Health Canada. Residential indoor air quality guidelines: Carbon dioxide. Government of Canada, 2021.
- [21] Henry P Oswin, Lucien Glachant, Savinda Arambawatta Lekamge, Behnaz Alinaghipour, Saima Bukhat Khan, and Lidia Morawska. Using indoor co₂ concentration thresholds to understand and improve the air quality of public buildings: A practical approach. *Energy and Buildings*, page 116254, 2025.

- [22] Guofeng Shen, Subinuer Ainiwaer, Yaqi Zhu, Shuxiu Zheng, Weiying Hou, Huizhong Shen, Yilin Chen, Xilong Wang, Hefa Cheng, and Shu Tao. Quantifying source contributions for indoor co2 and gas pollutants based on the highly resolved sensor data. *Environmental Pollution*, 267:115493, 2020.
- [23] William W Nazaroff. Residential air-change rates: A critical review. *Indoor Air*, 31(2):282–313, 2021.
- [24] Institute of Air Quality Management (IAQM). Indoor air quality guidance: Assessment, monitoring, modelling and mitigation. 2021.
- [25] Juliana P Sá, Maria Conceição M Alvim-Ferraz, Fernando G Martins, and Sofia IV Sousa. Application of the low-cost sensing technology for indoor air quality monitoring: A review. *Environmental Technology & Innovation*, 28:102551, 2022.
- [26] Marta Fonseca Gabriel, Gonçalo Marques, David Filipe, Fátima Felgueiras, João Pedro Cardoso, Joana Azeredo, Giannis Kazdaridis, Polychronis Symeonidis, Stratos Keranidis, Peter Conradie, et al. Implementation of an iot architecture for promoting healthy air quality in 84 homes of families with children. *Building and Environment*, 266:112040, 2024.
- [27] Scott D. Lowther, Sani Dimitroulopoulou, Kerry Foxall, Clive Shrubsole, Emily Cheek, Britta Gadeberg, and Ovnair Sepai. Low Level Carbon Dioxide Indoors—A Pollution Indicator or a Pollutant? A Health-Based Perspective. *Environments*, 8(11), 2021.
- [28] Sunday Nunayon and Lexuan Zhong. A mixed-method systematic review of indoor environmental quality (ieq) gaps and opportunities for future research. *Sustainable Cities and Society*, page 106535, 2025.
- [29] Philomena M Bluysen. The need to go beyond the comfort-based dose-related indicators in our ieq-guidelines. *Building and Environment*, page 113349, 2025.
- [30] Dong Hyun Kim and Philomena M Bluysen. Clustering of office workers from the officair study in the netherlands based on their self-reported health and comfort. *Building and Environment*, 176:106860, 2020.
- [31] Dadi Zhang, Marco A Ortiz, and Philomena M Bluysen. Clustering of dutch school children based on their preferences and needs of the ieq in classrooms. *Building and Environment*, 147:258–266, 2019.

- [32] Khaiwal Ravindra, Preety Rattan, Suman Mor, and Ashutosh Nath Aggarwal. Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment international*, 132:104987, 2019.
- [33] Stiftelsen NILU. Identification of chemical and biological determinants, their sources, and strategies to promote healthier homes in europe, 2024.
- [34] Michael J Crawley. *The R book*. John Wiley & Sons, 2012.
- [35] Frank Bretz, Torsten Hothorn, and Peter Westfall. *Multiple comparisons using R*. Chapman and Hall/CRC, 2016.
- [36] Caroline X Gao, Shengqi Wang, Ye Zhu, Myriam Ziou, Shu Mei Teo, Catherine L Smith, Derek Chiu, Aline Talhouk, Sue M Cotton, and Dominic Dwyer. Ensemble clustering: A practical tutorial. *PsyArXiv*, 2024.
- [37] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.
- [38] Simona Balzano, Giovanni C Porzio, Renato Salvatore, Domenico Vistocco, and Maurizio Vichi. *Statistical Learning and Modeling in Data Analysis*. Springer, 2021.
- [39] Alboukadel Kassambara. *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda, 2017.
- [40] pam: Partitioning around medoids. <https://www.rdocumentation.org/packages/cluster/versions/2.1.8.1/topics/pam>. Documentação do pacote cluster (versão 2.1.8.1), disponível no site RDocumentation.
- [41] Özlem Akay and Güzin Yüksel. Clustering the mixed panel dataset using gower’s distance and k-prototypes algorithms. *Communications in Statistics-Simulation and Computation*, 47(10):3031–3041, 2017.
- [42] Everitt S Brian, Landau Sabine, Leese Morven, and Stahl Daniel. Cluster analysis. Library of Congress Cataloging-in-Publication Data, 2011.
- [43] hclust: Hierarchical clustering. <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust>. Documentação do pacote stats (versão 3.6.2), disponível no site RDocumentation.

- [44] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295, 2014.
- [45] Friedrich Leisch. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18, 2004.
- [46] flexmixedruns: Fitting mixed gaussian/multinomial mixtures with flexmix. <https://www.rdocumentation.org/packages/fpc/versions/2.2-13/topics/flexmixedruns>. Documentação do pacote fpc (versão 2.2-13), disponível no site RDocumentation.
- [47] Madalina Olteanu and Nathalie Villa-Vialaneix. Using sombrero for clustering and visualizing graphs. Technical report, 2015.
- [48] Madalina Olteanu and Nathalie Villa-Vialaneix. Som bound to realize euclidean and relational outputs. *Neurocomputing*, 147:15–30, 2015.
- [49] Pinyan Liu, Han Yuan, Yilin Ning, Bibhas Chakraborty, Nan Liu, and Marco Aurélio Peres. A modified and weighted gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. *BMC Medical Research Methodology*, 24(1):305, 2024.
- [50] Pinyan Liu. Dafi-gower distance source code. <https://github.com/Pinyan-Liu/DAFI-Gower-Distance>, 2024. GitHub repository; accessed 9 July 2024.
- [51] Keyvan Golaripour, Ebrahim Akbari, Seyed Saeed Hamidi, Malrey Lee, and Rasul Enayatifar. From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence*, 104:104388, 2021.
- [52] Kurt Hornik. A clue for cluster ensembles. *Journal of Statistical Software*, 14:1–25, 2005.
- [53] Frédéric Ros, Rabia Riad, and Serge Guillaume. Pdbi: A partitioning davies-bouldin index for clustering evaluation. *Neurocomputing*, 528:178–199, 2023.
- [54] Christian Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1):258–271, 2007.
- [55] cl_dissimilarity: Dissimilarity between partitions or hierarchies. https://www.rdocumentation.org/packages/clue/versions/0.3-67/topics/cl_

dissimilarity. Documentação do pacote `clue` (versão 0.3-67), disponível no site RDocumentation.

- [56] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.
- [57] Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. *The elements of statistical learning*. Springer New-York, 2 edition, 2009.
- [58] J.A. Navarro Alberto. *Multivariate Statistical Methods: A Primer, Fourth Edition*. Chapman and Hall/CRC, 4th edition, 2016.
- [59] Michael A. A. Cox and Trevor F. Cox. *Multidimensional Scaling*, pages 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [60] Ciprian M Crainiceanu, Jeff Goldsmith, Andrew Leroux, and Erjia Cui. *Functional data analysis with R*. Chapman and Hall/CRC, 2024.
- [61] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, New York, second edition, 2005.
- [62] Simon N Wood. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2017.
- [63] Simon Wood and Maintainer Simon Wood. Package ‘mgcv’. *R package version*, 1(29):729, 2015.
- [64] Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36, 2011.
- [65] Tailen Hsing and Randall L Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. Wiley Online Library, 2015.
- [66] Simon N. Wood. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*, 2025. R package version 1.9-4.
- [67] Jonathan D Cryer and Kung-Sik Chan. *Time series analysis: with applications in R*. Springer, 2008.
- [68] Jacolien van Rij. Acf: Checking & handling autocorrelation. <https://cran.r-project.org/web/packages/itsadug/vignettes/acf.html>, 2016.

- [69] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.
- [70] James O Ramsay and Bernard W Silverman. *Functional data analysis*. Springer, 1997.
- [71] Ram C Kafle, Keshav P Pokhrel, Netra Khanal, and Chris P Tsokos. Differential equation model of carbon dioxide emission using functional linear regression. *Journal of Applied Statistics*, 46(7):1246–1259, 2019.
- [72] Shaoli Wang, Mian Huang, Xing Wu, and Weixin Yao. Mixture of functional linear models and its application to co2-gdp functional data. *Computational Statistics & Data Analysis*, 97:1–15, 2016.

A Pré-processamento

Evolução temporal da concentração de CO₂ antes da remoção das habitações com sensores descalibrados

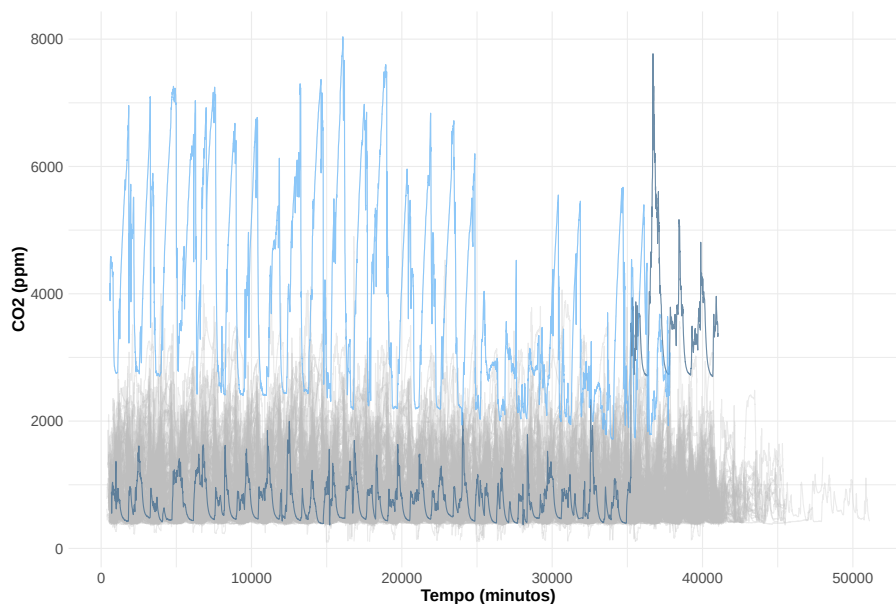


Figura A.1: Evolução temporal da concentração de CO₂ por habitação, ao longo de todo o período de monitorização, antes da remoção das habitações com sensores descalibrados. As linhas a cinzento representam todas as habitações incluídas na análise, enquanto as linhas a azul destacam as duas habitações excluídas.

B Caracterização dos dados e seleção de variáveis

Valores omissos

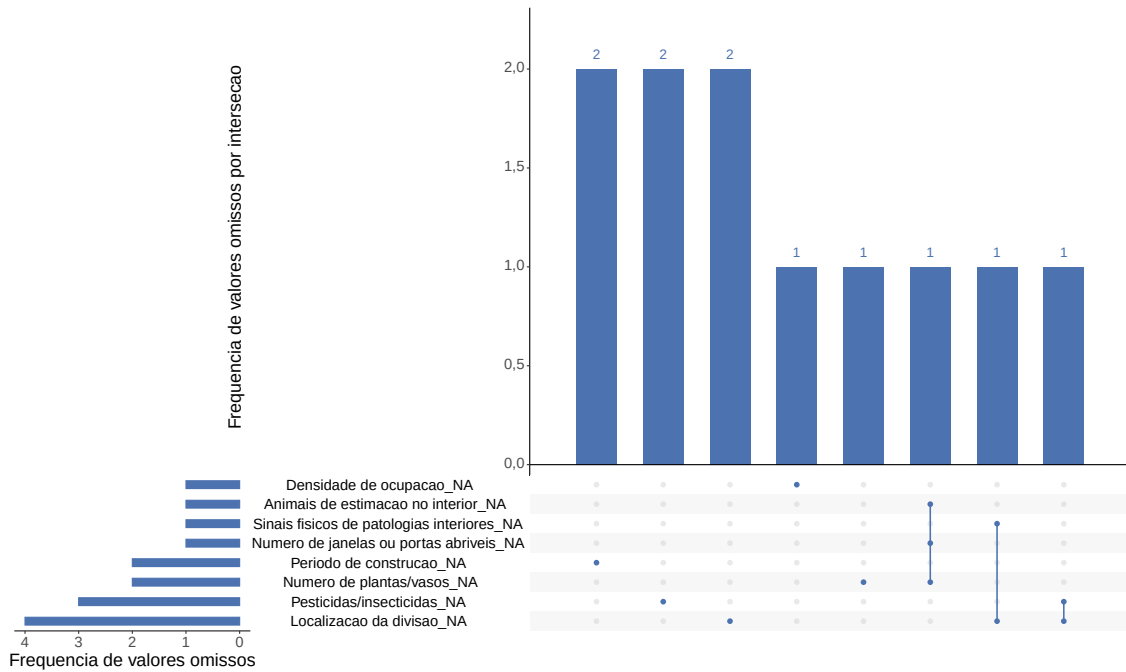
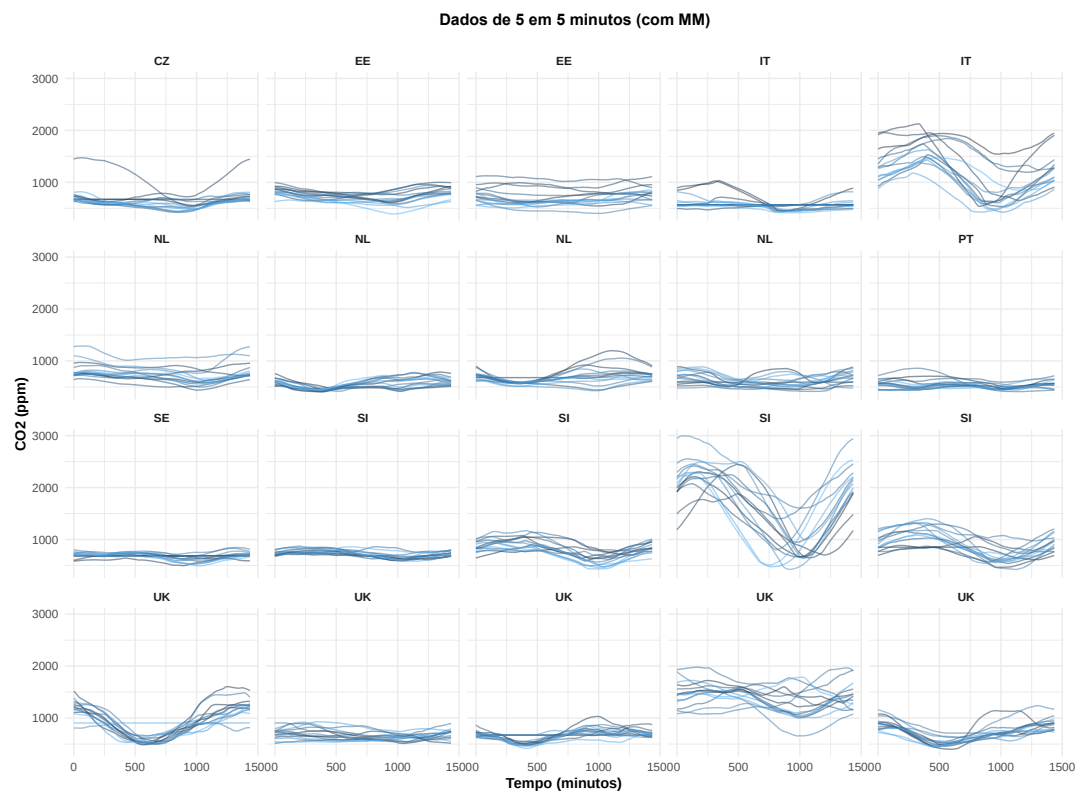
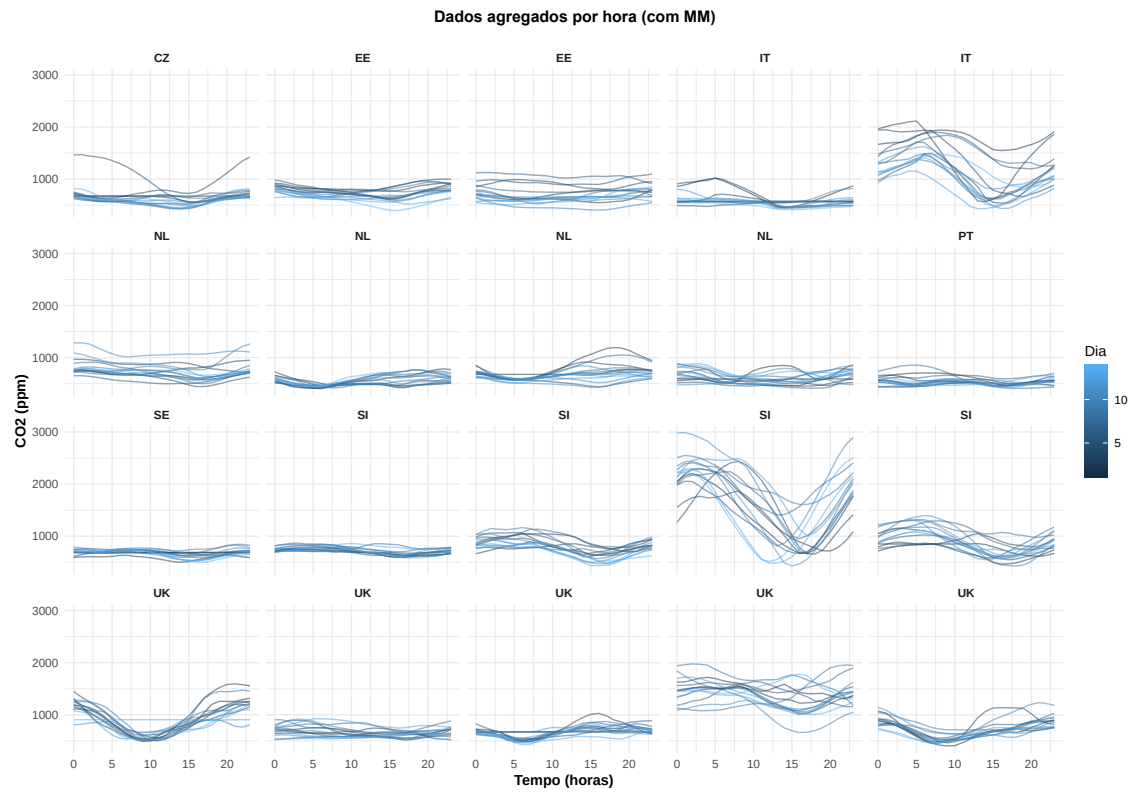


Figura B.2: Número de observações com valores em falta por variável e para cada combinação de ausência em múltiplas variáveis.

C Análise exploratória

Evolução temporal dos níveis de CO₂



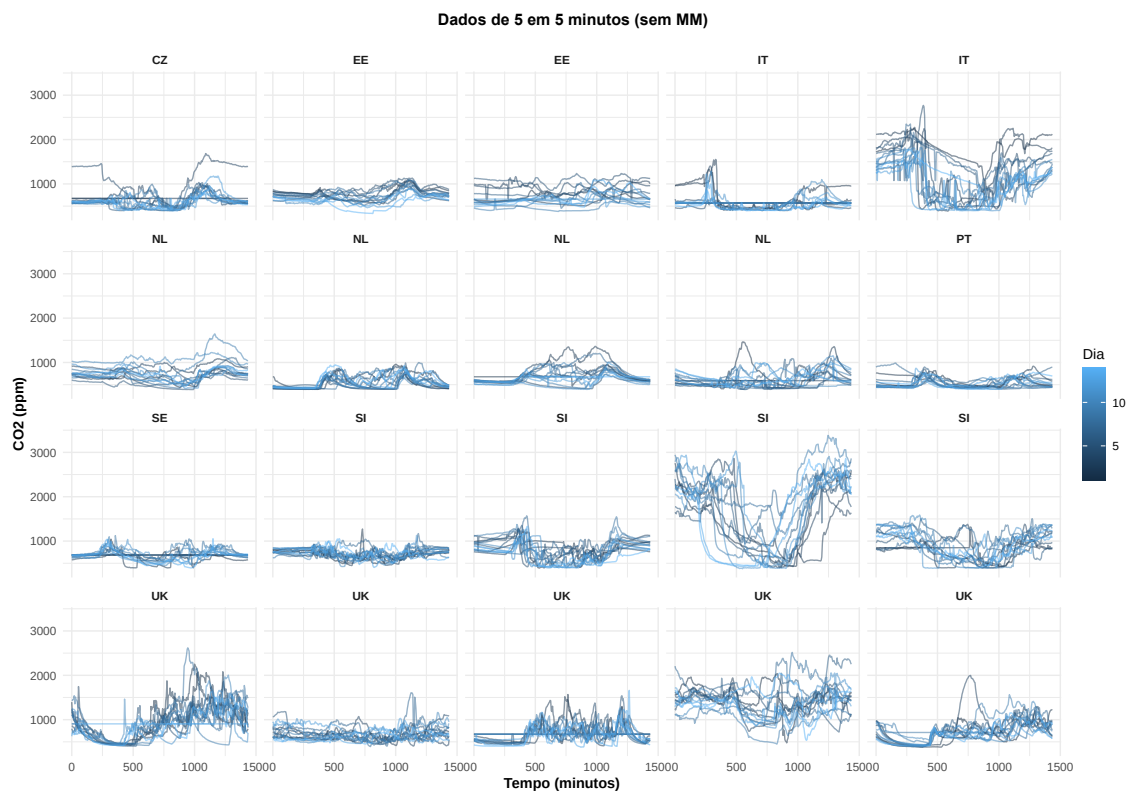
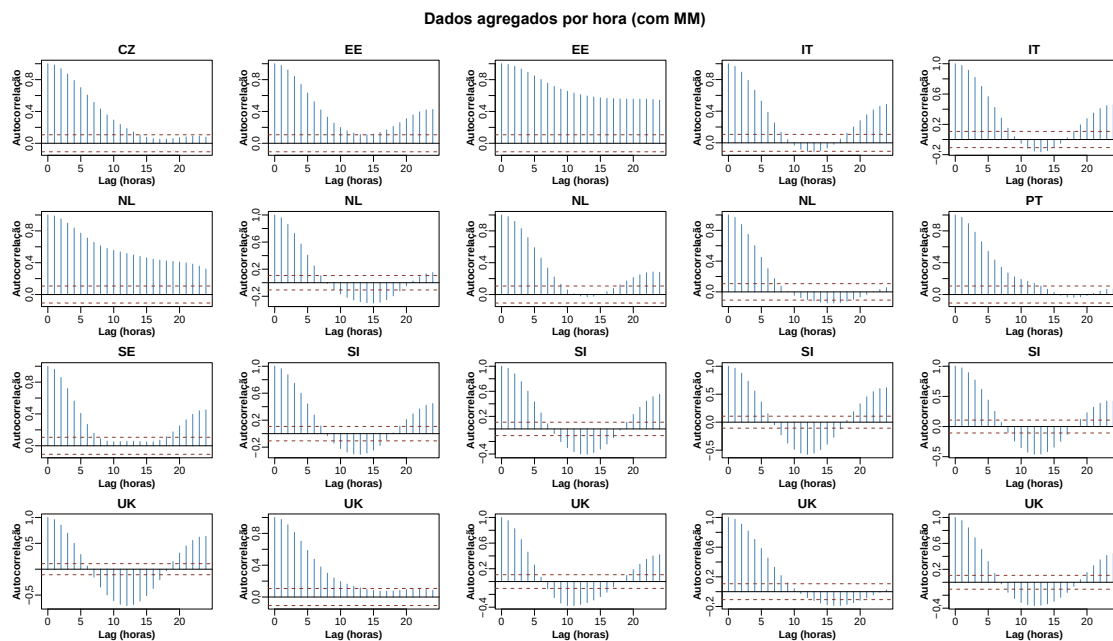


Figura C.1: Evolução temporal dos níveis de CO₂ para cada dia de cada habitação, num conjunto de 20 habitações seleccionadas aleatoriamente, considerando os diferentes conjuntos de dados.

ACF da concentração de CO₂



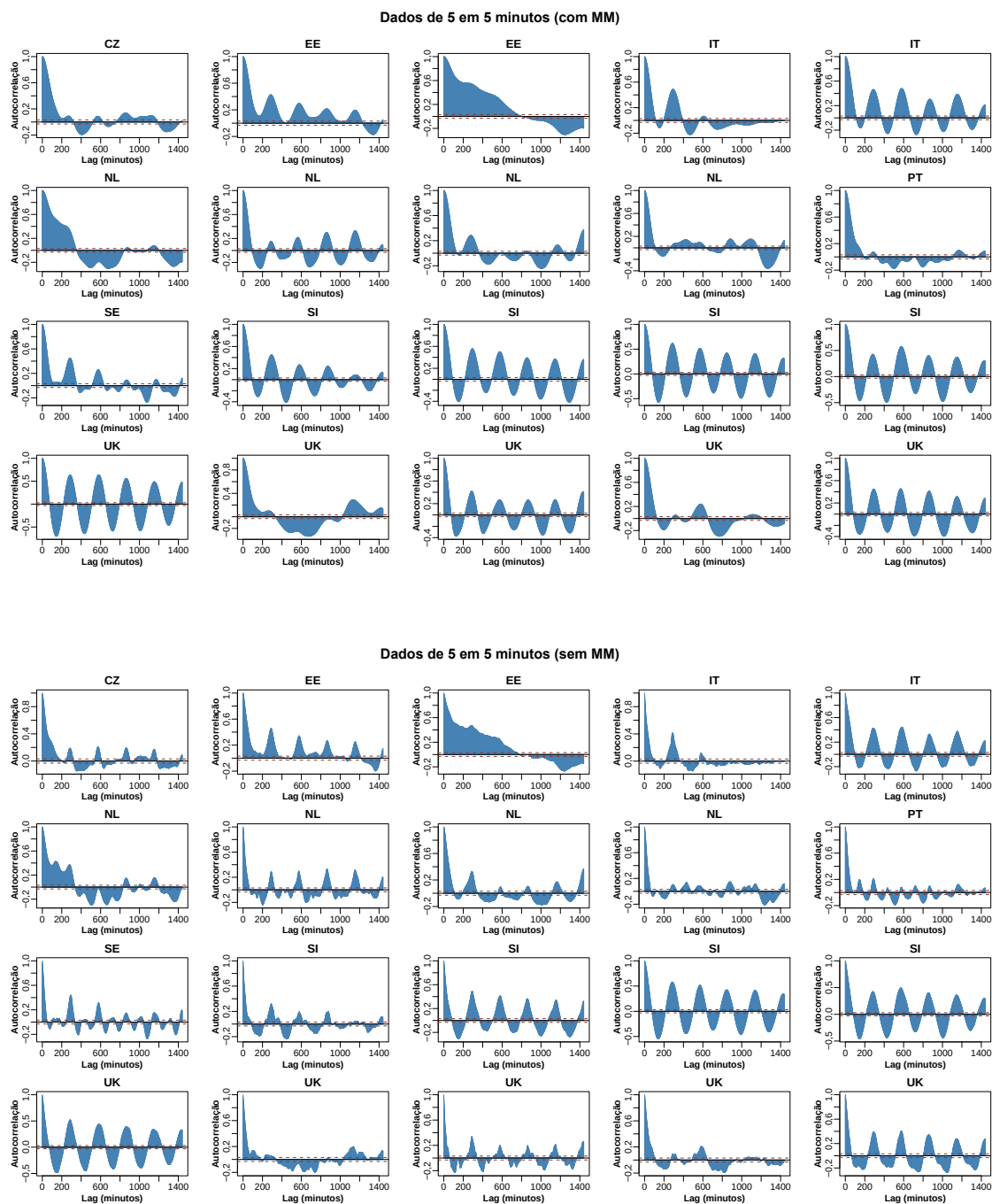


Figura C.2: ACF da concentração de CO₂ para 20 habitações selecionadas aleatoriamente, considerando os diferentes conjuntos de dados. As linhas tracejadas a azul representam o intervalo de significância a 95%, calculado como $\pm 1,96/\sqrt{n}$, onde n é o número de observações.

D Identificação de perfis habitacionais

Testes estatísticos para variáveis contínuas

Tabela D.1: Valores-p com correção de Holm dos testes estatísticos para as variáveis contínuas, incluindo diferenças entre todos os países (teste de Kruskal-Wallis) e comparações entre pares de países (teste de Wilcoxon-Mann-Whitney).

	CZ	EE	IT	NL	PT	SE	SI	UK
Período de construção	<0,0001*** ³							
CZ	-	-	-	-	-	-	-	-
EE	0,0673 ⁴	-	-	-	-	-	-	-
IT	1,0000 ⁴	0,0018** ⁴	-	-	-	-	-	-
NL	1,0000 ⁴	1,0000 ⁴	0,2147 ⁴	-	-	-	-	-
PT	0,9971 ⁴	1,0000 ⁴	0,0913 ⁴	1,0000 ⁴	-	-	-	-
SE	1,0000 ⁴	0,0128* ⁴	1,0000 ⁴	0,6524 ⁴	0,4135 ⁴	-	-	-
SI	1,0000 ⁴	0,0369* ⁴	1,0000 ⁴	0,9971 ⁴	0,4704 ⁴	1,0000 ⁴	-	-
UK	1,0000 ⁴	0,0008*** ⁴	0,6524 ⁴	0,0369* ⁴	0,0337* ⁴	0,2957 ⁴	0,2526 ⁴	-
Densidade de ocupação	<0,0001*** ³							
CZ	-	-	-	-	-	-	-	-
EE	1,0000 ⁴	-	-	-	-	-	-	-
IT	0,7111 ⁴	1,0000 ⁴	-	-	-	-	-	-
NL	<0,0001*** ⁴	0,0553 ⁴	0,0073*** ⁴	-	-	-	-	-
PT	<0,0001*** ⁴	0,0049*** ⁴	0,0004*** ⁴	0,6795 ⁴	-	-	-	-
SE	<0,0001*** ⁴	0,0520 ⁴	0,0023*** ⁴	1,0000 ⁴	0,7196 ⁴	-	-	-
SI	1,0000 ⁴	1,0000 ⁴	1,0000 ⁴	0,0426* ⁴	0,0089*** ⁴	0,0265* ⁴	-	-
UK	0,7111 ⁴	0,0547 ⁴	0,0169* ⁴	<0,0001*** ⁴	<0,0001*** ⁴	<0,0001*** ⁴	0,1325 ⁴	-
Número de plantas/vasos	<0,0001*** ³							
CZ	-	-	-	-	-	-	-	-
EE	1,0000 ⁴	-	-	-	-	-	-	-
IT	0,0003*** ⁴	0,0009*** ⁴	-	-	-	-	-	-
NL	1,0000 ⁴	0,1619 ⁴	<0,0001*** ⁴	-	-	-	-	-
PT	0,4370 ⁴	1,0000 ⁴	0,0966 ⁴	0,0466* ⁴	-	-	-	-
SE	0,0061** ⁴	0,0005*** ⁴	<0,0001*** ⁴	0,3506 ⁴	0,0001*** ⁴	-	-	-
SI	1,0000 ⁴	1,0000 ⁴	0,0035*** ⁴	0,2919 ⁴	1,0000 ⁴	0,0004*** ⁴	-	-
UK	<0,0001*** ⁴	0,0001*** ⁴	1,0000 ⁴	<0,0001*** ⁴	0,0141* ⁴	<0,0001*** ⁴	0,0003*** ⁴	-

Testes estatísticos: ³ Kruskal-Wallis; ⁴ Wilcoxon-Mann-Whitney
Diferença estatisticamente significativa: * valor-p < 0,05; ** valor-p < 0,01; *** valor-p < 0,001

Seleção do número de *clusters*

Tabela D.2: Coeficiente médio de silhueta e BIC para diferentes números de *clusters*.

Número de clusters	2	3	4	5	6	7	8
Coeficiente médio de silhueta							
PAM	0,2582	0,2081	0,2118	0,1822	0,1701	0,1964	0,1934
Hierárquico - Método de Ward	0,2847	0,2330	0,2570	0,2225	0,1879	0,1709	0,1617
Hierárquico - Ligação Completa	0,2847	0,2190	0,2194	0,1979	0,1880	0,1814	0,1862
SOM	0,2373	0,1821	0,1537	0,1335	0,1588	0,1440	0,1136
BIC							
FMM	7801	7658	7612	7657	7733	7757	7868

Clustering hierárquico

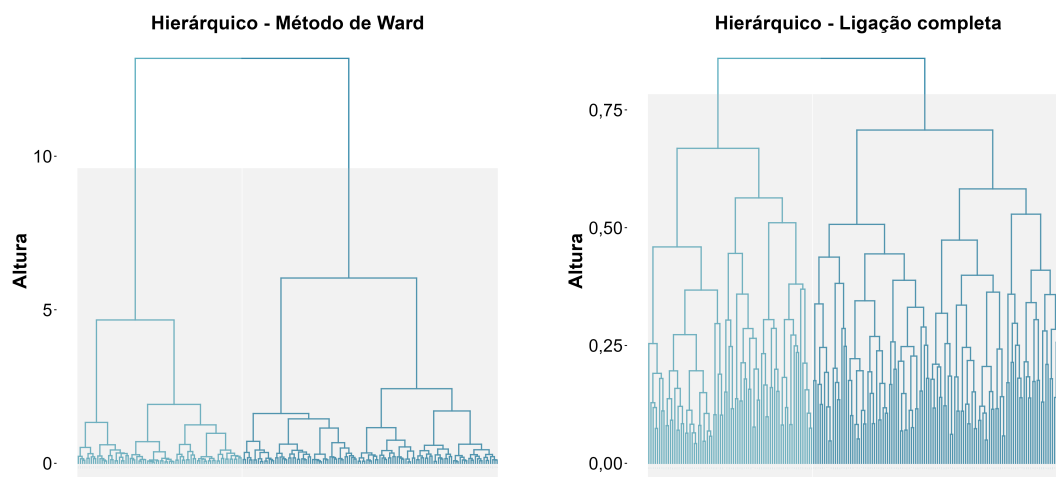


Figura D.1: Dendrogramas resultantes do *clustering* hierárquico com método de Ward (esquerda) e com ligação completa (direita). As regiões sombreadas destacam os pontos de corte para a formação dos dois *clusters*.

Self organizing map

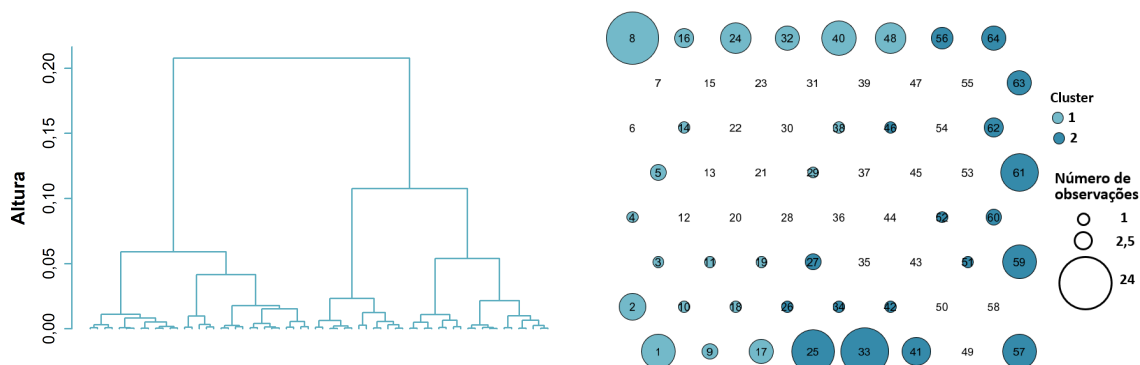


Figura D.2: Dendrograma resultante do *clustering* hierárquico com ligação completa aplicado aos neurónios (esquerda) e mapa de calor da rede de neurónios (direita), ambos referentes ao SOM. Neste último, cada neurónio é representado por um círculo cujo tamanho é proporcional ao número de observações que lhe estão associadas, enquanto a cor identifica o *cluster* ao qual pertence.

Vizualização dos *clusters*

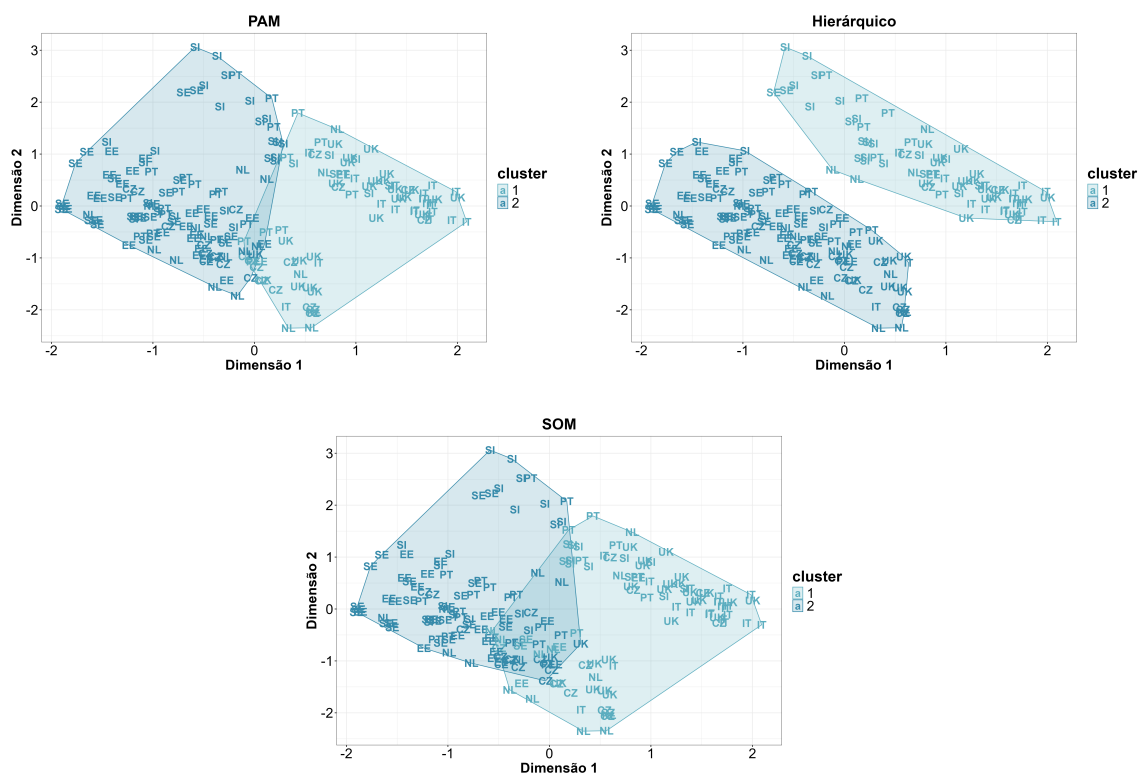


Figura D.3: *Clusters* obtidos pelos métodos PAM, hierárquico e SOM, representados nas duas primeiras dimensões após redução da dimensionalidade.

E Modelação do dióxido de carbono em função das características habitacionais

Significância dos termos suaves

Tabela E.1: Sumário dos termos suaves estimados pelo modelo múltiplo para os três conjuntos de dados.

Dados agregados por hora (com MM)					Dados de 5 em 5 minutos (com MM)				
	EDF	Ref.df	F	p-value		EDF	Ref.df	F	p-value
s(t)	18,893	22,000	76,135	<0,0001***	s(t)	21,354	22,000	944,593	<0,0001***
s(t):PaísCZ	15,930	22,000	13,949	<0,0001***	s(t):PaísCZ	21,332	22,000	214,243	<0,0001***
s(t):PaísEE	13,304	22,000	48,680	<0,0001***	s(t):PaísEE	20,184	22,000	971,342	<0,0001***
s(t):PaísIT	16,966	22,000	146,321	<0,0001***	s(t):PaísIT	21,385	22,000	3128,498	<0,0001***
s(t):PaísNL	17,200	22,000	248,060	<0,0001***	s(t):PaísNL	21,308	22,000	5170,703	<0,0001***
s(t):PaísSE	12,731	22,000	14,988	<0,0001***	s(t):PaísSE	20,466	22,000	263,003	<0,0001***
s(t):PaísSI	16,024	22,000	87,976	<0,0001***	s(t):PaísSI	21,118	22,000	1680,969	<0,0001***
s(t):PaísUK	17,350	22,000	230,143	<0,0001***	s(t):PaísUK	21,323	22,000	4692,257	<0,0001***
s(t):Renovação	6,939	22,000	14,218	<0,0001***	s(t):Renovação	16,429	22,000	324,392	<0,0001***
s(t):Tipologia	14,645	22,000	43,638	<0,0001***	s(t):Tipologia	20,642	22,000	583,881	<0,0001***
s(t):PatologiasHumidade	10,263	22,000	10,240	<0,0001***	s(t):PatologiasHumidade	19,895	22,000	126,933	<0,0001***
s(Densidade)	3,088	3,095	12,496	<0,0001***	s(Densidade)	3,102	3,104	12,527	<0,0001***
s(habituação)	183,376	189,000	157,156	<0,0001***	s(habituação)	183,281	189,000	2179,926	<0,0001***
s(dia,habituação)	1110,798	2674,000	0,785	<0,0001***	s(dia,habituação)	2132,569	2674,000	7,407	0,2290
EDF = Graus de liberdade efetivos; Ref.df = Graus de liberdade de referência; F = Estatística de teste					EDF = Graus de liberdade efetivos; Ref.df = Graus de liberdade de referência; F = Estatística de teste				
Diferença estatisticamente significativa: * p-value < 0,05; ** p-value < 0,01; *** p-value < 0,001					Diferença estatisticamente significativa: * p-value < 0,05; ** p-value < 0,01; *** p-value < 0,001				

Dados de 5 em 5 minutos (sem MM)				
	EDF	Ref.df	F	p-value
s(t)	16,470	22,000	41,423	<0,0001***
s(t):PaísCZ	19,110	22,000	17,964	<0,0001***
s(t):PaísEE	16,771	22,000	21,309	<0,0001***
s(t):PaísIT	18,932	22,000	24,726	<0,0001***
s(t):PaísNL	19,993	22,000	62,303	<0,0001***
s(t):PaísSE	15,306	22,000	8,826	<0,0001***
s(t):PaísSI	14,961	22,000	32,990	<0,0001***
s(t):PaísUK	19,788	22,000	30,503	<0,0001***
s(t):Renovação	4,359	22,000	4,560	<0,0001***
s(t):Tipologia	11,344	22,000	51,441	<0,0001***
s(t):PatologiasHumidade	14,339	22,000	12,575	<0,0001***
s(Densidade)	3,007	3,016	11,785	<0,0001***
s(habituação)	183,494	189,000	105,544	<0,0001***
s(dia,habituação)	786,558	2674,000	0,4430	<0,0001***
EDF = Graus de liberdade efetivos; Ref.df = Graus de liberdade de referência; F = Estatística de teste				
Diferença estatisticamente significativa: * p-value < 0,05; ** p-value < 0,01; *** p-value < 0,001				

Pressupostos

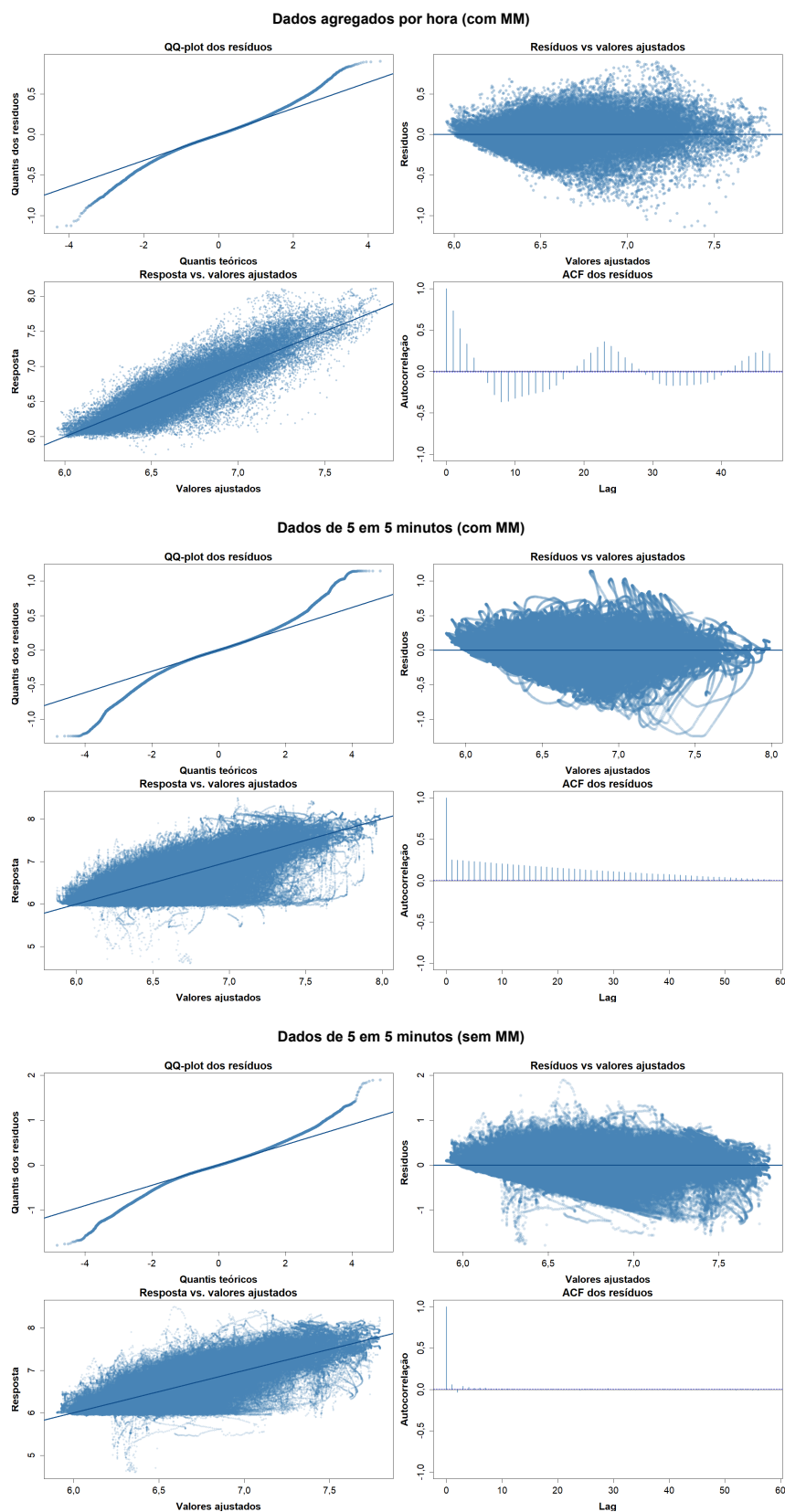


Figura E.1: Avaliação dos pressupostos do modelo múltiplo: normalidade dos resíduos (QQ-plot), homocedasticidade dos resíduos (resíduos vs valores ajustados), linearidade (resposta vs valores ajustados) e independência das observações (ACF).