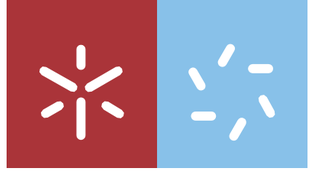
Modelos de previsão com Big Data proveniente de transações fianceiras

Joana Patrícia da Silva Simões

UMinho | 2019



Universidade do Minho
Escola de Ciências

Joana Patrícia da Silva Simões

Modelos de previsão com Big Data proveniente de transações financeiras

outubro de 2019

Universidade do Minho
Escola de Ciências

Joana Patrícia da Silva Simões

Modelos de previsão com Big Data proveniente de transações financeiras

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação de
Professora Doutora Cecília Castro
Professor Doutor Pedro Campos

outubro de 2019

Direitos de autor e condições de utilização do trabalho por terceiros

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.
Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Agradecimentos

À Professora Doutora Cecília Castro pelo tempo dispensado, pela ajuda e conhecimentos transmitidos ao longo da elaboração desta dissertação.

Ao Professor Doutor Pedro Campos pelo apoio, orientação e acolhimento no INE e no projeto.

Aos meus pais que me deram toda a liberdade e apoio.

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

Título: Modelos de previsão com Big Data proveniente de transações financeiras

A troca de serviços, por determinado período de tempo, com compensação monetária ou outra, através de plataformas digitais é um fenómeno bastante recente, designado por economia colaborativa. Esta realidade é ainda pouco compreendida, e o tipo de trocas/ transações incluídas neste conceito, ainda não são consideradas no cálculo de indicadores macroeconómicos como, por exemplo, o PIB. No entanto, há necessidade de estudar mais pormenorizadamente este tipo de economia para poder englobá-la no cálculo de indicadores de atividade económica, ou outros, já existentes. É este o principal objetivo do projeto ESSNet Big Data II – Financial Transactions Data, onde este trabalho se insere, sob a alçada do INE Porto.

Nesta tese, utilizam-se variáveis que podem ser consideradas dentro de um conceito de economia colaborativa. Tais variáveis foram introduzidas em modelos de efeitos fixos e em modelos de efeitos aleatórios, conseguindo explicar o PIB além de proporcionarem modelos com elevado poder preditivo. Uma vez que o foco se encontra na previsão, propõem-se aqui modelos de *machine learning* bastante recentes, árvores de regressão com inclusão de efeitos aleatórios, que demonstram também elevado poder preditivo, embora em comparação com os modelos de efeitos mistos apresentados, ficam ligeiramente aquém pela natureza linear dos dados utilizados.

Para a execução deste trabalho, recorreu-se a dados de levantamentos nacionais em caixas de multibanco, de compras através de terminais de pagamento automático e de dormidas nos estabelecimentos hoteleiros, ou seja, dados de transações financeiras que, em abstrato, são dados de economia colaborativa, pelo menos numa definição lata deste paradigma. Estes dados encontram-se agregados por regiões NUTS III e por ano, o que impõe que sejam tratados como dados em painel, tendo em conta a heterogeneidade entre as regiões.

Palavras-Chaves: economia colaborativa, dados em painel, modelos de efeitos mistos, árvores de regressão, amostra de treino/ amostra de teste.

Abstract

Title: Forecasting models from Big Data financial transactions

A new paradigm arises in economy, consisting in the exchange of services, for a certain period of time, with monetary compensation or other, through digital platforms. It's a recent phenomenon, called collaborative economy.

This reality is still poorly understood, and the type of transactions included in this definition are not yet considered in the calculation of macroeconomic indicators such as GDP.

However, it is necessary to study this type of economy in more detail in order to be able to include it in indicators of economic activities. This is the focus of the ESSNet Big Data II project - Financial Transactions Data, where this work is included under the guidance of NSI, Porto, Portugal.

In this thesis, we use variables that can be considered within a concept of collaborative economy. These variables were introduced in fixed effects models and random effects models, being able to explain the GDP and provide models with high predictive power. Since the focus is on prediction, very recent machine learning models are used here, like regression trees with random effects, which also show high predictive power. Although these models compared with mixed effects models, are slightly less powerful because of the linear nature of the data handled.

In this work, we used data from national withdrawals at ATMs, purchases through automatic payment terminals and overnight stays in hotel establishments, i.e. financial transaction data that, in abstract, are "collaborative economy" data, at least in a broad definition of this paradigm. These data are aggregated by regions and by year, which requires the use of a panel data approach, taking into account the heterogeneity between regions.

Keywords: collaborative economy, panel data, mixed models, regression trees, train/test set

Conteúdo

Lista de Figuras

Lista de Tabelas

1

Introdução

A procura de serviços partilhados, que se tem vindo a observar de uma forma cada vez mais incisiva, facilitada pela ligação entre pessoas continuamente *online*, através do acesso a redes *web*, a partir de computadores e *smartphones*, é uma realidade que afeta todos e que define um novo modelo de economia: a economia colaborativa, ou partilhada – *partilha* de bens.

Este tipo de economia, promove o sentido de comunidade, potencia o capital humano, gera confiança (incluindo em estranhos), privilegia a escolha e a conveniência, diminui a pegada de carbono mas também gera rendimentos.

Os modelos de negócios associados à economia de partilha não são os mesmos que os ligados à era de consumo de massa. É fundamental associar, de forma responsável e adequada aos modelos digitais e de uso partilhado, uma regulamentação. Há uma necessidade premente de busca de dados que permitam decidir com confiança.

Esta tese surge, assim, da necessidade de estudar e propor indicadores de contas nacionais, mercado de trabalho, turismo, transportes ....

É urgente a definição de indicadores de economia partilhada ou colaborativa, sendo importante complementar e corrigir estruturas de dados já existentes, com vista a uma análise fundamentada e rigorosa dos dados.

Com vista a um enquadramento do problema, é necessária uma definição de economia de partilha, o que não tem sido fácil pois existem diversos pontos de vista igualmente pertinentes.

Por exemplo, no que diz respeito às contas nacionais, o relatório *European Comission, Note on Measuring the digital collaborative economy* (2018), indica que apenas devem ser consideradas transações com compensação e taxas pagas à plataforma. Já em

relação ao mercado de trabalho, o mesmo relatório indica a necessidade de existir uma distinção entre trabalhadores independentes e trabalhadores da economia de partilha.

O INE faz parte de um projeto global, promovido pelo Eurostat, Gabinete de Estatística da União Europeia, designado por ESSnet Big Data II WPG (Workpackage G) – Financial Transactions Data, que tem como principal objetivo conhecer as fontes e a infraestrutura dos dados de transações financeiras dos países participantes.

Entendida a forma como os dados se encontram armazenados e organizados, e a disponibilidade de acesso pelos diferentes Institutos Nacionais de Estatística, fica possibilitado o acesso às fontes e ao seu potencial estatístico, de modo a avaliar e melhorar a qualidade de estatísticas já existentes e propor novas estatísticas de economia partilhada.

O *workpackage* G em causa está dividido em várias tarefas. Numa primeira fase é necessário investigar a existência e aceder aos dados de transações financeiras já existentes. Explorar e analisar estes dados é uma tarefa que naturalmente se impõe.

Numa segunda etapa, o objetivo é repetir este processo apenas para dados de plataformas de economia colaborativa.

Inserido neste trabalho global em que o INE está envolvido, trataram-se, nesta tese, dados que podem ser entendidos como fazendo parte de economia partilhada, sendo eles dados de dormidas em diversos estabelecimentos turísticos, apesar de não haver indicação sobre a forma como a transação foi efetuada, ou seja, não se sabe se as pousadas ou hotéis que forneceram estes dados de dormidas (através de inquéritos promovidos pelo INE), tiveram acesso aos clientes através de plataformas digitais, às quais terá sido atribuída, ou não, uma compensação. Dados sobre levantamentos em caixas multibanco nacionais e comprassss em terminais de pagamento automático, dados de transações financeiras, existem taxas que devem ser pagas, quer pelos utilizadores, quer pelas empresas, existindo, também aqui, uma troca de serviços que pode ser incluída dentro do conceito de partilha (bancos – empresas – consumidores).

O Capítulo 2 deste trabalho de tese é iniciado com uma breve síntese sobre diversos entendimentos do conceito e das implicações da economia partilhada.

No Capítulo 3 desta dissertação apresentam-se os Objetivos, a Metodologia e os Dados tratados neste trabalho.

O Capítulo 4 contém os Resultados dos modelos considerados para a resolução do problema.

No Capítulo 5 apresentam-se as principais conclusões do trabalho.

2

Economia Colaborativa

Com o impacto que a *internet* tem na vida das pessoas, o conceito de partilha deve também englobar comunicações e toda a partilha de bens não físicos (Stanoevska-Slabeva et al., 2017).

A partilha de bens em plataformas *web* que incluem o *Youtube*, o *Facebook* ou a *Wikipedia*, serviços como `Uber` [1] e `Airbnb` [2] entre muitos outros, consistem em *vender* ideias, conhecimento, serviços, fotografias, vídeos e outras informações de diverso tipo, é uma realidade bastante recente que necessita de ser compreendida e tratada para que se possa regulamentar.

Esta partilha envolve, na maioria dos casos, um pagamento, um retorno, um lucro.

Ora, alguns autores não concordam com a aplicação do termo partilha à economia, uma vez que partilha não deve envolver um pagamento (por definição) e, por isso, defendem que a economia partilhada não pode ser considerada uma verdadeira partilha (Stanoevska-Slabeva et al., 2017). Pode falar-se em *pseudo-partilha*, conforme Belk (2014), sendo caracterizada pela falta de sentimento de comunidade e reciprocidade, e motivada pelo lucro.

Vários autores consideram que apenas transações que envolvem algum tipo de compensação monetária (como por exemplo aluguer para férias) ou não monetária (por exemplo troca de casas) fazem parte da economia partilhada, enquanto que, para outros, trocas gratuitas (eg `Couchsurfing` [3]) estão incluídas neste conceito (Nguyen and Llosa, 2018).

Um primeiro aspeto onde existe desacordo relativamente ao conceito de economia

[1] https://www.uber.com/
[2] https://www.airbnb.pt/
[3] https://www.couchsurfing.com/

partilhada, é se esta deve apenas incluir trocas entre indivíduos (P2P, *peer-to-peer*) (eg `Blablacar` [4]) ou também entre empresas e indivíduos (B2C, *business-to-customer*) (eg `Zipcar` [5]).

Ainda não existe um consenso no que diz respeito à definição de economia partilhada, existindo vários termos para designar práticas muito semelhantes, por exemplo, *gig economy*, *mesh economy*, *peer-to-peer markets*, *collaborative economy*. Para alguns, estes termos definem o mesmo fenómeno enquanto que, para outros, os conceitos mencionados referem práticas distintas (Nguyen and Llosa, 2018).

Para além disso, alguns investigadores consideram que a economia partilhada apenas diz respeito a trocas mediadas por uma plataforma digital, enquanto que outros consideram as trocas feitas local ou pessoalmente, entre amigos, familiares ou conhecidos, também devem ser consideradas parte da economia partilhada.

Também não há concordância relativamente ao acesso dos bens, ou seja, para muitos a economia partilhada tem como base o aluguer a curto prazo mas, para outros, tanto este acesso temporário como a mudança de proprietário (eg `eBay` [6]) devem ser incluídos.

Serviços de aluguer prestados por empresas a consumidores, ou a partilha de bens entre amigos e familiares, não são fenómenos recentes, já existiam antes do conceito de economia partilhada. O que é novo é a troca de bens ou serviços entre indivíduos a uma escala global através de plataformas *web*.

As plataformas digitais facilitam estas transações, ao combinar dados relativos à oferta e à procura de serviços prestados por indivíduos, tornando possível que estranhos prestem e usufruam de serviços, tais como, partilhar carro (eg `Deboleia` [7], `Boleia.net` [8]) ou alugar casa de férias (eg `Airbnb`, `Homeaway` [9]). Estas trocas através das plataformas podem ser gratuitas (eg `Couchsurfing`) ou ter uma taxa associada (eg `Blablacar`).

Dentro do projeto global europeu referido na introdução, e de que o INE faz parte, é consensual que o termo economia partilhada seja substituído por economia colaborativa.

Este tipo de economia assenta numa plataforma colaborativa *online*, que facilita o contacto e transações entre indivíduos ou empresas.

A economia colaborativa envolve três componentes:

[4] https://www.blablacar.pt/
[5] https://www.zipcar.com
[6] https://www.ebay.com/
[7] http://www.deboleia.com/
[8] https://www.boleia.net/
[9] https://www.homeaway.pt/

1. Os provedores, indivíduos ou empresas que oferecem bens, recursos, tempo e serviços. Estes podem ser indivíduos a providenciar um serviço ou prestadores profissionais de serviços.

2. Os consumidores, indivíduos ou empresas que usufruem dos bens ou serviços prestados.

3. As plataformas, que servem como intermediário entre os participantes enunciados antes. As plataformas podem ter um papel mais passivo, apenas enumerando proprietários e quem procura os bens ou serviços, ou mais controlador ao monitorizar as transações que ocorrem. Além disso, as plataformas podem promover fins lucrativos ou não lucrativos.

Como se pode ver, na abordagem seguida por este projeto, a economia colaborativa engloba transações não só entre indivíduos mas também entre empresas e indivíduos, mas exclui transações em que os bens ou serviços são oferecidos para venda, ou seja, onde existe mudança de proprietário (eg `eBay`), considerando apenas trocas efetuadas por meio de uma plataforma digital.

O estudo já efetuado e acedido pelo projeto europeu em causa, teve como objetivo saber o desenvolvimento da economia colaborativa nos 28 países membros da União Europeia. Constatou-se que a maioria das plataformas operavam com base em relações apenas entre indivíduos, mas algumas também consideravam empresas como clientes.

As plataformas de economia colaborativa foram diferenciadas pelo tipo de objeto que está a ser partilhado, acomodação/alojamento, transporte, empréstimos e angariações (setor financeiro), serviços por profissionais (setor *online skills*).

Observou-se que a maior parte das plataformas operam no setor das finanças, seguido por plataformas que facilitam serviços prestados por profissionais e pelo setor do transporte.

No setor do alojamento verificou-se a existência de menos plataformas, talvez porque a plataforma `Airbnb` domina neste setor em todos os países membros. Também é de destacar que a maior parte das plataformas têm lucro contra uma pequena percentagem de plataformas sem lucro.

A economia partilhada oferece novas oportunidades de emprego, horários de trabalho flexíveis e novas fontes de rendimento. Para além disso, há maior conveniência no acesso aos bens ou serviços, maior flexibilidade traduzida pela poupança de tempo e esforço na procura e facilidade de pagamento. A oferta de novos serviços a preços mais acessíveis

(devido a maior competitividade) é também um ponto a favor da economia de partilha (*European Comission, Revenue and employment created by the collaborative economy, 2018*). O consumidor também tem uma redução nos custos ao não ser o proprietário, pois não tem gastos associados à reparação e manutenção dos bens (Oliveira, 2017).

Relativamente a benefícios ambientais, uma melhor utilização pode traduzir-se numa diminuição do uso de recursos naturais, por exemplo, a partilha de carro poderá reduzir o consumo de combustíveis fósseis, a troca de bens e a venda em segunda mão, poderá reduzir necessidades de produção (*European Comission, Note on Measuring the digital collaborative economy*, 2018).

Esta forma de economia parece ser mais sustentável do que a economia tradicional, uma vez que promove a reutilização dos bens (há uma utilização temporária de um bem que é propriedade de outro). Esta reutilização traduz-se numa diminuição do desperdício e do impacto ambiental causado pelo excesso de produção (Oliveira, 2017).

Contudo, a economia partilhada acarreta problemas, há dificuldade em distinguir entre consumidor e provedor, aquele que proporciona e promove o serviço, empregado ou trabalhador por conta própria e que serviços são prestados por profissionais, ou não. Relativamente a este último aspeto, cada país usa/define critérios diferentes para distinguir entre serviços profissionais e serviços prestados no âmbito da economia partilhada (dar boleias, alugar quartos ...). Para além disso, uma vez que ainda não existe uma definição consensual para este tipo de economia, existem atividades económicas que poderão, ou não, pertencer à economia colaborativa ou à economia tradicional, o que torna difícil identificar e calcular os indicadores económicos para medir a economia colaborativa (*European Comission, Note on Measuring the digital collaborative economy*, 2018). Por este motivo, há ainda problemas de regulamentação no que diz respeito a este novo paradigma de economia.

Baker (2015) escreveu a este respeito enumerando quatro principais tipos de problemas de regulamentação. São eles a regulação laboral, a proteção do consumidor, a proteção de propriedade e regras contra a discriminação. No primeiro tipo, os trabalhadores da economia partilhada são vistos, maioritariamente, como trabalhadores independentes e por isso não usufruem dos direitos de proteção e segurança dos restantes trabalhadores. Em segundo lugar, os serviços e bens prestados no âmbito da economia de partilha devem respeitar as leis de qualidade e segurança já estipuladas para atividades da economia tradicional. Na terceira categoria, o autor defende que os problemas referentes à proteção de propriedade ocorrem na sua maioria com serviços de aluguer como a `Airbnb`, enfati-

zando a possibilidade de falta de eficiência nas legislações que não permitem o aluguer a terceiros. Por último, as leis que proíbem a discriminação (por raça, género) e garantem o acesso a serviços a pessoas com deficiências, têm que ser ajustadas para garantir que as atividades dentro da economia partilhada não as contornem.

Há também o receio de que atividades que começaram com o intuito de partilhar bens e serviços a uma grande escala, se tornem em negócios focados no lucro em detrimento do altruísmo da partilha (Schor et al., 2016).

De um ponto de vista estatístico, o desafio que advém da economia partilhada é como complementar bases de dados em estruturas já existentes, como proceder ao estudo estatístico desses dados, nomeadamente utilizando algoritmos mais eficientes para o tratamento de dados de elevada frequência, com um número de preditores (ou variáveis explicativas) demasiado alargado, que não permitem muitas vezes evitar questões de multicolinearidade e, assim, usar técnicas tradicionais de *machine learning* com o elevado *expertise* existente na Estatística.

3

Metodologia e Dados

Um problema clássico dos economistas consiste em estabelecer uma relação entre o produto interno bruto e indicadores de consumo num país ou região.

O produto interno bruto, PIB, é o indicador, por excelência, da atividade económica de um país, do seu comportamento global e da sua economia.

De acordo com os Dados do Banco Mundial sobre contas nacionais e arquivos de dados da OCDE de Contas Nacionais, o PIB "é a soma do valor agregado bruto de todos os produtores residentes na economia mais quaisquer taxas de produtos e menos quaisquer subsídios não incluídos no valor dos produtos".

Em termos concetuais, o PIB deve ser um indicador exaustivo da economia, ou seja, deve englobar todo o tipo de atividade económica, mesmo que esta seja considerada ilegal. No entanto, o trabalho voluntário e atividades de prestação de serviços que não incluam uma remuneração não são consideradas aquando do cálculo do PIB. Esta é uma das críticas dirigidas ao PIB, que não está de acordo com o seu conceito e não inclui de forma exaustiva toda a atividade económica (INE, 2018).

A partir dos valores do PIB podem-se compreender as grandes assimetrias entre as regiões do país. A título de exemplo, a região de Lisboa e Vale do Tejo tem um PIB muito elevado relativamente a qualquer outra, o que pode estar relacionado com a maior produtividade desta região, por ter uma população em idade ativa mais numerosa, enquanto que, por exemplo, a região do Tâmega apresenta o valor de PIB mais baixo, possivelmente, devido ao facto de parte da população residente nesta região trabalhar em regiões vizinhas e os setores de atividade económica serem de baixa produtividade (Ramos and Rodrigues, 2001).

Neste trabalho, os indicadores de consumo devem incluir, tanto quanto possível,

variáveis da economia colaborativa.

As variáveis relativas a consumo são diversas. Na abordagem efetuada apenas se teve acesso a três, levantamentos nacionais em caixas multibanco, compras em terminais de pagamento automático e número de dormidas em alojamentos turísticos, que como foi referido na Introdução podem ser consideradas dentro de um conceito de economia colaborativa.

No sentido de tentar prever o valor do produto interno bruto em função das variáveis de consumo atrás identificadas, recorreu-se, naturalmente, a análise de regressão (no caso para dados longitudinais) e modelação com árvores de regressão, uma vez que o objetivo é encontrar, por um lado, um modelo que descreva bem os dados e, por outro, prever, com precisão, os valores da variável resposta.

3.1 Metodologia

Os dados a que se teve acesso foram organizados por regiões NUTS III e agregados por ano. Assim, o efeito desta organização em painel deve ser incluído nos modelos, uma vez que são dados longitudinais.

Neste capítulo é feita uma descrição sucinta dos métodos de explicação e previsão de dados longitudinais.

3.1.1 Análise de Regressão

A análise de regressão é um método analítico que visa estabelecer uma relação entre uma variável dependente e várias variáveis independentes de maneira a explicar determinado fenómeno.

Esta relação é expressa através de um modelo que associa a variável dependente/resposta com uma ou mais variáveis independentes/explicativas que, no caso de serem numéricas, se designam por covariáveis. A variável resposta é usualmente denotada por Y. As variáveis explicativas são usualmente denotadas por $\mathbf{X}$=(X_1, X_2,..., X_p), onde p é o número de variáveis independentes do modelo.

A relação entre a variável resposta e as variáveis explicativas pode ser representada pela equação (3.1).

$$Y = f(x_1, x_2, ..., x_p) + \epsilon, \tag{3.1}$$

onde a função $f(x_1, x_2, ..., x_p)$ designa a relação entre Y e $x_1, x_2, ..., x_p$, parte determinística do modelo, e ϵ diz respeito ao erro do modelo, parte aleatória.

A parte determinística, formada por uma ou mais variáveis observáveis é considerada fixa, enquanto que a parte aleatória ϵ admite uma distribuição de probabilidade.

No modelo (3.1), f pode ser uma função linear nos parâmetros, dizendo-se neste caso que o modelo é de regressão linear, ou não linear nos parâmetros (modelo não linear), podendo depender apenas de uma variável explicativa ou de várias (regressão múltipla).

A variável resposta Y pode ser quantitativa ou qualitativa, discreta ou contínua. No caso em análise trata-se de uma variável quantitativa contínua, pelo que o modelo é o de regressão linear múltipla, (3.2).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon, \tag{3.2}$$

onde $\beta_0, \beta_1, ..., \beta_p$ são os parâmetros ou coeficientes de regressão que se pretende estimar a partir dos dados.

As estimativas dos coeficientes de regressão são usualmente denotadas por $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$.

O valor $\hat{y}$ corresponde ao valor estimado. O i-ésimo valor estimado, $\hat{y}_i$, é dado por

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip}, \quad i = 1, 2, ..., n, \tag{3.3}$$

onde n é o número de observações, $x_{i1}, x_{i2}, ..., x_{ip}$ designam os valores das p variáveis explicativas para a i-ésima observação.

Quando se recorre à equação (3.3) para prever valores da variável resposta com base em valores observados das variáveis independentes, $\hat{y}$ fala-se em valor previsto.

O método dos mínimos quadrados, usualmente utilizado para estimar os coeficientes de regressão β, consiste em minimizar a soma dos quadrados dos resíduos (ver Chatterjee and Hadi, 2015, pg.89).

As propriedades dos estimadores de mínimos quadrados, assim como inferências estatísticas aplicadas a um determinado modelo de regressão, apenas são válidas se alguns pressupostos forem satisfeitos.

Os pressupostos usuais do modelo de regressão são:

- Normalidade dos erros, $\epsilon_i, i = 1, 2, ..., n$ tem uma distribuição Normal, com média nula.

- Homocedasticidade dos erros, também conhecido como o pressuposto da variância

constante ou da homogeneidade, isto é, os erros $\epsilon_1, \epsilon_2, ..., \epsilon_n$ têm a mesma variância desconhecida σ^2.

- Independência dos erros, ϵ_i e ϵ_j são independentes, para $i \neq j$.
- Não colinearidade das variáveis explicativas, os vetores $X_1, X_2, ..., X_p$ devem ser independentes.

Uma forma simples e eficiente de detetar anomalias na análise destes pressupostos é através da inspeção de gráficos dos resíduos.

Qualidade e seleção de modelos

Os métodos para averiguar a qualidade de ajuste de um modelo e proceder à seleção de modelos são vários e, tradicionalmente, são métodos "in sample", isto é, usam os mesmos dados que foram usados para a modelação.

De seguida, fala-se dos principais métodos para aferir a qualidade de ajustamento, assim como para a seleção de modelos.

Coeficiente de determinação – mede a relação entre a variável resposta Y e as variáveis explicativas $X_1, X_2, ..., X_p$, e é usualmente denotado por R^2.

Pode ser interpretado como a percentagem da variabilidade de Y que é explicada pelo conjunto das variáveis independentes.

Este coeficiente varia entre zero e um. Quando o modelo descreve bem os dados, obtém-se um valor de coeficiente próximo de um. Por outro lado, se não houver uma associação linear entre Y e as variáveis explicativas, R^2 será próximo de zero.

Uma medida relacionada com R^2 é o coeficiente de determinação ajustado, R^2_a. Este último coeficiente é usado para comparar modelos que têm um número diferente de variáveis independentes, uma vez que o valor de R^2 aumenta quando se adicionam mais variáveis independentes ao modelo. Ao contrário do coeficiente de determinação, o coeficiente ajustado não pode ser interpretado como a percentagem da variabilidade de Y explicada pelo modelo. Para mais detalhes sobre estes coeficientes consultar (Chatterjee and Hadi, 2015, cap3).

Critério de informação de Akaike – este método avalia cada modelo por si só. Num modelo com p coeficientes, o estimador de máxima verosimilhança da variância é

dado por

$$\widehat{\sigma}_p^2 = \frac{\text{SSE}_p}{n}$$

onde SSE_p representa a soma de quadrados residual do modelo com p coeficientes de regressão.

Akaike sugeriu medir a qualidade de ajustamento para os modelos de regressão, balançando o erro de ajustamento com o número de parâmetros do modelo, definindo o indicador:

$$\text{AIC} = -2\log L(M) + 2p(M) = n\log(\widehat{\sigma}_p^2) + 2p(M)$$

onde $L(M)$ é a função de log-verosimilhança dos parâmetros do modelo e $p(M)$ é o número de parâmetros do modelo.

O valor de p que minimiza o AIC especifica o melhor modelo. A ideia é penalizar a variância do erro por um fator proporcional ao número de parâmetros. A escolha do termo de penalização não é única, havendo várias na literatura tais como, AICc e BIC, em que o primeiro usa como fator de penalização $\frac{n+p}{n-p-2}$ e o segundo $\frac{p\log n}{n}$ (ver Chatterjee and Hadi, 2015, pg.305).
O critério de informação de Akaike assim como AICc e BIC, avaliam a qualidade de ajustamento de um modelo ao compará-lo com outros (ver Rawlings et al., 2001, pg.225). Ao usar estas medidas para selecionar um modelo, considera-se o melhor modelo aquele que apresentar menor valor.

3.1.2 Dados em Painel

Os dados em painel, também designados por medidas repetidas, ou longitudinais, dizem respeito a observações de indivíduos/objetos/sujeitos que são medidos repetidamente em diversas unidades de tempo. Medidas repetidas podem envolver medições efetuadas na mesma unidade de análise ao longo do tempo, ou medições efetuadas na mesma unidade alterando as condições experimentais. Dados transversais dizem respeito à medição de cada indivíduo sem ter em conta o caráter longitudinal dos dados (Diggle et al., 2002).

Neste contexto, quando se faz referência a indivíduos pode-se também estar a fazer referência a agregados familiares, empresas, regiões, países, entre outros, ou seja, unidades estatísticas.

Quando cada indivíduo é observado em todos os tempos do estudo, tem-se um estudo balanceado, pelo contrário, se existem observações em falta ou diferentes tempos de estudo, o estudo designa-se não balanceado.

A principal vantagem deste tipo de dados, comparativamente aos dados de corte transversal, é a flexibilidade em modelar diferenças de comportamento entre indivíduos (inter) e dentro de cada indivíduo (within) (Greene, 2003), permitindo medir efeitos que não são detetados em dados de corte transversal. Os dados em painel permitem mais graus de liberdade e mais eficiência. De facto, a variância total dos dados é decomposta na variância entre indivíduos e na variância dentro dos indivíduos.

Nos dados em painel observam-se muitos indivíduos em múltiplos períodos, pelo que se consegue explicar e prever os diferentes caminhos que uma variável resposta pode tomar ao longo do tempo para os vários indivíduos (Sela and Simonoff, 2012).

Os dados em painel requerem métodos de análise especiais, uma vez que as observações para um mesmo indivíduo podem estar correlacionadas (Diggle et al., 2002).

3.1.3 Modelos de Efeitos Mistos

Os modelos de regressão de efeitos mistos de regressão são os adequados para o tratamento de dados em painel, uma vez que permitem considerar a heterogeneidade entre indivíduos e as correlações dentro de cada indivíduo.

As diferenças entre objetos são representadas por efeitos aleatórios, as relações ao nível da população são representadas por efeitos fixos.

Define-se o modelo linear de efeitos mistos, com interceção aleatória e relação ao nível da população f, função conhecida linear nos parâmetros, como

$$y_{it} = \mathbf{b}_i + f(\mathbf{x}_{it}) + \epsilon_{it}, i = 1, ..., n, t = 1, ..., T$$

onde$i$ representa o indivíduo, neste caso a região, t o instante de tempo, neste caso o ano, e $\mathbf{b}_i$ é um vetor independente do tempo, com distribuição Normal.

No caso de apenas a interceção variar entre os sujeitos, $\mathbf{b}_i$ é a interceção específica do objeto.

O modelo linear de efeitos mistos assume uma forma paramétrica para a relação a nível da população $f = \mathbf{X}\boldsymbol{\beta}$ onde $\boldsymbol{\beta}$ é o vetor de efeitos fixos.

O termo de erro $\boldsymbol{\epsilon}_i$ é Normal multivariado com dimensão n, com vetor valor médio 0_n e matriz de covariância Λ_i. Os erros são independentes dos efeitos aleatórios $\mathbf{b}_i$. Num

estudo longitudinal, a matriz de covariância associada aos erros não tem que ser diagonal, refletindo o facto destes poderem não ser estatisticamente independentes uns dos outros, incorporando autocorrelação.

Em geral, o modelo de regressão linear de efeitos mistos é dado por:

$$\boldsymbol{y} = \mathbf{Z}\,\mathbf{b} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

onde $\mathbf{X}$ é uma matriz de desenho de efeitos fixos $n \times p$ e $\mathbf{Z}$ uma matriz de desenho de efeitos aleatórios de dimensão $n \times q$, onde q é o número de variáveis associadas aos efeitos aleatórios e n o número de indivíduos. No caso de apenas um efeito aleatório, a matriz de desenho reduz-se a um vetor de comprimento n.

Parâmetros do modelo

- Os p coeficientes de efeitos fixos, $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$, e a componente de variância do efeito aleatório, os elementos de Λ_i.

Os métodos mais usuais para estimação dos parâmetros num modelo linear de efeitos mistos são o método de máxima verosimilhança ou o método de máxima verosimilhança restrita (ver Diggle et al., 2002, pg.64 a 69).

O método clássico de máxima verosimilhança produz estimadores enviesados dos parâmetros da covariância e, por isso, há necessidade de recorrer ao método de máxima verosimilhança restrita, REML (Diggle et al., 2002).

Na estimação usando o método de máxima verosimilhança restrita, o interesse é na estimação dos efeitos aleatórios e não dos efeitos fixos.

O espaço de parâmetros é restrito aos efeitos fixos acima de um determinado patamar. Neste espaço restrito são procurados os valores dos parâmetros de efeitos aleatórios, neste caso a variância, num conjunto que maximiza a log-verosimilhança dos dados.

Uma vez que este método depende dos valores dos parâmetros de efeitos fixos, não pode ser usado para comparar modelos que sejam diferentes na estrutura de efeitos fixos.

Pressupostos do modelo de efeitos mistos

1. O modelo de efeitos mistos contém pelo menos mais uma variável aleatória que o modelo de regressão linear.

2. O erro de um modelo de efeitos mistos inclui a hipótese de que as observações dentro do mesmo nível (sujeito/indivíduo) estão potencialmente correlacionadas.

3. Os modelos de efeitos mistos estão desenhados para incluírem esta correlação sem violarem a hipótese de independência das observações.

4. As observações são independentes das outras observações exceto no que diz respeito às autocorrelações específicas dos erros.

5. Existe ainda uma outra hipótese de independência. Os efeitos associados à variável sujeito são não correlacionados com as médias dos efeitos fixos.

6. Todas as outras hipóteses dos modelos de efeitos mistos são as dos modelos lineares.

Na análise de regressão as variáveis explicativas assumem-se fixas e o erro é a única parcela que explica efeitos não observados. Assume-se ainda que os erros do modelo são independentes e normalmente distribuídos com variância constante.

Quando se trata de dados longitudinais, os modelos que retratam este tipo de dados podem conter mais do que uma parcela aleatória para ter em consideração efeitos que não são explicados pelas variáveis explicativas. Este problema de conseguir explicar variáveis não observáveis é um dos motivos para se recorrer aos dados longitudinais.

Efeitos aleatórios vs Efeitos fixos

Se $\mathbf{b}_i$ é tomado como fixo, potencialmente correlacionado com as covariáveis, então, trata-se de um modelo linear de efeitos fixos. Caso contrário, e sob as mesmas condições sobre f, se se assume que os efeitos $\mathbf{b}_i$ são não correlacionado com as covariáveis, tem-se um modelo linear de efeitos aleatórios, designado também por efeitos mistos, uma vez que os parâmetros em $\mathbf{X}\boldsymbol{\beta}$ são efeitos fixos.

Os modelos de efeitos mistos, quando apropriados, são mais eficientes do que os modelos de efeitos fixos, porque o número de parâmetros estimados num modelo de efeitos fixos aumenta com a inclusão de mais objetos/indivíduos.

3.1.4 Árvores de Regressão

As árvores são um método de estimação baseado em algoritmos de *machine learning*, que tem sido bastante usado para a previsão em problemas mais complexos, com

comportamento não linear, assim como com dados de alta frequência com um elevado número de variáveis.

O facto de não estar subjacente um modelo estatístico aos erros de previsão, a facilidade de interpretação e a estrutura intuitiva de uma árvore, permitem decidir que variáveis são mais importantes para explicar o fenómeno em estudo, de que modo estão relacionadas, independentemente de eventuais problemas de multicolinearidade. Tal tem permitido resolver problemas, classicamente resolvidos com metodologia estatística, quando o objetivo é a previsão.

Em suma, os modelos baseados em árvores são usados para tomar decisões, explorar os dados e fazer previsões.

Caso a variável resposta seja qualitativa tem-se uma árvore de decisão em que o output é uma categoria. Por outro lado, numa árvore de regressão, tem-se uma variável quantitativa como variável dependente, obtendo-se um escalar como resultado. Ou seja, a estrutura da árvore é a mesma, apenas diferindo no resultado.

Uma árvore consiste num nó raiz, ramos, nós (locais onde os ramos são divididos) e folhas. Cada nó interno, que não é uma folha, pode ser partido em dois ou mais ramos. Nas árvores binárias cada nó interno é partido em apenas dois ramos. Cada um desses ramos corresponde a uma instrução `if-else; true-false`.

Uma árvore é uma estrutura hierárquica em que cada nó particiona os dados resposta com base numa determinada característica preditora, de forma a que as respostas sejam mais homogéneas entre si. Para tal, torna-se necessário considerar uma medida de "impureza"que, no caso dos dados resposta serem numéricos, é, tipicamente, medida pela variância dos dados.

O nó raiz e os nós internos estão associados a condições de teste, binárias, e cada folha está associado a um resultado, categórico ou numérico, consoante a árvore é de decisão ou de regressão.

A primeira utilização desta técnica remonta a 1963 (Ferreira, 1999) no âmbito das ciências sociais por Morgan and Sonquist (1963).

No entanto, foram os trabalhos desenvolvidos por Quinlan (1986) e Breiman et al. (1984) que tiveram um contributo decisivo na popularização do uso das árvores em problemas de classificação e de regressão. A aplicação do método das árvores a problemas de regressão foi iniciado em Morgan and Sonquist (1963) com o algoritmo AID (Automatic Interaction Detection), mas apresentando bastantes falhas. Mais tarde, Breiman et al. (1984) desenvolveram o programa CART (Classification and Regression Trees) que

se encontra implementado em `R` e tem sido usado como base de inúmeros algoritmos.

O algoritmo CART proporciona um método não paramétrico de modelação da relação de base populacional, função f (3.1), com base num procedimento de *machine learning* de *busca gulosa*. Este método processa de forma exaustiva todas as possíveis partições, terminando, apenas quando as folhas são puras.

Como é evidente este procedimento pode conduzir a árvores de dimensão extremas, em que muitos ramos conduzem a situações de menos pureza que os anteriores, pelo que se torna necessário proceder a *podagem* da árvore, de modo a prevenir *overfitting* (Hajjem et al., 2014)

Os métodos de podagem, no algoritmo CART, são baseados em *cross validation*.

Além desta questão, é necessário definir parâmetros que condicionem o tamanho das árvores. Entre estes encontra-se o *complexity parameter*, `cp`, que vai permitir escolher o critério ótimo para a poda da árvore.

Para covariáveis contínuas, as partições tomam a forma $x \leq c$, onde c é um ponto de corte/ separação específico (Hajjem et al., 2014). As partições continuam até se atingir um determinado valor de `cp`, que controla a proporção de variabilidade explicada pela árvore.

Uma vantagem dos algoritmos implementados nos métodos de árvores de regressão é que estão preparados para lidar com observações em falta nos preditores e não requerem uma preparação dos dados. Neste caso, os dados foram logaritmizados apenas para poder ser possível efetuar comparação entre os vários modelos.

A variabilidade associada às árvores, ou seja, uma pequena alteração nos dados pode resultar em partições completamente diferentes e, consequentemente, em árvores diferentes, é um problema que tem vindo a ser objeto de estudo levando à consideração, por exemplo, de *Random Forests* fora do âmbito deste trabalho.

Amostra de treino e amostra de teste

Um problema que pode ocorrer quando se faz modelação, usando todos os dados disponíveis, é o problema de *overfitting*, obtendo-se um resultado bastante otimista para o modelo (o modelo com um ajuste muito bom) mas, por vezes, com um fraquíssimo poder preditivo. Além disso, as estatísticas de qualidade de ajustamento e comparação de modelos utilizadas neste caso (quando todos os dados disponíveis são usados para a modelação), são "in sample".

Uma opção para solucionar esta questão, quando o interesse está na previsão, é

separar os dados em dois conjuntos. O primeiro conjunto, chamado conjunto de treino, é usado para construir o modelo, e o segundo conjunto, o conjunto de teste, é usado para testar o modelo e avaliar o seu poder preditivo. A forma usual consiste em usar 80% dos dados para o conjunto de treino e 20% para o conjunto de teste (Breiman et al., 1984).

Árvores de regressão para dados em painel

É possível ajustar uma árvore de regressão a dados longitudinais ignorando a estrutura longitudinal dos dados, no entanto podem obter-se resultados enganadores na medida em que não está incluída a potencial relação dentro de cada observação, ao longo do tempo.

De acordo com Loh et al. (2013) várias tentativas foram feitas de modo a adaptar o algoritmo CART para dados longitudinais.

O primeiro algoritmo desenvolvido para árvores de regressão para dados longitudinais foi de Segal (1992).

Sela and Simonoff (2012) propuseram uma metodologia que combina a estrutura de modelos de efeitos mistos para dados longitudinais com a flexibilidade de métodos de estimação com base em árvores, designando o algoritmo por árvore RE-EM.

Árvores RE-EM são um método de *data mining* vocacionado para introduzir a estrutura de autocorrelação e de efeitos aleatórios nos dados longitudinais com variável resposta quantitativa.

O método proposto usa uma estrutura de árvore do tipo CART para estimar f (3.1), incorporando os efeitos aleatórios $\mathbf{b}_i$ específicos do indivíduo. Neste método, os nós podem ser divididos com base num qualquer atributo de maneira a que diferentes observações do mesmo objeto possam ser colocadas em diferentes nós. Para além disso, o método assegura que a estrutura longitudinal dos erros é preservada.

Uma vez que, nem os efeitos fixos nem os efeitos aleatórios são conhecidos, alterna-se entre modelar f (3.1) usando a metodologia de árvore de regressão, assumindo que as estimativas dos efeitos aleatórios estão corretas, e estimar os efeitos aleatórios usando REML dos modelos de efeitos mistos, assumindo que o modelo em árvore de regressão para f (3.1) está correto. A alternância na estimação dos parâmetros fixos e aleatórios em dois passos, justifica a designação Random Effects/EM tree, ou RE-EM tree, deste algoritmo. No entanto, não se utiliza o algoritmo EM (expectation-maximization), pelo que as propriedades usuais deste algoritmo não podem ser aplicadas (Sela and Simonoff, 2012).

Medidas de precisão

Num modelo de previsão é fundamental avaliar a qualidade das previsões. Uma vez que a metodologia utilizada neste trabalho usa um conjunto de treino para a modelação e um conjunto de teste para a avaliação das previsões, as medidas utilizadas para avaliar o desempenho dos modelos são

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

O MAE, Mean Absolute Error, é a distância média absoluta entre os valores observados (atuais) e os valores previstos.

O $RMSE$, Root Mean Square Error, corresponde à raiz quadrada da distância média ao quadrado entre os valores observados e os valores previstos.

A principal diferença entre estas duas medidas é que o $RMSE$ dá um peso maior aos erros de maior magnitude, enquanto que o MAE, sendo um parâmetro linear, as diferenças individuais têm todas o mesmo peso na média.

Neste trabalho usaram-se essencialmente dois *packages* e dois métodos – *package* `rpart` e `REEMtree` com os métodos correspondentes.

3.2 Dados

Para a realização deste projeto acederam-se a quatro bases de dados com o propósito de criar uma variável dependente explicada por três variáveis independentes.

Os dados tratados estão agregados por NUTS, Nomenclatura das Unidades Territoriais para Fins Estatísticos. Esta designação foi adotada pelo Eurostat (Gabinete de Estatísticas da União Europeia) para facilitar o desenvolvimento de estatísticas regionais.

Existem três níveis nesta nomenclatura, NUTS I, NUTS II e NUTS III. Mais especificamente, NUTS I diz respeito ao território do continente e às Regiões Autónomas dos Açores e da Madeira. NUTS II é constituído por sete regiões, cinco no continente e as Regiões Autónomas dos Açores e da Madeira. Por último, NUTS III engloba vinte e cinco unidades, designadas de sub-regiões.

Em 2015 entrou em vigor uma nova divisão regional em Portugal, NUTS 2013. Os dados trabalhados encontram-se agregados de acordo com esta última atualização, dentro

das NUTS III.

Os dados dos levantamentos nacionais em caixas de multibanco por localização geográfica (NUTS III) e os dados das compras através de terminais de pagamento automático por localização geográfica (NUTS III) foram fornecidos pela SIBS (empresa responsável pela gestão das Redes ATM Express e Multibanco). Estes dados, constantes na página do INE, encontram-se agregados por mês e o seu valor é dado em euros.

Com base no inquérito à permanência de hóspedes na hotelaria e outros alojamentos levado a cabo pelo INE, conseguiram-se os dados das dormidas nos estabelecimentos hoteleiros por localização geográfica (NUTS III) assim como o tipo de estabelecimento hoteleiro. É apresentado o número total de dormidas em estabelecimentos hoteleiros, que englobam, hotéis, pensões, estalagens, pousadas, motéis, hotéis-apartamentos e apartamentos turísticos. Estes dados encontram-se agregados por ano.

Os dados que dão origem à variável resposta correspondem ao produto interno bruto calculado a preços correntes. Estes dados estão agregados por ano.

Uma vez que os dados do produto interno bruto e das dormidas estão agregados por ano e os dados dos levantamentos e das compras se encontram agregados por mês, estes dois últimos foram transformados de maneira a estarem de acordo com os primeiros, isto é, por ano.

4

Resultados

4.1 Análise Exploratória

Os dados analisados são dados em painel, longitudinais. Existem várias observações para um mesmo indivíduo, região, ao longo do tempo. As 25 sub-regiões, NUTS III, foram medidas durante 7 anos, de 2011 a 2017, com periodicidade anual. Existe um total de 175 observações.

Assume-se independência entre as regiões e uma eventual correlação das observações dentro de cada região.

O estudo é balanceado uma vez que as regiões NUTS III foram medidas nos mesmos instantes de tempo e não existem dados em falta.

Em todo este capítulo *pib*, *levant*, *compras*, *dorm*, designam, respetivamente, os valores, em euros, de PIB, levantamentos nacionais em caixas de multibanco, compras em terminais de pagamento automático e o número total de dormidas em estabelecimentos hoteleiros.

Ao longo do texto usa-se PIB, levantamentos, compras e dormidas para designar estas variáveis.

Começa-se por apresentar algumas estatísticas sumárias dos dados (Tabela 4.1).

Verifica-se que a variável *dorm* apresenta uma escala de valores muito menor comparativamente às outras variáveis, que, por sua vez, têm uma escala de valores semelhante.

As transformações dos dados visam compatibilizar variáveis com escalas e dispersões muito diferentes. Neste caso, optou-se por aplicar uma transformação logarítmica a todas as variáveis.

Na Tabela 4.2 apresentam-se estatísticas sumárias das variáveis depois de aplicado

Tabela 4.1: Medidas de localização e de dispersão das variáveis

Variável	**Mínimo**	**1ºQuartil**	**Mediana**	**Média**	**3ºQuartil**	**Máximo**	**Desvio Padrão**
pib	$1,005 \times 10^9$	$2,379 \times 10^9$	$3,610 \times 10^9$	$7,131 \times 10^9$	$5,480 \times 10^9$	$6,998 \times 10^{10}$	12930606078
levant	$1,578 \times 10^8$	$3,482 \times 10^8$	$5,460 \times 10^8$	$1,031 \times 10^9$	$8,914 \times 10^8$	$8,157 \times 10^9$	1631975118
compras	$9,482 \times 10^7$	$3,170 \times 10^8$	$6,000 \times 10^8$	$1,298 \times 10^9$	$9,468 \times 10^8$	$1,465 \times 10^{10}$	2484145412
dorm	90046	246698	461979	1993988	926838	20207151	4074193

o logaritmo. Observa-se, agora, menos heterogeneidade e mais semelhança na dispersão dos dados.

Tabela 4.2: Estatísticas sumárias das 4 variáveis após transformação logarítmica

Variável	**Mínimo**	**1ºQuartil**	**Mediana**	**Média**	**3ºQuartil**	**Máximo**	**Desvio Padrão**
log(*pib*)	20,730	21,590	22,010	22,100	22,420	24,970	0,886
log(*levant*)	18,880	19,670	20,120	20,220	20,610	22,820	0,878
log(*compras*)	18,370	19,570	20,210	20,250	20,670	23,410	1,039
log(*dorm*)	11,410	12,410	13,040	13,330	13,740	16,820	1,348

Para além disso, calcularam-se as médias dos valores de cada variável em cada ano do estudo. Os resultados estão disponíveis na Tabela 4.3.

Tabela 4.3: Média das variáveis por ano

	2011	2012	2013	2014	2015	2016	2017
pib	7039604000	6730784000	6806024000	6921496000	7186324000	7453284000	7778448000
levant	1032174507	1007028027	1011215591	1016067000	1026356642	1049996279	1070848233
compras	1198731631	1148376260	1155349193	1224741491	1327068492	1440821285	1593863328
dorm	1577613	1587242	1741326	1948455	2122967	2364906	2615408

O ano de 2017 apresenta valores ligeiramente mais altos para todas as variáveis em análise. Ademais, observa-se que para as primeiras três variáveis ocorre inicialmente um decréscimo dos valores, entre o ano 2011 e o ano 2012, seguido de um aumento das médias do ano 2013 até ao ano 2017. Verifica-se sempre um aumento do número de dormidas ao longo dos anos. Este aumento é ligeiramente superior entre 2016 e 2017 e inferior entre 2011 e 2012.

É muito interessante visualizar a distribuição empírica dos valores do PIB usando as

25 regiões, ver Figura 4.1, na medida em que é clara a enorme assimetria induzida nesta distribuição pelas regiões Área Metropolitana de Lisboa e Área Metropolitana do Porto.

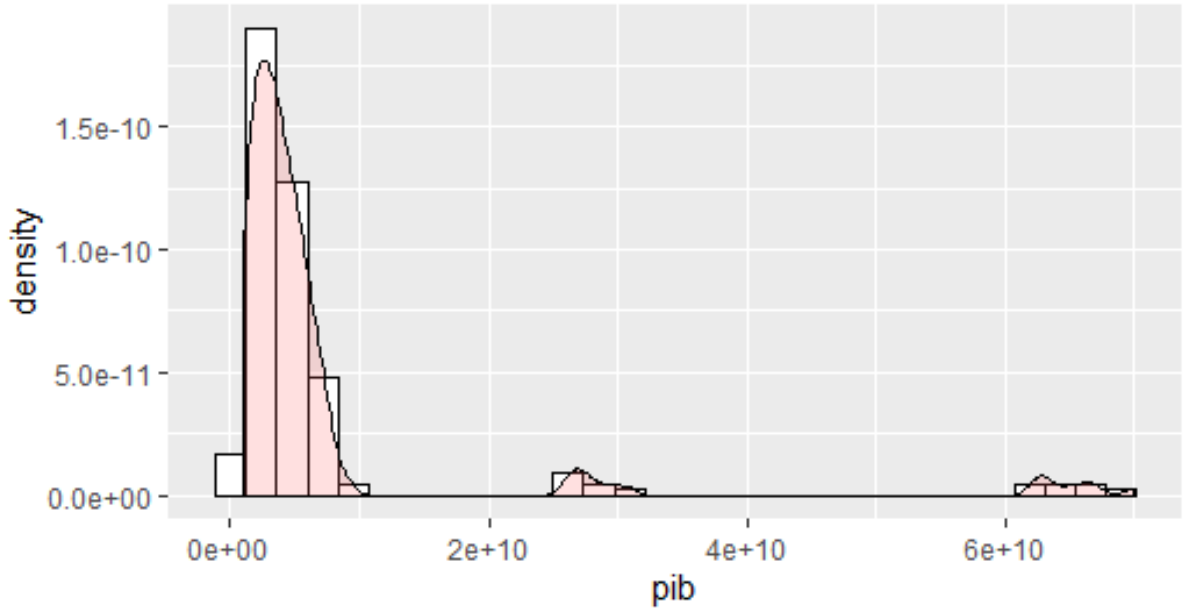


Figura 4.1: Distribuição empírica do PIB

Considerando a mesma distribuição sem estas duas Áreas Metropolitanas, o resultado, ver Figura 4.2, apresenta ainda uma assimetria positiva acentuada mas bastante mais ligeira.

Como os dados utilizados se encontram transformados pelas razões já enumeradas, apresenta-se a Figura 4.3, em que é visível uma uniformização dos valores sem, contudo, se perder o efeito de assimetria promovido pelas regiões de Lisboa e Porto.

A distribuição dos dados apresentados da variável resposta apresenta algumas especificidades, tais como valores baixos do PIB muito frequentes seguidos de valores mais elevados menos frequentes, o que pode ser indicador de uma mistura de distribuições, Figura 4.2. Este comportamento poderia ser sujeito a uma análise mais fina, mas porque não é relevante para o estudo que se faz neste trabalho não foi efetuada.

Na Figura 4.4 estão representadas as caixas-com-bigodes para a distribuição do PIB em cada ano. Como se pode constatar, a variação destes valores ao longo do tempo é praticamente inexistente. No ano de 2017 o valor do PIB é ligeiramente superior e apresenta uma maior dispersão em linha com os resultados da Tabela 4.3.

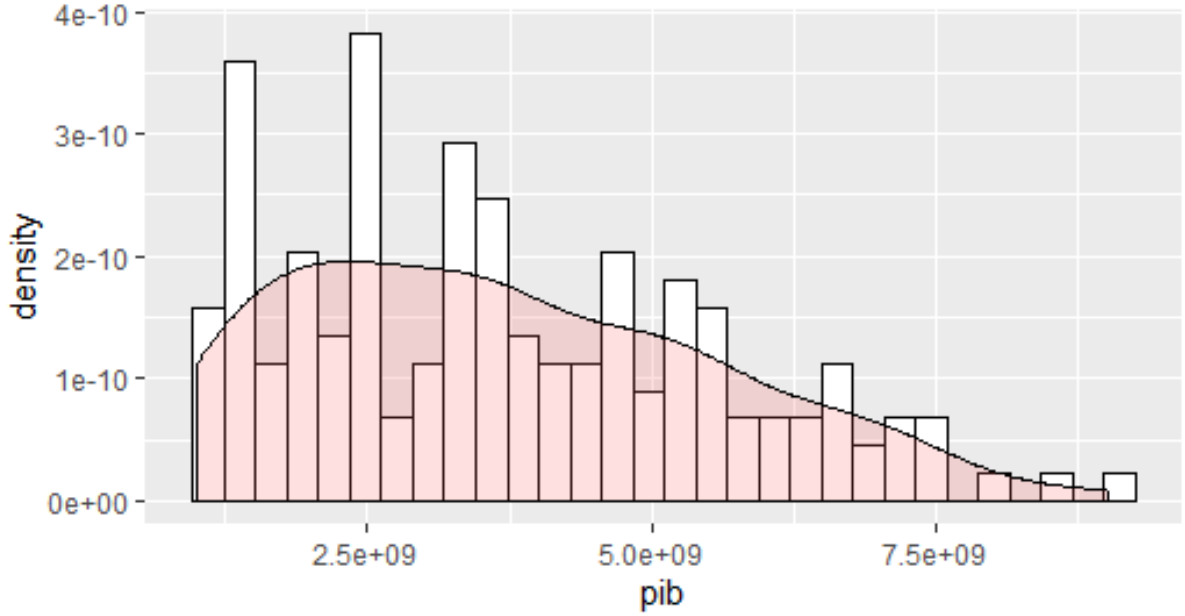


Figura 4.2: Distribuição empírica do PIB em 23 regiões

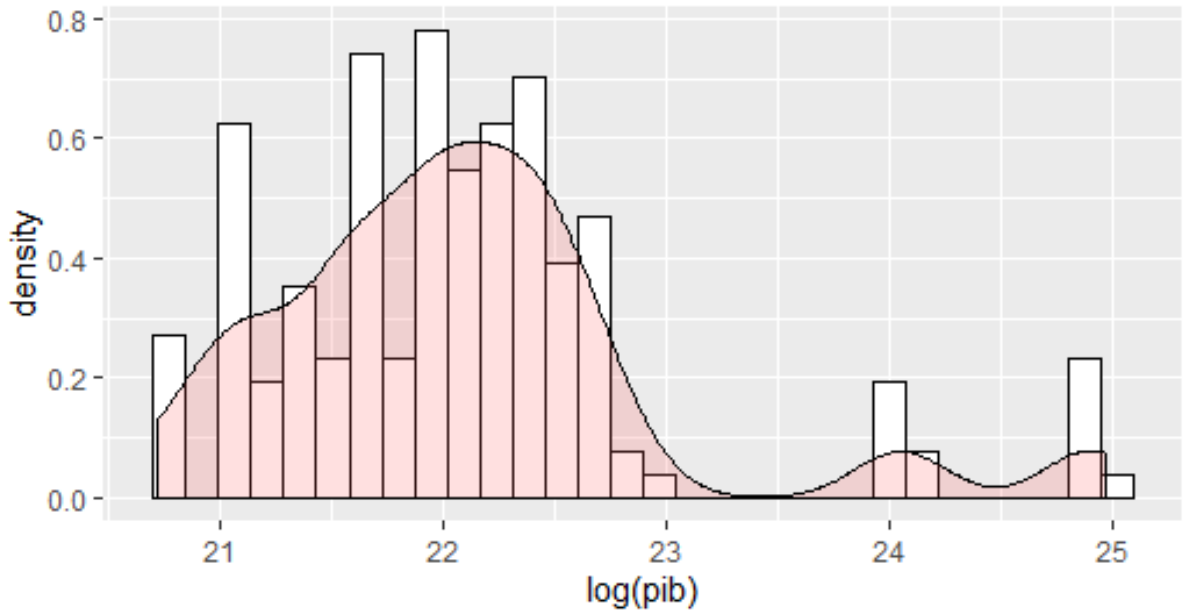


Figura 4.3: Distribuição empírica do log(PIB) em 25 regiões

Relativamente às restantes variáveis, conforme Figura 4.5, compras e dormidas apresentam um crescimento ao, longo do tempo, mais acentuado do que levantamentos, cujos valores mantêm-se idênticos ao longo do tempo.

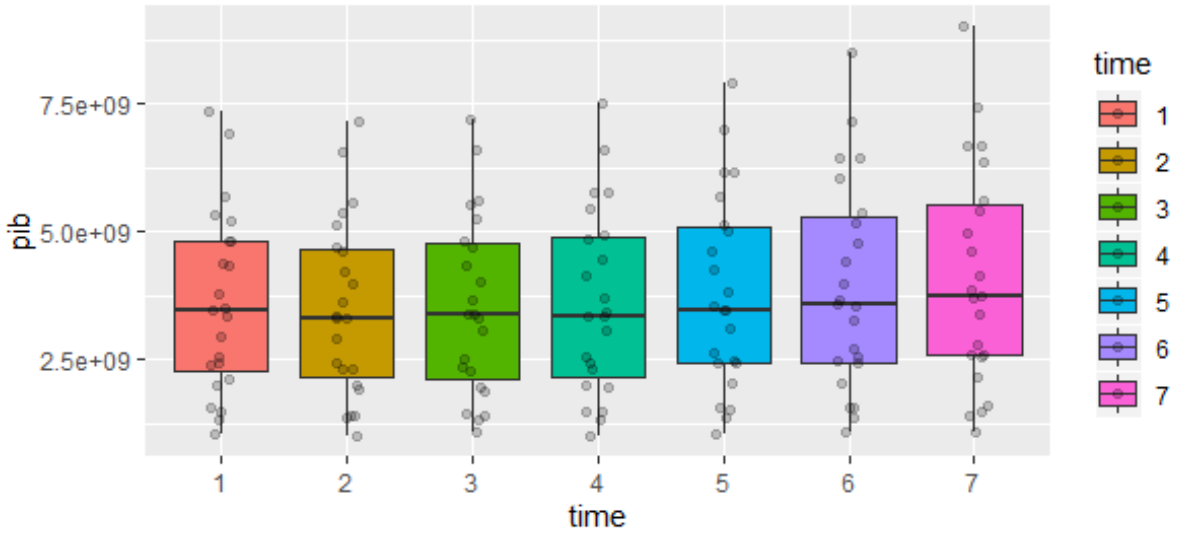


Figura 4.4: Distribuição do PIB em cada ano

Estes comportamentos são visíveis nos gráficos da Figura 4.5, onde se encontram ilustradas as médias dos valores de cada variável com intervalos de confiança de 95%.

Considere-se a distribuição dos valores do produto interno bruto, dos levantamentos e das compras em terminais, Figuras A.1 e A.2, respetivamente, dos Anexos, de 2011 até 2017 por sub-região (NUTS III), exceto Área Metropolitana de Lisboa e Área Metropolitana do Porto, conforme Figura 4.6.

A Área Metropolitana de Lisboa e Área Metropolitana do Porto apresentam valores muito mais elevados para todas as variáveis, não figurando na Figura 4.6 por questões de legibilidade (ver Figura A.3 nos Anexos).

Conforme se pode ver na Figura 4.6, a região do Algarve destaca-se das restantes 22, não considerando Lisboa e Porto. Por outro lado, com os valores mais baixos de PIB, (levantamentos e compras) encontram-se as sub-regiões de Alto Tâmega, Beira Baixa, Alto Alentejo e Terras de Trás-os-Montes.

Na Figura 4.6 destaca-se um PIB mediano bastante diferente entre algumas sub-regiões como, por exemplo, e a título representativo de classe, Alto Minho, Oeste, Cávado e Algarve, existindo ainda uma grande heterogeneidade na variabilidade dos dados conforme a região.

A variável correspondente às dormidas denuncia um cenário diferente, de assinalar, conforme Figura 4.7, onde o seu valor mais elevado encontra-se na sub-região do Algarve

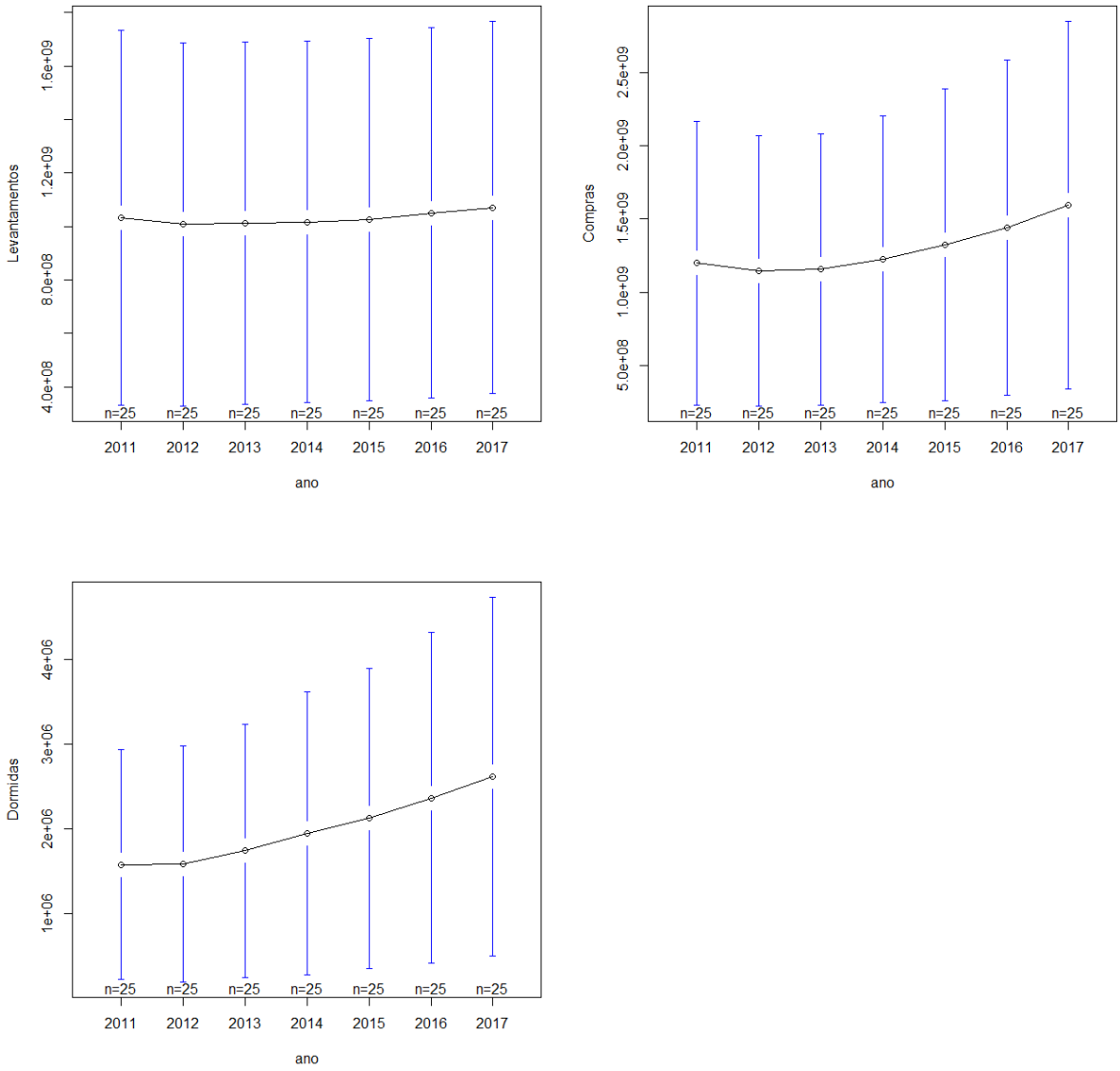


Figura 4.5: Variação ao longo dos anos

seguida da Área Metropolitana de Lisboa e da Região Autónoma da Madeira. Em relação aos valores mais baixos, destacam-se a Beira Baixa, Lezíria do Tejo e Terras de Trás-os-Montes.

Na Figura 4.4 a evolução dos valores do PIB ao longo do tempo é pouco significativa. Este efeito global verifica-se nas diferentes regiões mas com exceções, por exemplo, Algarve, conforme Figura 4.8.

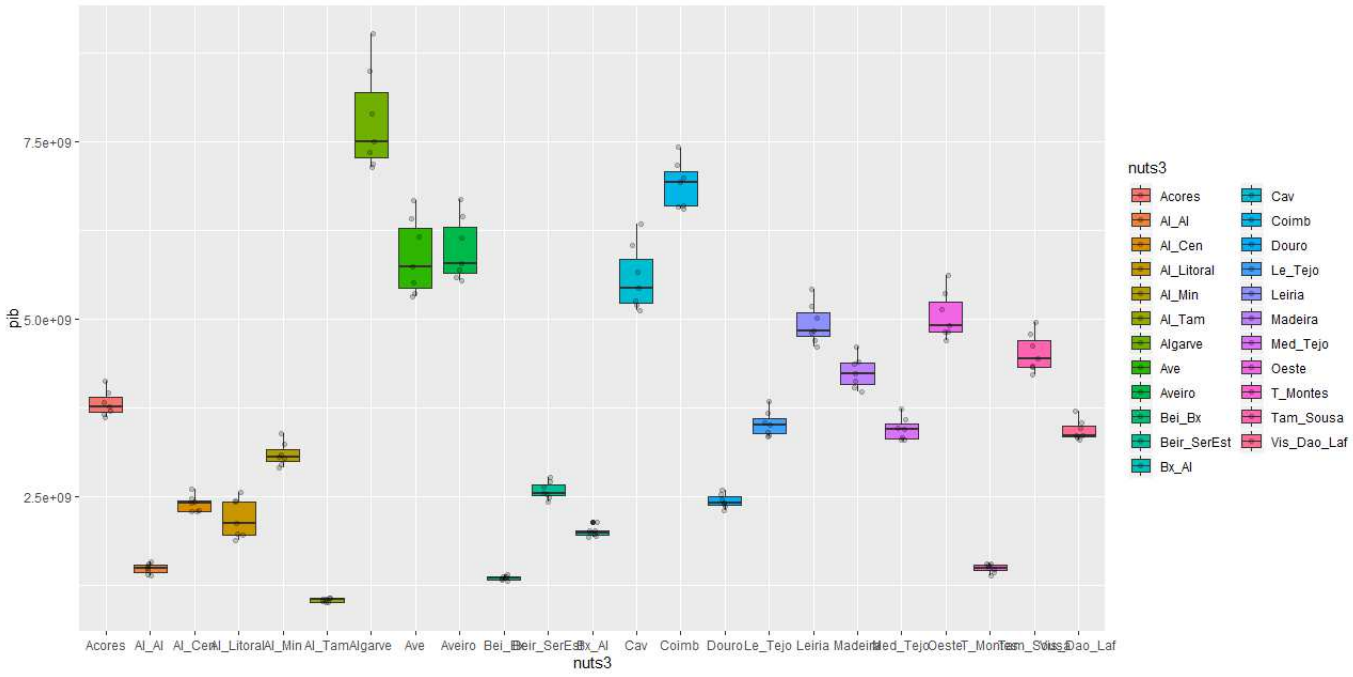


Figura 4.6: PIB por região

4.1.1 Correlações

É importante estudar as relações lineares entre as quatro variáveis de interesse.

Na Tabela 4.4 figura a matriz de correlação empírica das variáveis.

A variável PIB apresenta um coeficiente de correlação muito elevado com levantamentos e compras, estando estas também fortemente associadas positivamente. De facto, todas as correlações apresentadas são elevadas.

Tabela 4.4: Coeficientes de correlação de Pearson

	$\log(pib)$	$\log(levant)$	$\log(compras)$	$\log(dorm)$
$\log(pib)$	1,000	0,991	0,983	0,727
$\log(levant)$	0,991	1,000	0,977	0,711
$\log(compras)$	0,983	0,977	1,000	0,780
$\log(dorm)$	0,727	0,711	0,780	1,000

As Figuras 4.9, 4.10 e 4.11 revelam relações lineares, em termos globais, com comportamentos diferentes. No entanto, dentro de cada sub-região, percebe-se que as relações são de associação linear positiva com razões de crescimento diferentes conforme as variáveis em causa.

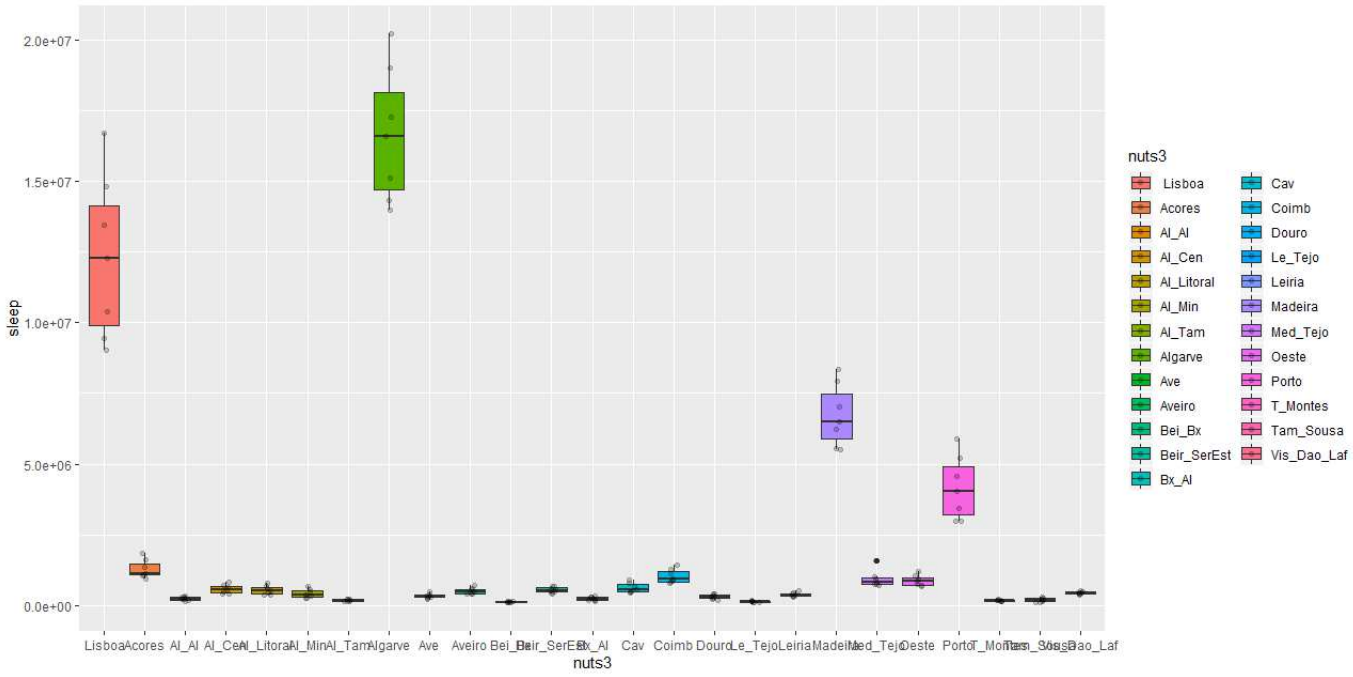


Figura 4.7: Dormidas por região

Na relação entre o PIB e os levantamentos, tem-se uma relação linear de base mas o comportamento observado não é linear dentro das regiões ainda que a correlação seja positiva.

Por exemplo, na região do Alto Tâmega, Beira Baixa, Alto Alentejo, Terras de Trás-os-Montes e Ave, as relações entre levantamentos e PIB não são lineares conforme gráfico 4.9.

Este efeito pode ser importante quando se opta por uma metodologia de modelos lineares mistos em detrimento de uma metodologia de árvores de regressão em painel como se verá no desenvolvimento deste trabalho.

A Figura 4.10 indica que a relação de fundo entre PIB e número de dormidas, sendo linear, não é tão evidente. Mesmo dentro de cada região é fácil dar exemplos de regiões em que a relação entre as duas variáveis é praticamente inexistente (Alto Tâmega) e outros em que a relação é quase perfeita (Algarve).

Por outro lado, a Figura 4.11 revela uma forte associação entre levantamentos e compras, o que vai fazer com que haja necessidade de incluir efeitos de interação nos modelos de regressão linear.

Estas conclusões com base na observação dos gráficos estão de acordo com os coefi-

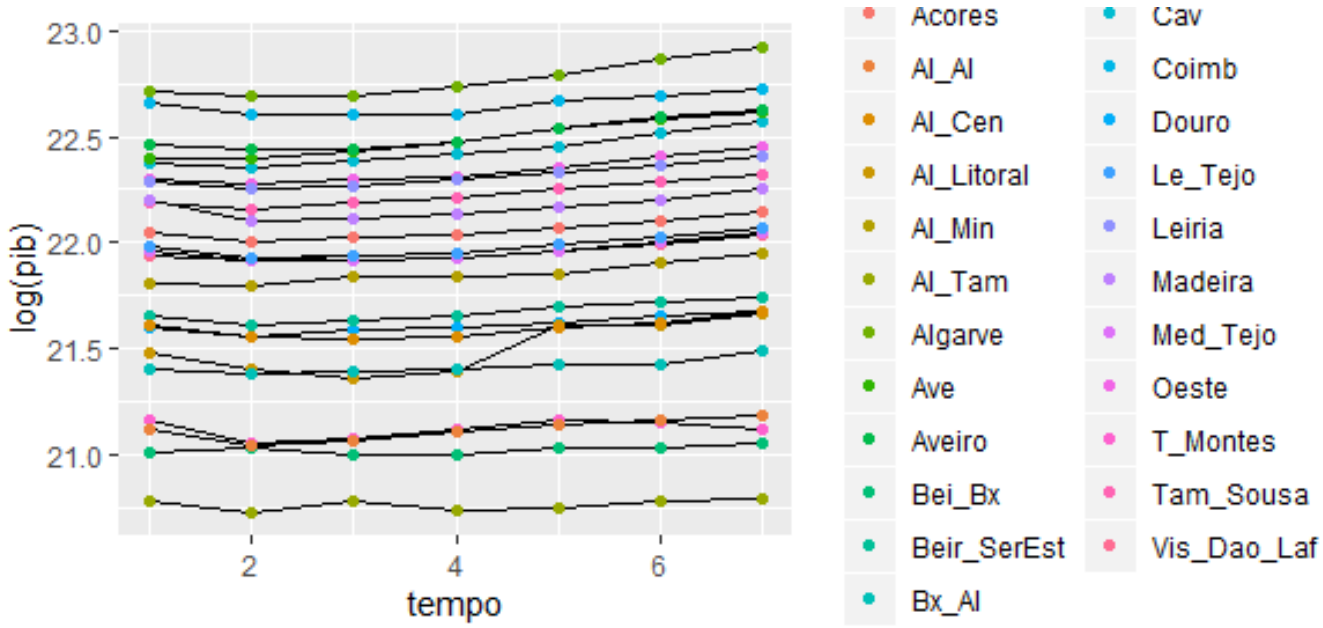


Figura 4.8: Evolução do PIB

cientes apresentados na Tabela 4.4.

Todos estes resultados são fundamentais na secção seguinte aquando do ajustamento de modelos.

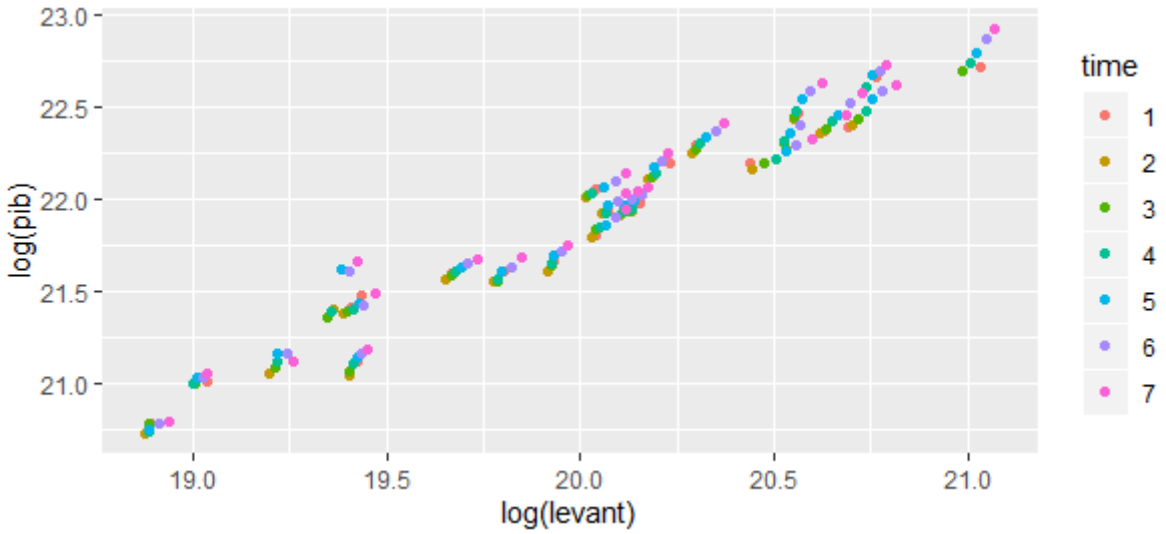


Figura 4.9: Relação entre PIB e levantamentos

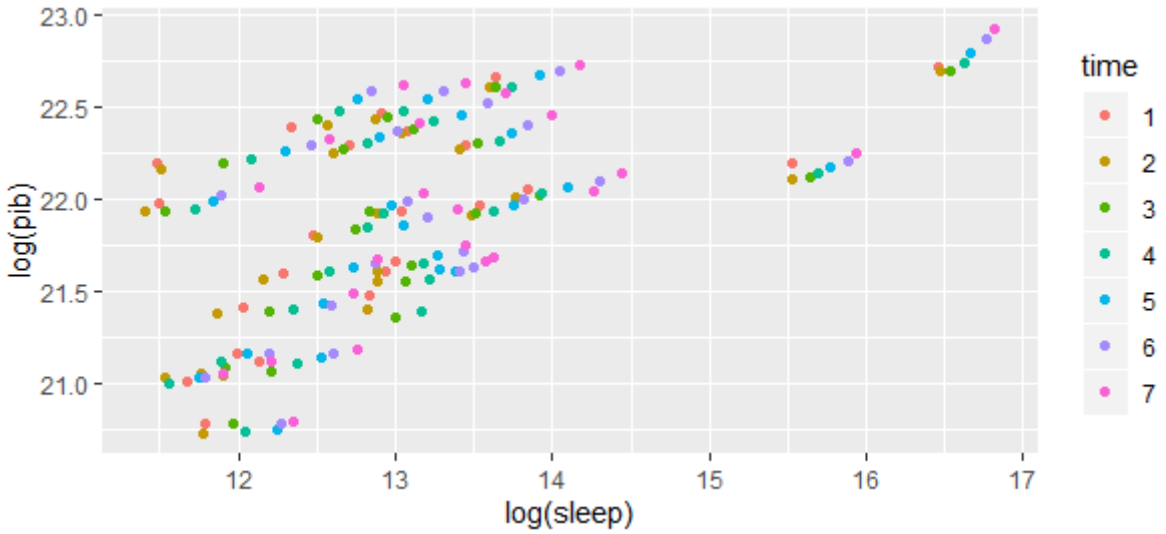


Figura 4.10: Relação entre PIB e dormidas

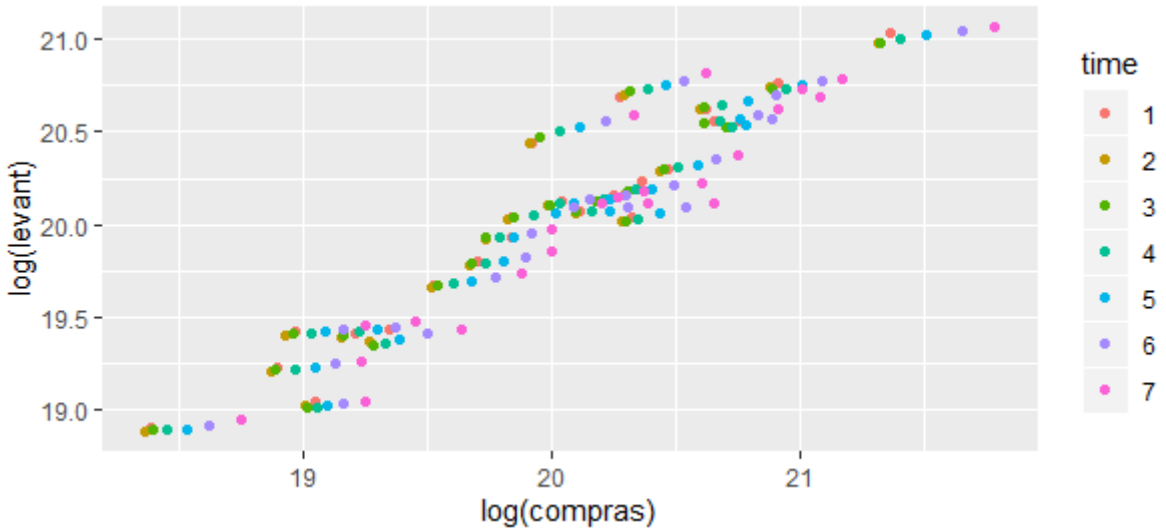


Figura 4.11: Relação entre levantamentos e compras

4.2 Modelos

Como já foi referido, o objetivo principal deste trabalho é prever o valor do PIB, variável resposta, a partir da observação de três variáveis preditoras, dormidas, levantamentos e compras, cujos valores influenciam os valores da variável resposta, conforme a análise exploratória dos dados efetuada anteriormente indica.

Trata-se de um problema clássico de regressão.

Em termos gerais, a regressão pode ter dois propósitos.

Exploratório ou Explicativo – obter uma relação matemática que indique, mas que não prova, uma relação de causa-efeito entre a variável dependente, resposta, e as variáveis independentes, variáveis explanatórias, ou explicativas.

Preditivo – obter uma relação que permita, em futuras observações das variáveis independentes, preditoras, predizer os valores correspondentes da resposta, sem ter que a medir.

Neste contexto é a função preditiva da regressão que interessa.

Neste capítulo, o processo de modelação é iniciado usando modelos lineares múltiplos com interação entre as variáveis explicativas e também com a variável categórica NUTS III, que é tratada como um efeito fixo.

De seguida, incorpora-se esta variável categórica num modelo de efeitos mistos, como fator aleatório, considerando ainda interações entre algumas variáveis independentes. Finalmente, uma vez que o objetivo é prever, usa-se modelos de árvores de regressão, para dados longitudinais, no processo de modelação.

4.2.1 Modelos Lineares

Os modelos lineares são o tipo de regressão mais frequentemente utilizado, que permitem explicar a mudança média na variável dependente, dada uma unidade de variação em cada uma das variáveis independentes, mantendo todas as outras fixas. Neste problema em concreto, não há razão para considerar outro tipo de regressão uma vez que a variável resposta é numérica, quantitativa.

A seleção de modelos foi efetuada com base em Estatísticas F, ANOVAs, AIC, AICc e BIC, apesar de se ter considerado sempre um conjunto de treino e não a totalidade dos dados.

A análise exploratória efetuada anteriormente indica que existem correlações entre as variáveis explicativas, o que leva a considerar condições de interação entre variáveis nos modelos de regressão.

No caso de variáveis numéricas, existindo interação das variáveis X_1 e X_2 com a variável resposta Y, o gráfico de Y em função de $X_1 * X_2$ deve ser linear. Apresenta-se, a título de exemplo, na Figura 4.12 a interação entre compras e levantamentos com o PIB.

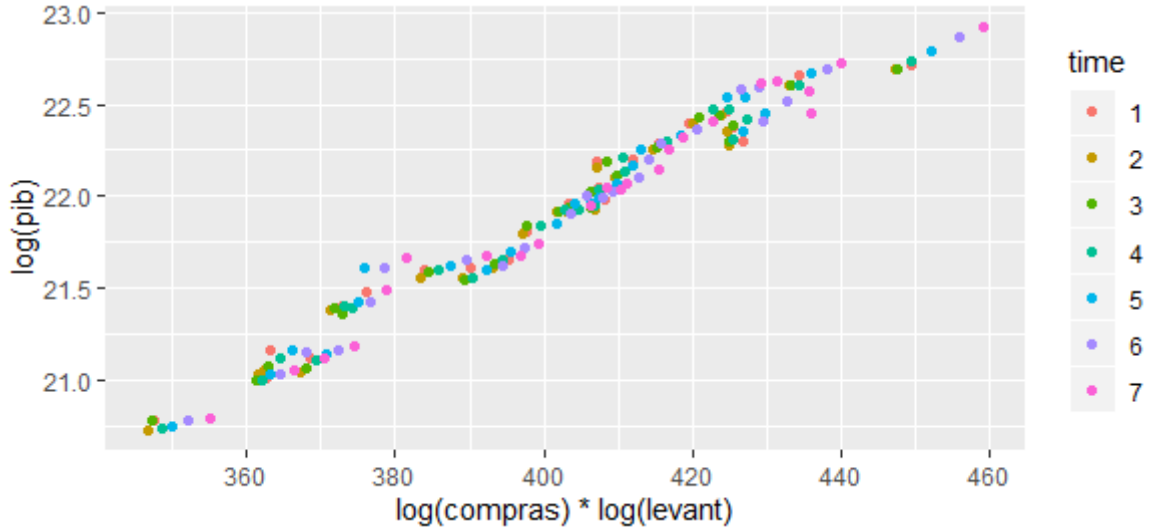


Figura 4.12: Interação entre levantamentos e compras

No caso em que uma das variáveis explicativas é categórica, digamos X_1, o gráfico de Y vs X_2 deve apresentar valores medianos com magnitudes e tendências diferentes conforme as categorias de X_1. Esse comportamento encontra-se ilustrado, por exemplo, na Figura 4.13 onde se apresenta a interação de levantamentos com as NUTS III nos valores do PIB.

Regressão Linear Múltipla – Modelo de efeitos fixos

Nas secções que se seguem apresentam-se modelos com parâmetros estimados. A notação adotada não inclui o símbolo usual, $\widehat{\ }$, por questões estéticas.

Começa-se por apresentar um modelo em que as NUTS III são consideradas efeitos fixos.

O modelo de regressão selecionado para explicar os valores da variável resposta, $\log(pib)$, em função das restantes variáveis quantitativas e da variável categórica $nuts3$, se-

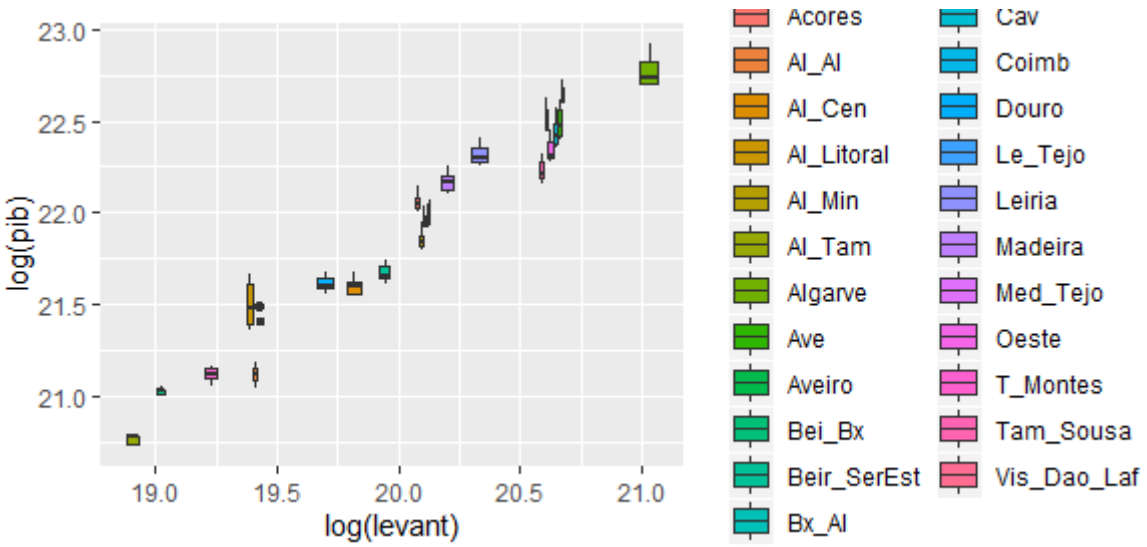


Figura 4.13: Interação entre levantamentos e NUTS III

lecionou como variáveis significativas $\log(levant)$ além das interações $\log(levant) : \log(compras)$ e $\log(levant) : nuts3$.

A fórmula do modelo (com os dados de treino) é a seguinte:

$$\log(pib) = 6,738 + 0,516\log(levant) + 0,012\log(compras) \times \log(levant) - 0,008\log(levant) \times Porto$$

Este modelo apresenta um $R^2 = 0,999$ (ajustado), um $AIC = -606,226$. O valor para o $RMSE = 0,025$ (para efeitos preditivos).

O comportamentos dos resíduos é apresentado na Figura 4.14.

Modelo de Efeitos Mistos

A existência de dados longitudinais implica que o modelo inclua as eventuais correlações dentro de cada indivíduo ou região, além da heterogeneidade entre os indivíduos.

O modelo de efeitos mistos selecionado, com fatores aleatórios NUTS III, para incluir o efeito da heterogeneidade entre regiões, é mais simples do que o anterior e inclui compras em vez de levantamentos. De facto, esta alteração tem a ver com a elevada interação que existe entre compras e levantamentos, já mencionada anteriormente e não se estaria à espera que o modelo incluísse as duas variáveis.

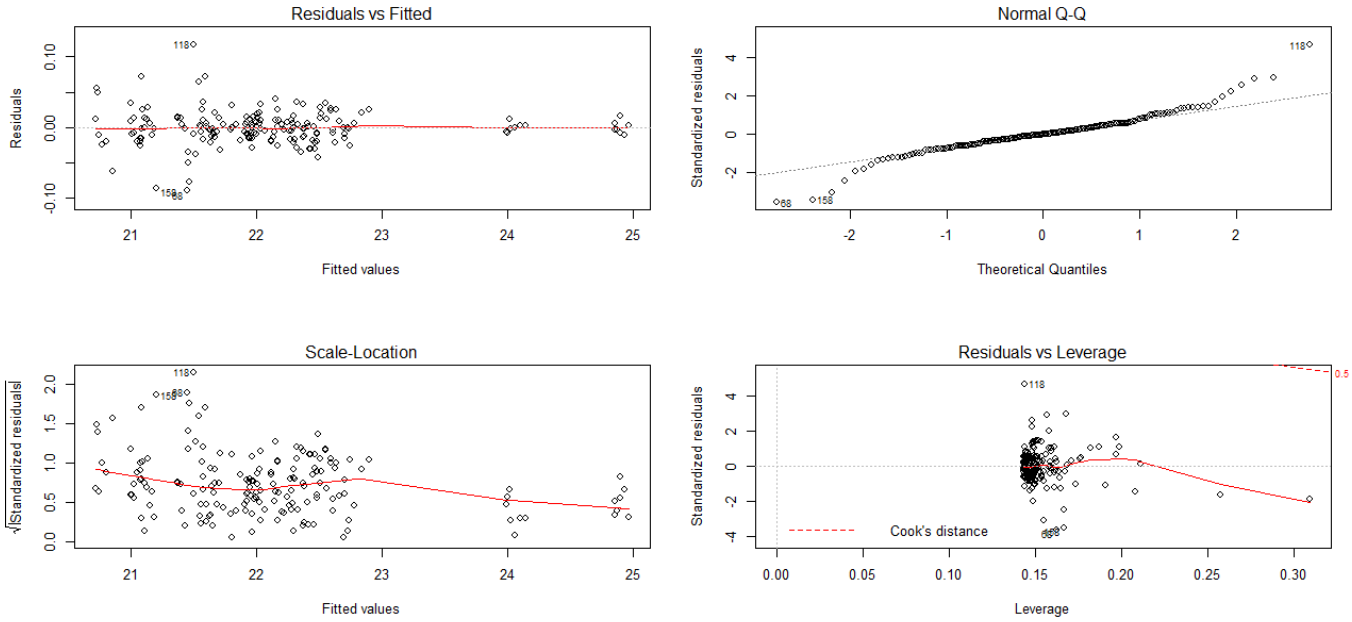


Figura 4.14: Comportamento dos resíduos no modelo de efeitos fixos

$$\log(pib) = 16,746 - 0,398\log(compras) + 0,033\log(compras) \times \log(levant)$$

Em termos da variabilidade total dos dados explicada pela introdução dos efeitos aleatórios, o modelo associa um desvio-padrão de $0,089$ a estes efeitos, passando o desvio-padrão da parte residual apenas para $0,028$, o que é indicador que o fator aleatório é importante no modelo, explicando grande parte da variabilidade residual.

Quanto ao poder preditivo, obtém-se $RMSE = 0,023$, um $AIC = -502,689$ e um $R^2 = 0,999$ (ajustado). Note-se que este valor não pode ser comparado com o anterior, uma vez que o método utilizado para a estimação dos parâmetros, REML, pressupõe que a parte fixa se mantenha para comparação.

Em termos gráficos a precisão nas previsões deste modelo pode ser visualizada na Figura 4.15.

Análise dos Resíduos

Os modelos apresentados estão de acordo com os pressupostos do modelo de regressão linear. Os resíduos distribuem-se aleatoriamente em torno de zero tanto no modelo

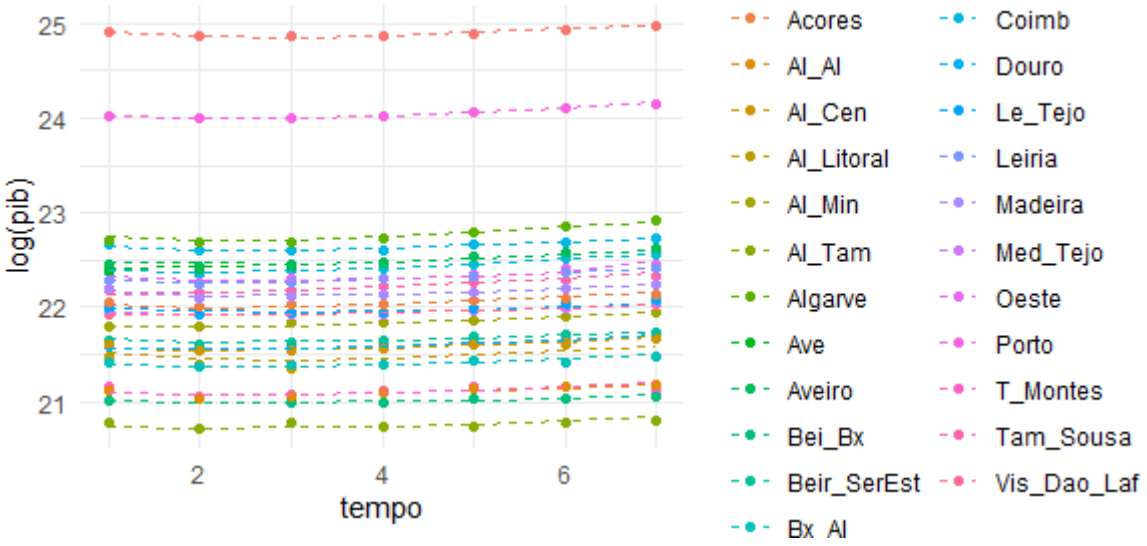


Figura 4.15: Precisão das previsões

global como em relação a cada variável, ver Figura 4.16. Para além disso, para a validação da significância estatística dos parâmetros, os resíduos devem ter um comportamento Normal, ver Figura 4.17.

Árvores de Regressão

Como já foi visto, os algoritmos de árvores de classificação e regressão, CART, consistem num conjunto de condições do tipo "se-então"que permitem prever ou classificar casos. Este tipo de algoritmos pode ser usado para modelar a função f (3.1) que relaciona a variável resposta com os preditores. No caso de a variável resposta ser contínua os algoritmos dizem-se de regressão.

A função obtida para f (3.1) é uma função descontínua em tantos pontos quantas as folhas da árvore, que não inclui parâmetros, pelo que este modelo é não paramétrico.

As árvores de regressão são obtidas por partição sucessiva do espaço preditor em subconjuntos nos quais a distribuição da variável resposta é cada vez mais homogénea. A homogeneidade é medida em termos de "impureza"dos nós. As medidas de impureza, no caso de árvores de regressão, incluem o cálculo da variância do subconjunto de dados resposta em cada nó.

Os algoritmos de machine learning implementados no R têm como base a busca "gulosa"em que há necessidade de recorrer a poda e validação cruzada, sob pena do processo de busca só terminar quando as folhas forem puras o que, no limite, implica que cada

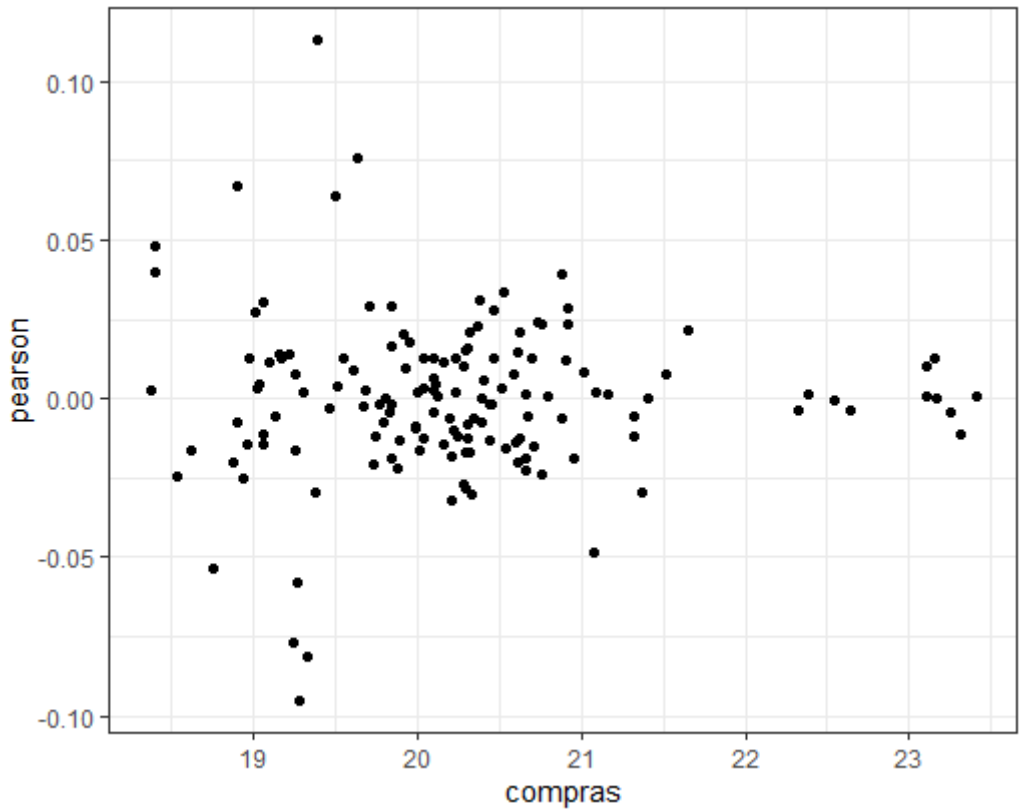


Figura 4.16: Resíduos de Pearson relativos a compras

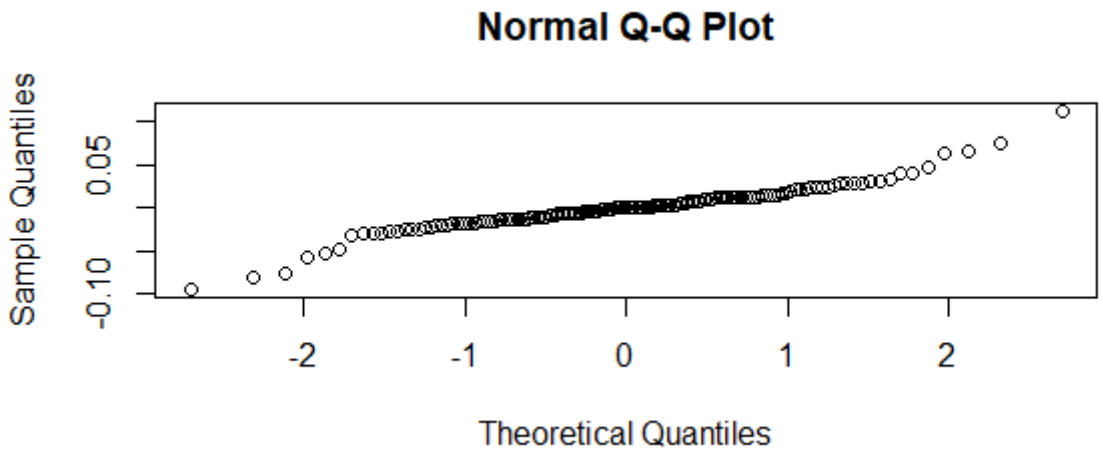


Figura 4.17: Normalidade dos resíduos do modelo de efeitos mistos

folha tenha apenas uma observação.

Em primeiro lugar apresenta-se uma árvore sem a inclusão das NUTS III como efeitos aleatórios, mas como uma variável categórica – tal como foi feito nos modelos lineares múltiplos de efeitos fixos acima descritos.

De realçar que neste tipo de abordagem, todas as variáveis de interesse são incluídas no modelo, e o algoritmo escolhe, com critérios próprios, que variável vai usar em cada divisão. Não é necessário fazer uma seleção de variáveis nem tão pouco haver a preocupação de incluir interações entre variáveis, porque tal não faz qualquer sentido.

Utiliza-se a função `rpart` do package `rpart`, e o modelo utilizado inclui $\log(pib)$ em função de $\log(levant)$, $\log(dorm)$, $\log(compras)$ e ainda $nuts3$. O resultado encontra-se na Figura 4.18.

Da análise da árvore resulta que as únicas variáveis selecionadas foram levantamentos e NUTS III. De acordo com este esquema obtém-se cinco classes de previsões correspondendo a cinco folhas na árvore. O nó raiz contém 100% das observações (147 observações) das quais 136 apresentam um valor de $\log(levant)$ inferior a 21,627.

Por exemplo, se $\log(levant)$ é inferior a 19,990 e se as NUTS III forem Alto Alentejo, Alto Tâmega, Beira Baixa, Terras de Trás-os-Montes, o valor previsto para o PIB é $exp(21,007)$.

O valor preditivo deste modelo medido pelo $RMSE$ é 0,233, bastante superior a qualquer modelo dos apresentados anteriormente.

Árvores de Regressão em Painel

As árvores de regressão do tipo CART com inclusão de efeitos aleatórios para modelação de dados em painel, designadas aqui por árvores RE-EM, permitem, tal como nos modelos mistos, incorporar no modelo final a eventual correlação dentro dos indivíduos.

Como se viu anteriormente, a parte de efeitos fixos do modelo é ajustada usando um algoritmo do tipo CART, e só depois são modelados os efeitos aleatórios conforme descrito no método RE-EM.

Usando o package `REEMtree` e o método com o mesmo nome obtém-se a árvore na Figura 4.19, onde as variáveis selecionadas são agora compras, levantamentos e tempo. O tempo surge pela primeira vez como uma variável importante. Como se viu anteriormente, nos modelos de regressão clássicos não foi incorporado o efeito tempo, o que não surpreende porque, tal como foi visto na análise exploratória dos dados, a evolução dos valores do PIB com o tempo, e em termos globais, não se mostra muito significativa (ver Figura 4.4)

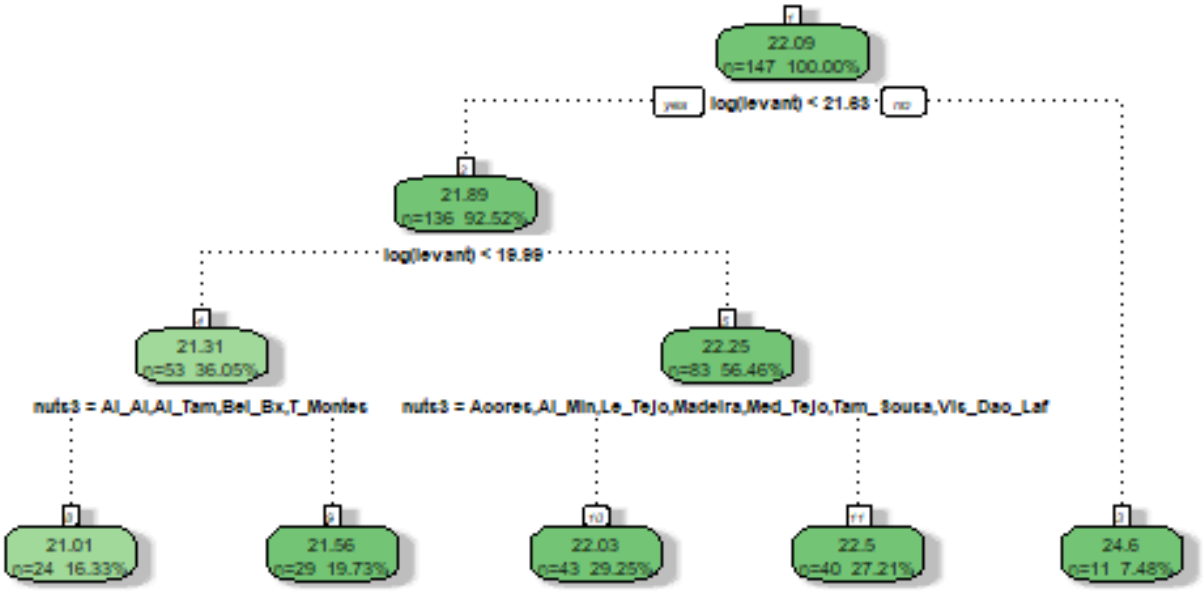


Figura 4.18: Árvore de regressão

apesar de, dentro de algumas regiões, o tempo poder ser importante, ver Figura 4.8 – o modelo RE-EM incorpora este efeito.

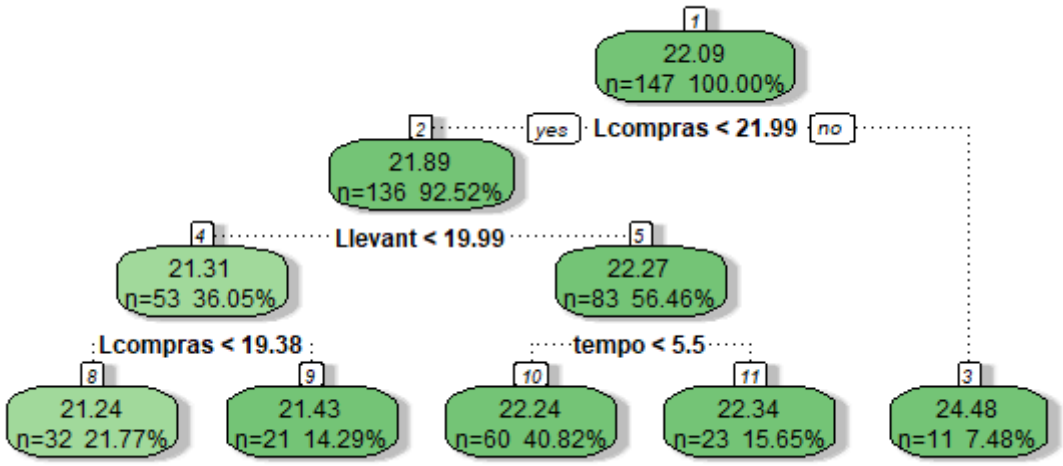


Figura 4.19: Árvore REEM

Quanto ao poder preditivo deste modelo é inferior aos modelos lineares de regressão,

apresentando um valor de $RMSE = 0,043$, o que é compatível com os estudos efetuados na literatura, uma vez que com estes dados, a relação entre as variáveis é obviamente linear.

A fim de se efetuar uma comparação entre os valores atuais (observados), constantes no conjunto teste, os valores previstos pelo modelo de efeitos mistos apresentado e o modelo em árvore com efeitos aleatórios, efetuou-se uma ANOVA clássica.

O valor de prova do teste de Bartlett de homogeneidade de variâncias, `p-value = 0.997` leva à não rejeição da hipótese de igualdade das variâncias, e sendo o valor `F=0.0003` com (2, 81) graus de liberdade, conclui-se que os três conjuntos de dados independentes são estatisticamente iguais.

Esta conclusão valida a utilização de qualquer um dos modelos para efeitos de previsão.

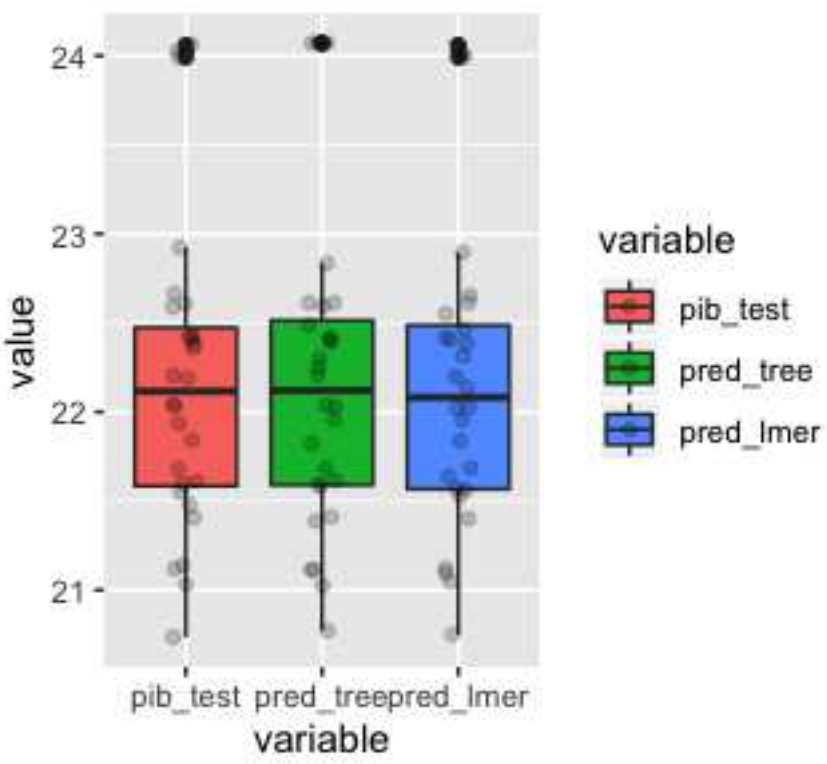


Figura 4.20: Valores de teste vs Valores previstos pelos modelos

5

Conclusões

Começa-se por enumerar as conclusões gerais do trabalho, passando depois às conclusões mais específicas.

1. Os modelos de regressão usam-se para prever e explicar relações. Estes modelos não devem ser usados para extrapolar para além do domínio dos dados.

2. Os modelos obtidos por aplicação dos algoritmos de *machine learning* de regressão em árvore focam-se na predição. Acresce que a medida de impureza usada nas folhas tem como objetivo principal aumentar o poder preditivo do modelo.

3. A medida utilizada para aferir a qualidade das previsões, RMSE, não é única mas é a mais usual dentro das medidas de precisão "out of sample".

Enfatiza-se a obtenção de modelos com boa qualidade de ajustamento, que explicam os valores do PIB em função de variáveis de economia partilhada tais como, compras, levantamentos e dormidas em estabelecimentos hoteleiros e similares.

A qualidade dos modelos de efeitos mistos obtida e a sua bondade de ajustamento são bastante satisfatórios, acrescendo ainda o facto de terem um elevado poder preditivo. Obtiveram-se valores de R^2 ajustado da ordem dos 99%, o que é manifestamente bom.

Este facto é por si muito importante, uma vez que abre a possibilidade de, em dados de elevada frequência, estas relações se continuarem a verificar, com uma pertinência e utilidade bastante maiores do que no presente estudo.

Pode-se ainda concluir que:

1. As relações entre os valores do PIB e o tempo, dentro de cada região, nem sempre são lineares. Este facto é tanto mais curioso quanto o facto das regiões com PIB mais baixo terem comportamentos bastante longe da linearidade ao contrario das regiões com valores de PIB mais elevados (ver Figura A.4 em Anexo).

2. Existe uma relação de base linear entre PIB e dormidas. No entanto, existem regiões de quase aleatoriedade entre estas duas variáveis tais como Alto Tâmega e Terras Trás os Montes, o que pode explicar o facto desta variável não ter sido incluída nos modelos mistos nem nas árvores.

3. As relações de linearidade entre as variáveis explicativas e a variável resposta incrementam a qualidade de ajustamento e previsão dos modelos de regressão mas condicionam a qualidade das previsões dos modelos de árvores de regressão em painel, o que está de acordo com vários estudos de simulação efetuados e constantes na literatura da área (Sela and Simonoff, 2012).

4. Apenas quatro, em vinte e três, das regiões consideradas apresentam uma não linearidade entre levantamentos e tempo, o que é refletido, por um lado, na não inclusão da variável tempo nos modelos mistos mas, por outro, a sua inclusão nos modelos de árvores de regressão, ver Figura 4.8.

5. As regiões de Lisboa, Alto Alentejo, Alentejo Litoral, Alto Tâmega, Aveiro, Coimbra, Médio Tejo, Terras de Trás-os-Montes e Viseu Dão Lafões, conforme Figura A.4 (Anexos), apresentam um comportamento não linear com o tempo, que pode ter tido influência na qualidade da previsão no modelo de regressão de árvores para dados longitudinais, uma vez que este inclui o tempo como variável importante.

Conclui-se que estas metodologias podem ser usadas em paralelo para aferir a qualidade de indicadores de economia partilhada ou colaborativa, e complementar estruturas de dados existentes, com vista a uma análise mais completa e rigorosa, uma vez que, quanto maior o conjunto de dados disponível, maior o conjunto de teste utilizado, e maior a confiança nas qualidades da previsão.

Em termos de trabalho futuro, uma vez que se pretende utilizar dados da economia colaborativa para fazer previsões acerca do PIB e de outros indicadores macroeconómicos, por ser um fenómeno em expansão e com efeito direto na economia dos países, a potencialidade que existe nos modelos em árvores de regressão abre a possibilidade de se atingir esse objetivo com qualidade e rigor.

De facto, este trabalho mostrou grande precisão na previsão mesmo com modelos de base linear, muito poucas observações e um número muito baixo de variáveis explicativas, o que deixa antever e potencia o seu uso em Big Data.

Bibliografia

Baker, D. (2015). The opportunities and risks of the sharing economy. *Testimony before the Subcommittee on Commerce, Manufacturing, and Trade of the US House of Representatives Committee on Energy and Commerce. Washington, DC, September*, 29.

Belk, R. (2014). Sharing versus pseudo-sharing in web 2.0. *The Anthropologist*, 18(1):7–23.

Breiman, L., Friedman, J. H., Oslhen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, Wadsworth.

Chatterjee, S. and Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Ferreira, M. d. F. M. (1999). Árvores de regressão e generalizações: Aplicações. *Tese de Mestrado, Universidade do Porto*.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Hajjem, A., Bellavance, F., and Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328.

INE, I. N. d. E. (2018). Como se calcula o pib.

Loh, W.-Y., Zheng, W., et al. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, 7(1):495–522.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434.

Nguyen, S. and Llosa, S. (2018). On the difficulty to define the sharing economy and collaborative consumption–literature review and proposing a different approach with the introduction of 'collaborative services'. *Journée de la Relation à la Marque dans un Monde Connecté, Centre de Rechercheen Gestion des Organisations, Nov 2018, Colmar, France*.

Oliveira, B. M. M. (2017). Mercados p2p e economia da partilha: Perfil e motivações de quem participa no consumo colaborativo. *Tese de Mestrado, Universidade do Porto*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Ramos, P. N. and Rodrigues, A. (2001). Porque é diferente o pib per capita das regiões portuguesas? *VIII Encontro da Associação Portuguesa para o Desenvolvimento Regional*.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media.

Schor, J. et al. (2016). Debating the sharing economy. *Journal of Self-Governance and Management Economics*, 4(3):7–22.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87(418):407–418.

Sela, R. J. and Simonoff, J. S. (2012). Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning*, 86(2):169–207.

Stanoevska-Slabeva, K., Lenz-Kesekamp, V., and Suter, V. (2017). Platforms and the sharing economy: An analysis. report for the eu horizon 2020 project ps2share: Participation, privacy, and power in the sharing economy.

A

Anexos

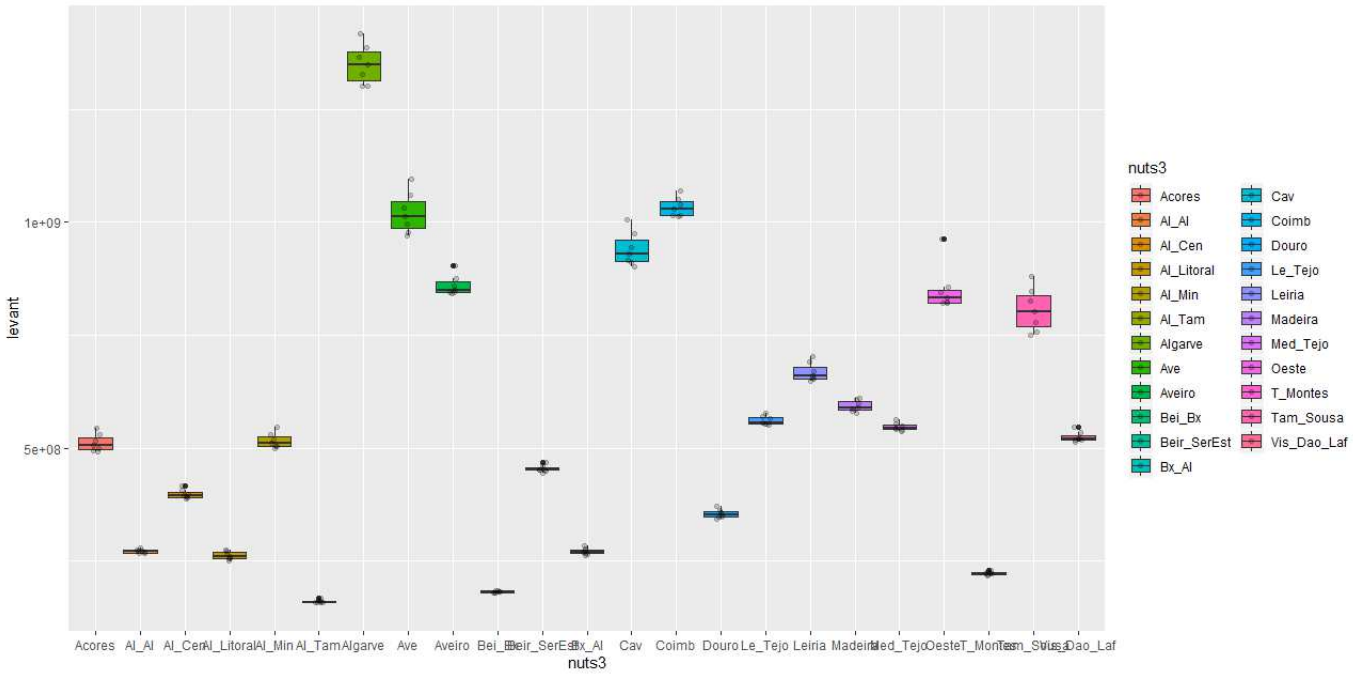


Figura A.1: Levantamentos por região

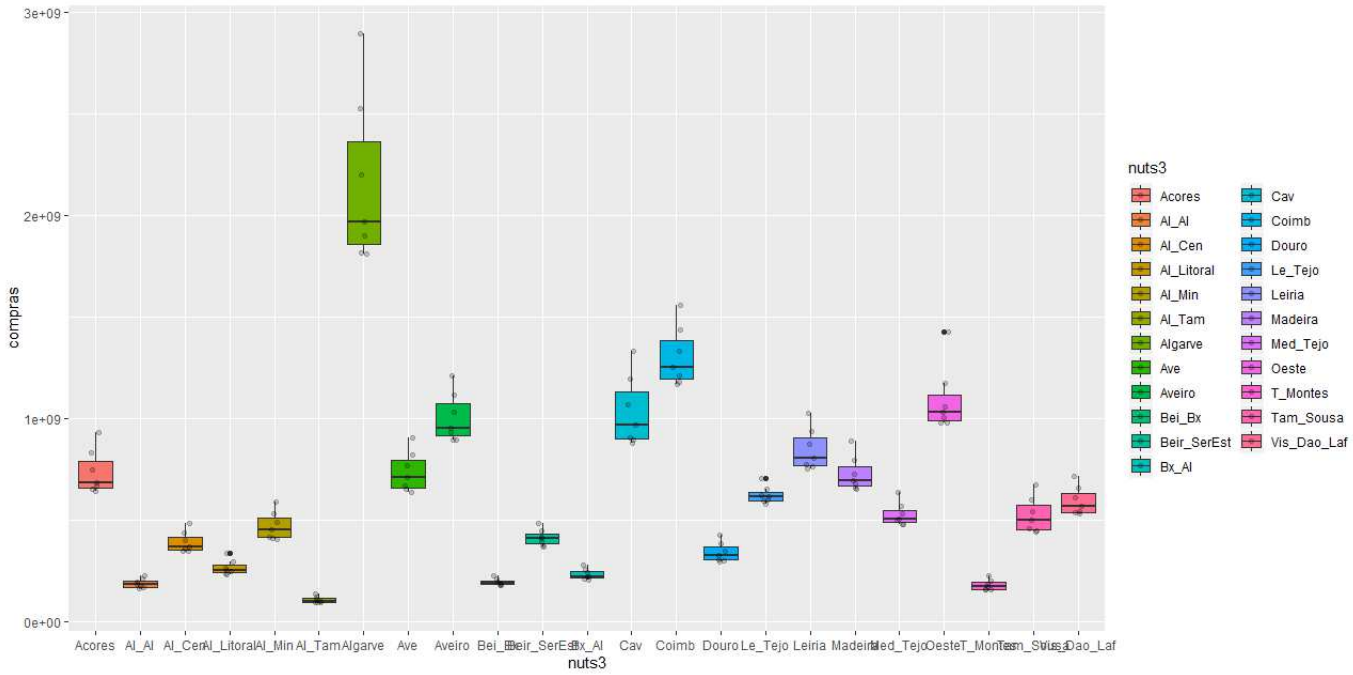


Figura A.2: compras por região

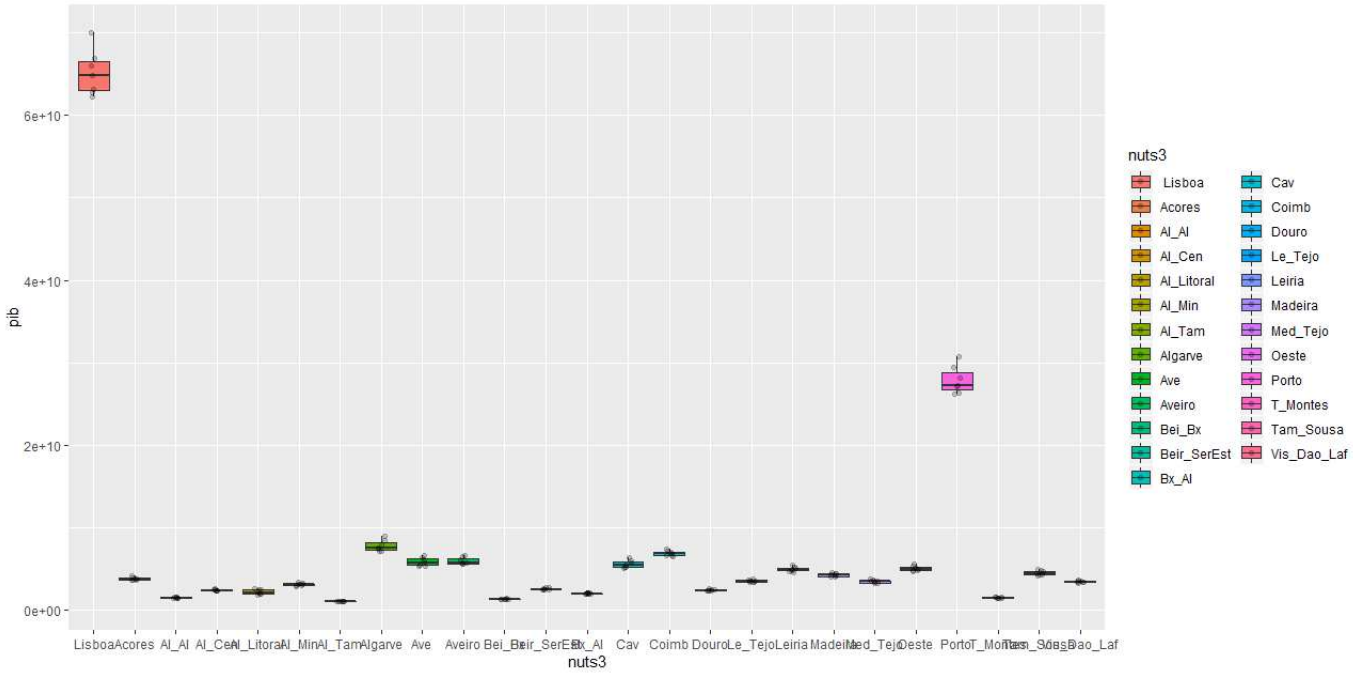


Figura A.3: pib por região nas 25 regiões

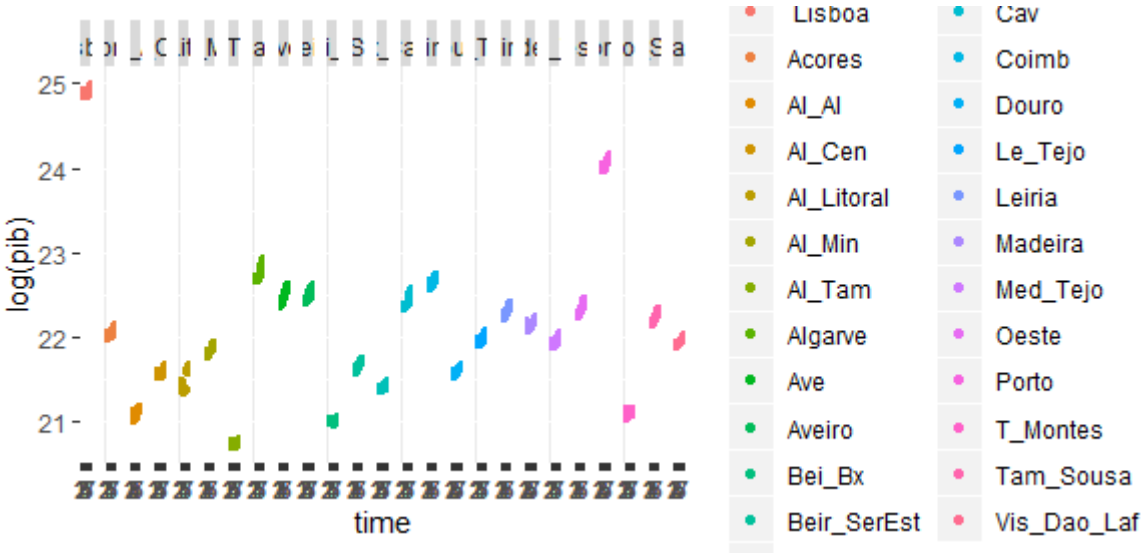


Figura A.4: Relação entre pib e tempo por região