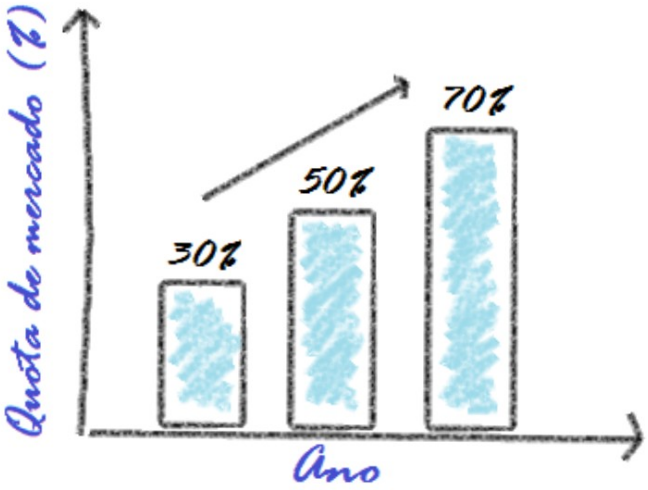


Modelação de Quotas de Mercado

Carla Sofia da Silva Gonçalves
Mestrado em Engenharia Matemática
Departamento de Matemática
2015

Orientador
Prof. Doutor Joaquim Pinto da Costa, FCUP
Co-orientadora
Profª Doutora Margarida Brito, FCUP

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, / /

Agradecimentos

Gostaria de agradecer ao Professor Joaquim Costa e à Professora Margarida Brito por terem aceite orientar o meu trabalho e pela disponibilidade e apoio prestados quer durante o decorrer do estágio quer durante a escrita da dissertação.

Aos meus pais um agradecimento especial por terem permitido que chegasse até aqui, apoiando sempre as minhas decisões, e à minha irmã, a Francisca, um obrigado por toda a sua energia contagiante.

A toda a equipa da SONAE SR que me acompanhou durante o estágio, em especial ao Rui Santos, Carla Araújo, Sandra Cardoso, Hugo Neves e Mafalda Pinto.

Deixo ainda um grande obrigada ao Ricardo, por todas as sugestões, paciência, apoio e carinho.

Por último, mas não menos importante, agradeço a todos os amigos e colegas que me acompanharam durante este percurso pela Faculdade de Ciências.

Resumo

Uma vez que as quotas de mercado são por vezes integradas em índices de desempenho duma empresa, a quota de mercado é frequentemente uma variável tão ou mais importante que as receitas e lucros. A análise de quotas de mercado é mais complexa que a análise de vendas, uma vez que estas não são o resultado do desempenho de apenas um produto, nem mesmo apenas da empresa em causa. Além disso, as ferramentas não estão tão desenvolvidas quanto à análise de vendas, nem os dados são tão compreensivos ou precisos, uma vez que englobam o comportamento das várias empresas que constituem o mercado. Esta dissertação é o resultado da colaboração com uma empresa líder mercado, inserida no contexto dum estágio curricular.

Vários métodos tradicionais foram explorados, nomeadamente de forma a modelar a resposta da quota a campanhas promocionais: regressões lineares, logísticas, redes neuronais e máquinas de suporte vetorial, assim como explorados modelos de atração, modelos não-lineares específicos para a modelação de quotas de mercado.

Foram também aplicados métodos menos ortodoxos Os modelos hierárquicos, normalmente aplicados a vendas, no geral melhoram a qualidade dum modelo quando se tem informação da série a vários níveis de agregação. A sua aplicação é menos imediata na aplicação a quotas de mercado e uma solução original é sugerida. Por último, foi implementado um modelo de escolha, uma metodologia premiada pelo Nobel de 2000, a qual tenta inferir parâmetros de escolha dos consumidores.

Durante o trabalho é feita uma digressão formal pelos vários modelos, juntamente com as respetivas aplicações em R. Espera-se desta forma que a dissertação tenha uma utilidade mais ampla que o caso em estudo.

Palavras-chave: quota de mercado; séries temporais; data mining; econometria; modelos hiérarquicos; modelos de atração; modelos de escolha.

Abstract

Since market shares are sometimes integrated into market performance indices of firms, the market share is often a variable as, if not more, importante than revenue and profits. Analyzing market shares is more complex than analyzing sales, since they are not the result of the performance of a single product, nor solely of the firm being studied. Besides, tools are not as developed as those to study sales, neither is data as comprehensive or accurate, since it encompasses the behavior of the several firms which make the market. This thesis is the result of a collaboration with a firm that is the market leader, within the scope of a curricular internship.

Several traditional methods were explored, namely as a way to model the market share response to promotional campaigns: linear regressions, logistic regressions, neural networks and vectorial support machines, and we explore attraction models, non-linear models specific for the modelling of market shares.

Other less orthodox methods are applied as well. Hierarchical models, usually applied to sales, in general improve the quality of the model when there are time series data at several levels of aggregation. Its application is less linear when applied to market shares and an original solution is proposed. Lastly, a choice model was implemented, a methodology awarded with the Nobel of 2000, which tries to infer choice parameters from consumers.

During this work, a formal digression is perfomed through the several models, together with their respective application in R. This thesis is therefore expected to have an wider application than the case study at hand.

Keywords: market share; time series; data mining; econometrics; hierarchical models; attraction models; choice models.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este trabalho foi realizado no seguimento de um estágio curricular numa empresa líder de mercado no retalho não alimentar; mais concretamente na área dos eletrodomésticos, eletrónica de consumo e entretenimento, inserida na SONAE SR.

Dado o crescimento do mercado e, consequentemente, o aumento da competição entre empresas, torna-se importante não só prever a quantidade e valor vendidos como também conhecer o quanto se vende em relação aos concorrentes. Neste sentido, surge o conceito de quota de mercado.

A quota de mercado permite avaliar melhor o impacto de variáveis de marketing do que as próprias vendas. Esta é, sem dúvida, uma variável importante para as empresas. No entanto, a sua definição não é trivial dada a ambiguidade do termo mercado.

No caso em estudo, a quota de mercado é dada pela fração de vendas brutas reais (em valor monetário) da empresa relativamente aos concorrentes, para um período mensal, num dado conjunto de produtos e determinada área geográfica. O mercado total é constituído pela empresa e pelos concorrentes com superfícies de venda de grande dimensão.

1.1 Problema

Como referido, a quota de mercado no caso em estudo é dada mensalmente para diferentes categorias de produtos e diferentes zonas geográficas. Assim sendo, o objetivo deste trabalho foi analisar o comportamento da variável quota de mercado, realizando a sua previsão no horizonte de um mês, para as diferentes combinações categoria de produtos/ região. Os dados disponíveis dizem respeito ao período compreendido entre janeiro de 2011 e novembro de 2014. Por razões de confidencialidade, estes dados foram transformados e ainda normalizados. Na prática, um dos interesses deste estudo é perceber qual o impacto das campanhas promocionais na quota.

Não há muitas variáveis com que trabalhar, o que limita o que pode ser feito. Os dados foram recolhidos de relatórios anuais produzidos por uma terceira parte, um consultor, que nos foram facultados ao longo do estágio em PDF. Depois de devidamente processados e digitalizados, estes incluem: a quota da Sonae a nível nacional, assim como regional, e também as suas vendas. Colegas na empresa forneceram informação sobre quatro campanhas publicitárias da Sonae, assim

como de outros dois concorrentes. Também conseguimos informação sobre a área e localização das várias lojas, que dividimos em regiões. Estes dados serão explorados em detalhe no Capítulo 2. No Capítulo 3 é feito um resumo do estado da arte, e a sua aplicação é realizada nos capítulos posteriores.

1.2 Estrutura da Dissertação

No **Capítulo 2**, é feita uma apresentação dos dados disponíveis para este trabalho. No **Capítulo 3** apresenta-se um resumo da literatura consultada, referente a problemas que envolvem o estudo de quotas de mercado. Os dados disponíveis não permitiram a aplicação de todos os modelos consultados. Os capítulos 4, 5 e 6 consistem numa revisão teórica aprofundada dos modelos apresentados no estado da arte, e a sua aplicação.

No **Capítulo 4** é tratada a análise de séries temporais, nomeadamente através do cálculo das autocorrelações e estimação das componentes tendência, sazonalidade e resíduo. É ainda considerada a modelação de séries utilizando modelos autoregressivos integrados de médias móveis (ARIMA).

No **Capítulo 5**, abordam-se modelos em data mining, nomeadamente regressão linear e logística, árvores de regressão, máquinas de suporte vetorial e redes neuronais. Estes modelos explicam a trajectória da quota como resposta de esforços promocionais.

O **Capítulo 6** trata um modelo Bayesiano, proposto por Chen e Yang (2007), cujo objetivo é simular o comportamento dos clientes utilizando dados agregados. Numa perspetiva um pouco diferente são apresentados também modelos de atração que vêm a quota como resultado do quociente entre a atração que uma empresa suscita nos consumidores e a soma da atração de todas as empresas que constituem o mercado (Cooper e Nakanishi, 1988).

Nos **capítulos 7 e 8** é feita, respetivamente, a apresentação e análise dos erros cometidos em cada método e conclusão do trabalho. No **Capítulo 7** são apresentados também modelos hierárquicos cujo objetivo é usar a hierarquia das séries temporais (uma vez que são consideradas 5 regiões que no seu conjunto formam Portugal Continental) para tentar obter resultados mais robustos combinando as previsões independentes de cada região.

Uma vez que a elaboração da tese foi uma oportunidade para aprender novos modelos e explorar conceitos do curso foi feita uma digressão formal pelos vários modelos, juntamente com as respetivas aplicações em R. Espera-se desta forma que a tese tenha uma utilidade mais ampla que o caso em estudo.

Capítulo 2

Descrição dos Dados

Neste capítulo é feita uma apresentação e primeira análise dos dados disponíveis:

2.1 Estrutura do Negócio e Divisão Geográfica

De forma a perceber como abordar o problema de previsão das quotas foi necessário entender de que forma a empresa é gerida e, por sua vez, a sua informação estruturada. A empresa é dividida em **regiões**, cada uma das quais é responsável pelas suas **lojas**, as quais estão, por sua vez, sub-divididas por **unidade de negócio**. Esta estruturação é ilustrada na Figura 2.1.

Os produtos comercializados na empresa estão divididos em 5 grupos designados por unidade de negócio. As regiões, para Portugal Continental, estão igualmente dividas em 5 grupos, detalhadas

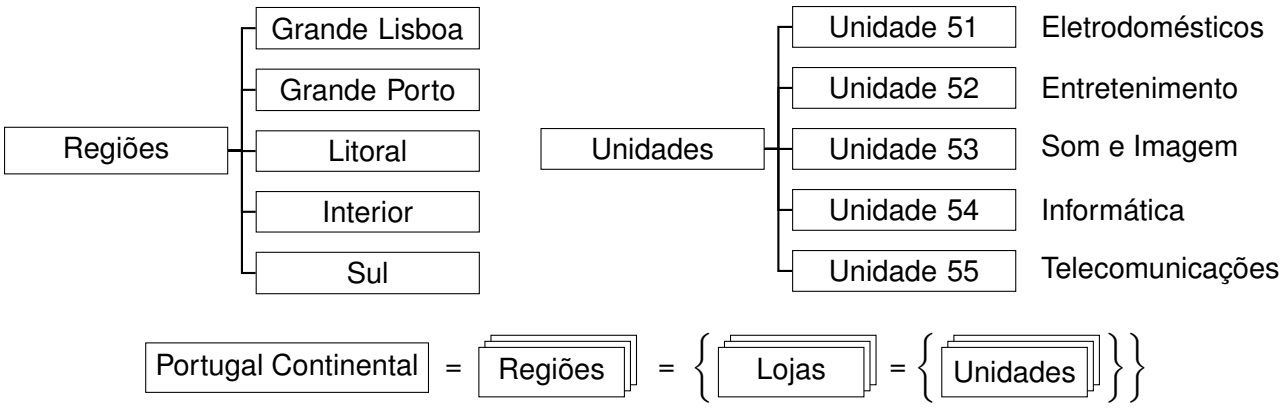


Figura 2.1: Visão geral da estrutura da empresa.

no mapa das lojas apresentado na Figura 2.2.

O estudo da quota de mercado é feito por uma segunda empresa de consultoria, especializada em estudos de mercado. Estes relatórios foram sendo fornecidos ao longo do estágio e compreendem as datas entre janeiro de 2011 e novembro de 2014 (47 meses). Para além disso, informação de outras variáveis foi fornecida por colegas, nomeadamente das campanhas publicitárias e áreas das várias lojas.

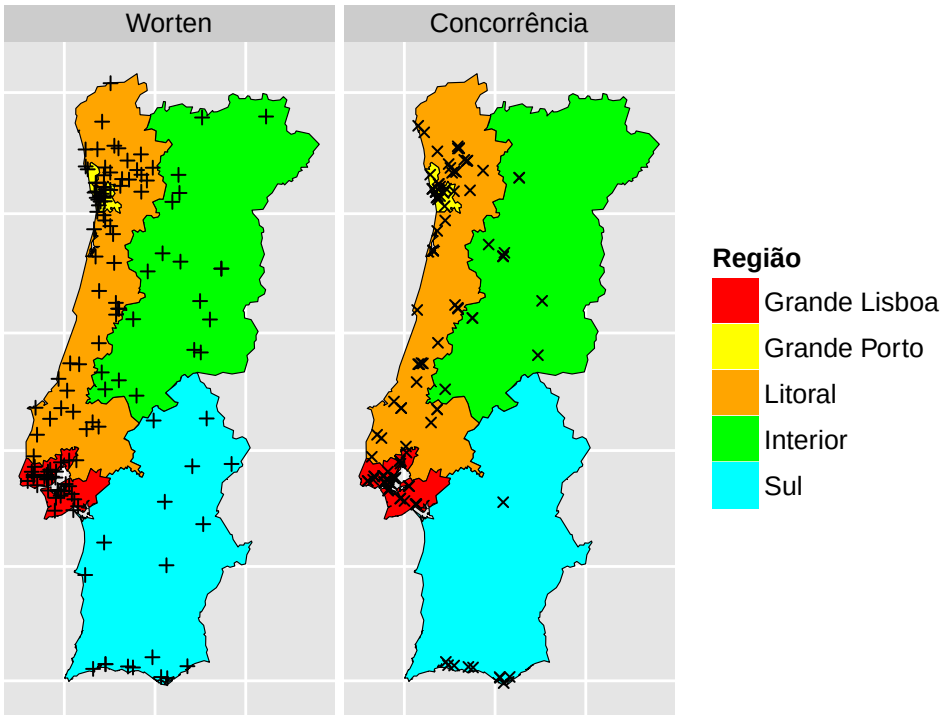


Figura 2.2: Divisão geográfica de Portugal Continental com representação de lojas.

2.2 Recolha dos Dados

Numa fase inicial do estágio foram reunidos todos os relatórios referentes ao período dos 47 meses. Os dados a que tivemos acesso combinavam as unidade 54 e 55, não tendo havido acesso a informações referentes à unidade 52. Isto deve-se ao facto da empresa e da consultora estruturarem o negócio de formas diferentes. Dos referidos relatórios foi possível extrair a seguinte informação:

- Quota total da empresa a nível de Portugal Continental;
- Quota disponível por região e a nível de Portugal Continental:
  - Total da região (excepto para a unidade 52), assim como para cada uma das suas unidades:
    - Unidade 51;
    - Unidade 53;
    - Unidade 54 e 55 (em conjunto).

Estas serão as variáveis independentes que tentaremos explicar e prever.

Durante a recolha de informação dos relatórios escritos foi colocada a questão de como complementar e explicar esta série temporal com dados da empresa, nomeadamente das várias lojas. Uma

vez que a informação sobre a quota está disponível ao nível das regiões, enquanto para as lojas ao nível da sua localização apenas se dispõe do código postal, a solução passou por associar a cada código postal o concelho, distrito e, a cada distrito, a região, dentro das 5 referidas. Isto foi possível com a utilização da ferramenta de gestão de bases de dados, Access do Microsoft Office.

Com base em informação que nos foi disponibilizada foi construído um modelo entidade-relação no Microsoft Access, conforme o diagrama da Figura 2.3. Além disso, a base de dados foi complementada com informação demográfica do INE (INE, 2011, 2013).

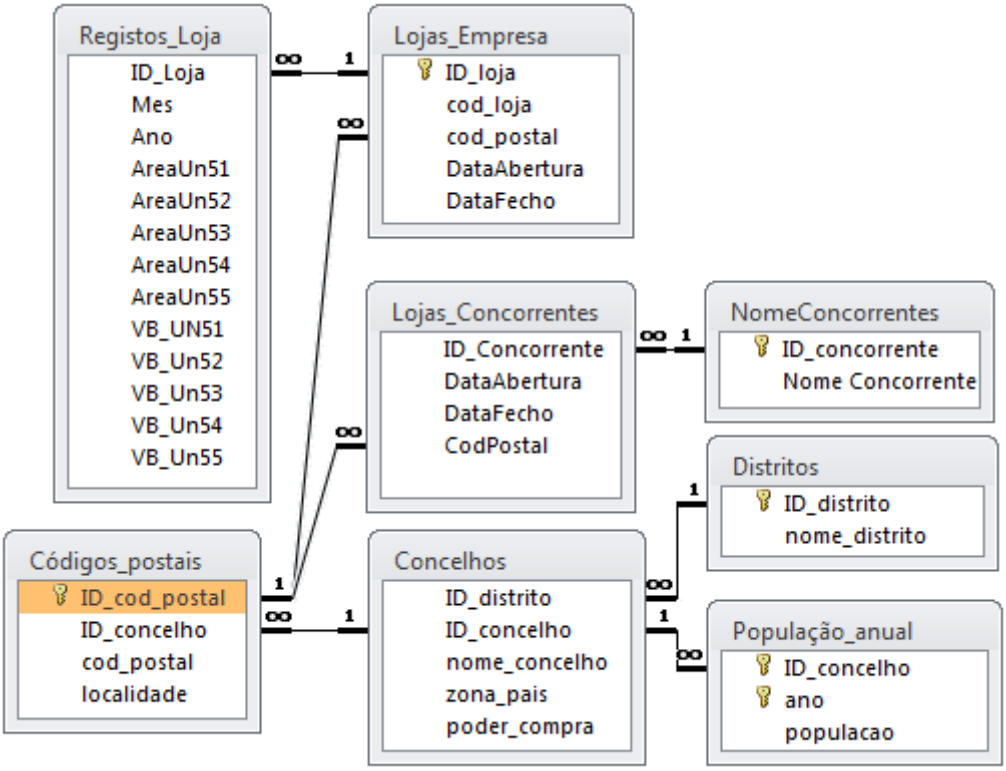


Figura 2.3: Diagrama Entidade-Relação do Microsoft Access.

Resumidamente, as tabelas inseridas foram:

- **tabela com distritos:** cada distrito tem um identificador (ID_distrito) que o distingue dos restantes;
- **tabela com a população anual por concelho**;
- **tabela com concelhos:** à semelhança dos distritos, cada concelho tem um identificador (ID_concelho) único; Além disso, por cada concelho é registada a região a que pertence (Grande Lisboa, Grande Porto, ...) e ainda a proporção de poder de compra;
- **tabela com códigos postais:** são registados todos os códigos postais; por cada código postal é registado o concelho a que este pertence.

Com estas tabelas foi possível aceder aos códigos postais de cada uma das regiões bem como a outras informações tais como a população no ano 2011 na Grande Lisboa, por exemplo. As tabelas restantes referem-se às lojas da empresa e dos concorrentes:

- **tabela com informação geral das lojas da empresa:** registo do código postal da loja, data de abertura e fecho;

- **tabela com informação mais especifica ao nível de cada loja:** registo mensal de vendas brutas e áreas de venda por unidade de negócio;
- **tabela com informação das lojas da concorrência:** registo do concorrente e código postal das lojas bem como datas de abertura/fecho das mesmas.

Uma vez trabalhada a informação num formato útil foi decidido combinar os dois meios de informação numa folha de cálculo única, uma estrutura também conhecida em estatística por *data frame*. Teve que ser adotado o denominador para o qual havia informação: sendo que a unidade da série temporal escolhida foi o mês, a unidade da série geográfica foi a região, e as unidade de negócio foram agrupadas dentro das três consideradas pela consultora.

Uma vez que a quota é um "jogo" entre empresas, tentamos obter mais variáveis exógenas referentes à competição para modelar a quota da empresa. Foram, por isso, recolhidos dados respeitantes ao histórico das **campanhas promocionais** no mesmo período, de janeiro de 2011 a novembro de 2014, tanto da própria empresa como das empresas concorrentes. Esta informação não proveio dos relatórios, mas de estimativas internas. Visto esta informação ser suscetível a erros foi feita uma tentativa de verificação da veracidade desta recolha. Os dados respeitantes à própria empresa, presumiu-se serem mais fiáveis, tendo sido apenas validadas as datas das campanhas. Apenas as principais campanhas promocionais foram consideradas; por principais entenda-se campanhas abrangentes que tenham sido transversais entre todos os produtos e regiões. Estas campanhas têm durações de apenas alguns dias, no máximo uma semana; no entanto, registou-se apenas os meses nas quais estas foram observadas.

Resumidamente, os dados disponíveis para este projeto são:

- os valores de quota referidos anteriomente;
- número de lojas da empresa e dos seus concorrentes, por região;
- histórico de principais campanhas promocionais da empresa e dos concorrentes;
- área de venda por unidade de negócio e por região (apenas as da empresa);
- vendas brutas por unidade de negócio e por região (apenas as da empresa);
- dados demográficos das várias regiões: proporção do poder de compra e número de habitantes.

2.3 Normalização das Variáveis

Ao longo da tese, por razões de confidencialidade, algumas estimações serão omitidas, algumas escalas nos gráficos não serão representadas, e promoções e concorrentes serão identificados e codificados por números.

Valores como vendas e quotas serão representados mas além de sofrerem uma transformação são ainda normalizados. Esta normalização será feita apenas durante a representação gráfica e não foi feita como pré-processamento. Dentro de várias funções de normalização de dados consideradas, optamos por escalar os valores das variáveis para o intervalo unitário usando a seguinte função:

$$f(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

Esta função conhecida por "featuring scale" é muito utilizada em data mining, habitualmente de forma a evitar trabalhar com variáveis de magnitudes muito diferentes (Aksoy e Haralick, 2001).

2.4 Descrição das Variáveis

Faz-se uma enumeração das variáveis longitudinais recolhidas para os 47 meses, janeiro de 2011 a novembro de 2014, bem como a sua designação no estudo. Além de variar ao longo do tempo, algumas variáveis variam também por região, $r \in \{\text{PC}, \text{GL}, \text{GP}, \text{L}, \text{I}, \text{S}\}$, e unidades de negócio, $u \in \{\text{UN51}, \text{UN53}, \text{UN54/55}, \text{total}\}$. A granularidade das variáveis é resumida no diagrama da Figura 2.4.

- **QUOTA**(t, r, u): quota nacional e das várias regiões para as várias unidades, excepto a unidade 52;
- **PROMO*k***(t): dentro da própria empresa são considerados 4 tipos distintos de campanhas promocionais k, de forma a melhor discernir a influência de cada tipo de campanha na quota. Trata-se de uma variável binária que toma valor 1 se há a campanha nesse mês e 0 caso contrário. Apenas as campanhas promocionais abrangentes são consideradas.
  De notar que as campanhas são iguais em todas as regiões consideradas e, excetuando a PROMO4, são comuns às três unidades de negócio consideradas.
- **CONCO*j***(t): campanha promocional da empresa concorrente j, com $j = \{1, 2\}$ para a unidade 51, $j = \{1, 2, 4\}$ para a unidade 53 e $j = \{1, 2, 4, 5\}$ para as unidades 54 e 55. Para cada unidade de negócio, a variável é igual em todas as regiões. Diferentes unidades de negócio apresentam diferente número de concorrentes e, portanto, há unidades de negócio com mais variáveis, relativas às campanhas da concorrência, que outras. Trata-se de uma variável binária. Não é considerada qualquer campanha do concorrente "3" uma vez que este não realizou nenhuma campanha abrangente nos últimos quatro anos.
- **AREA**(t, r, u): área total nas lojas da própria empresa para cada uma das unidades, em m^2.
- **VENDAS**(t, r, u): total de vendas brutas nas várias unidades das lojas da própria empresa, em euros.
- **NRLOJAS0**(t, r): número de lojas da própria empresa.
- **NRLOJAS*j***(t, r): número de lojas do concorrente j que contém as várias unidades de negócio. Tal como CONCOj, j varia com o número de concorrentes.
- **POP**(t, r): variável demográfica representativa da população na região, recolhida do INE (INE, 2013).
- **PPC**(t, r): variável demográfica representativa do poder de compra na região per capita em milhares de euros, recolhida do INE (INE, 2011).

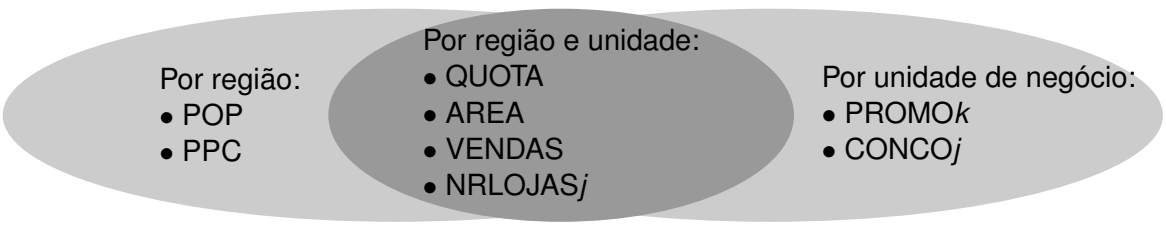


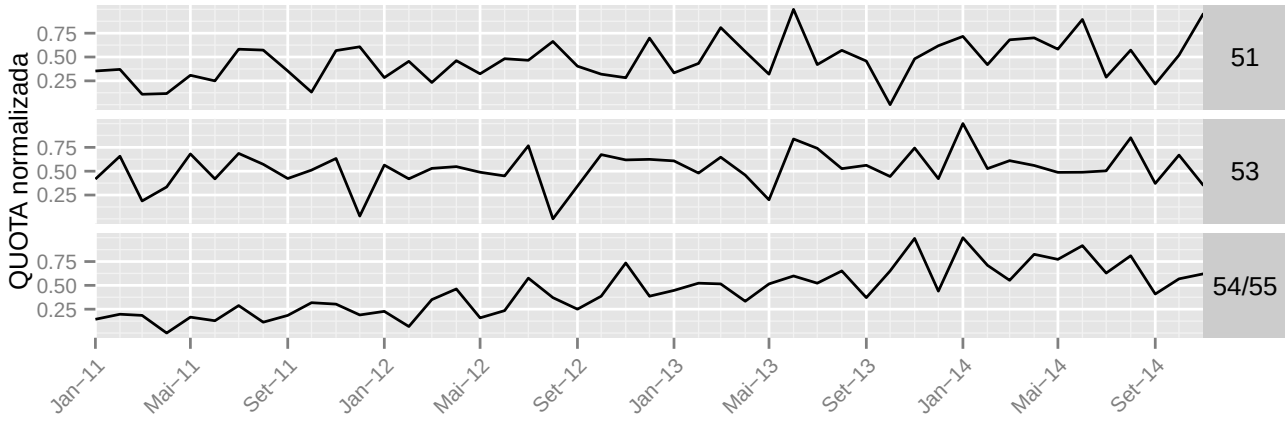
Figura 2.4: Esquema geral da granularidade das variáveis.

2.5 Primeira Análise dos Dados

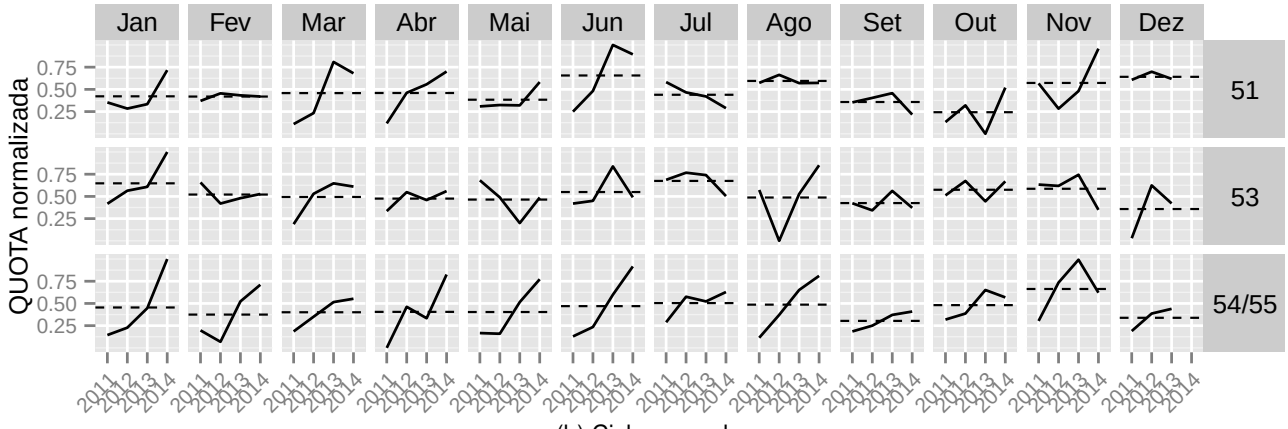
Construída a base de dados, procedemos à análise dos mesmos. Como referido, por questões de confidencialidade, algumas variáveis foram transformadas. Apesar de transformados, vamos supor que os dados continuam a refletir o comportamento dos valores reais.

Quota de Mercado. Esta variável foi recolhida, como referido anteriormente, para diferentes categorias de produtos e diferentes regiões. É apresentada a série temporal para as três unidades disponíveis na Figura 2.5, assim como o ciclo mensal que não demonstra qualquer efeito de sazonalidade. No capítulo 4 estas séries temporais serão estudadas com mais detalhe

Na figura é ilustrada a série temporal apenas para Portugal Continental; no entanto, é de notar que cada uma das regiões tem um comportamento concordante com a sua unidade. A diferença entre as séries é sobretudo entre unidades de negócio. A Figura 2.5b, conhecida como gráfico por ciclo mensal, representa a quota com as observações em cada mês com a respetiva média. Não são evidentes efeitos de sazonalidade dentro de cada mês, sendo a média muito semelhante.



(a) Série temporal



(b) Ciclo mensal

Figura 2.5: Representação da quota em série temporal e ciclo mensal para as três unidades de negócio para Portugal Continental.

Vendas Brutas. Tal como as áreas, esta variável é recolhida para cada uma das lojas sendo depois selecionadas as lojas por região e procedendo-se à soma de vendas por mês e categoria de produtos. A análise da Figura 2.6 na página ao lado permite observar que as vendas brutas nas diferentes regiões têm um comportamento semelhante. De observar ainda que o comportamento da

variável de ano para ano é bastante semelhante.

É de notar que, ao contrário das quotas, nas vendas existe um efeito sazonal, com a média das vendas muito mais pronunciada em dezembro. Este efeito sazonal não se reproduz nas quotas de mercado.

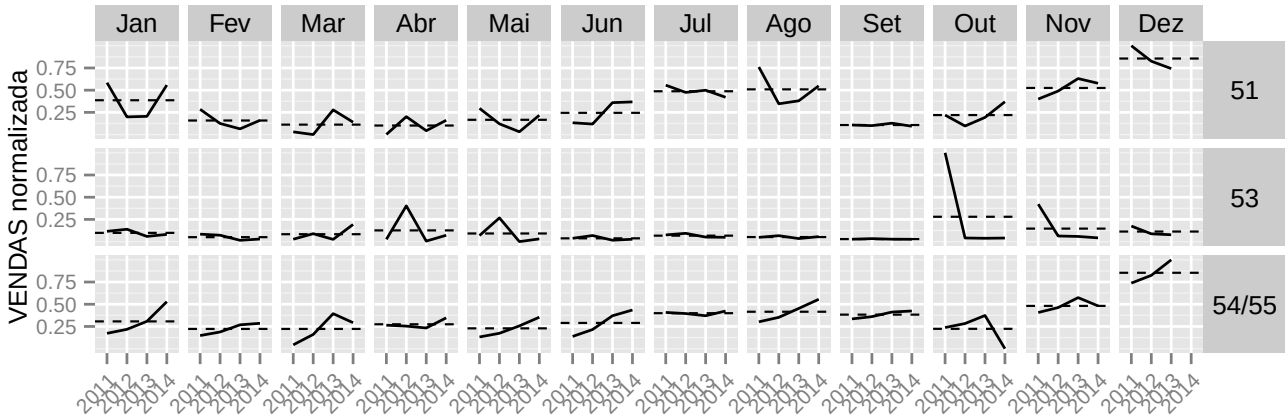


Figura 2.6: Vendas brutas normalizadas.

Área. Esta variável é medida ao nível de cada loja, sendo que a base de dados construída permite aceder às lojas por região; assim sendo, foram somadas as áreas para cada unidade de negócio ao longo dos meses. A análise da Figura 2.7 permite observar que em alguns períodos não houve alteração significativa nas áreas das lojas.

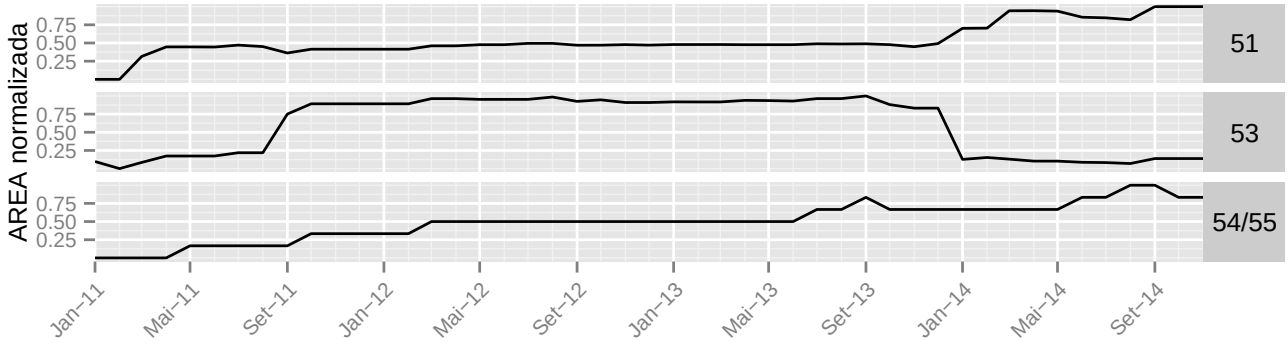


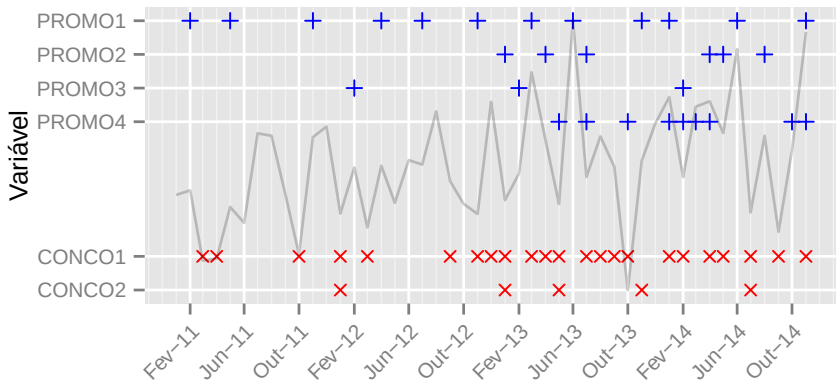
Figura 2.7: Áreas normalizadas para as várias unidades de Portugal Continental.

Número de Lojas. O número de lojas da empresa era conhecido com precisão. No que respeita aos concorrentes, a recolha teve de ser mais cuidada. Como visto na listagem de variáveis na secção anterior, para a unidade 51 apenas foram considerados três concorrentes. Para a unidade 53, foi considerada um quarto concorrente e para as unidades 54 e 55 um quinto. As lojas encontram-se representadas geograficamente na Figura 2.2.

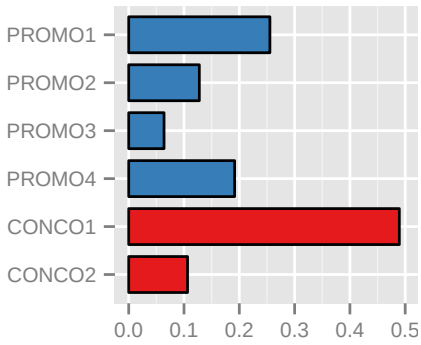
De notar que, em algumas regiões o número de lojas de algumas empresas se manteve inalterado no período em estudo. Infelizmente esta pequena variação nas áreas e no número de lojas durante os quatro anos em consideração significa que estas variáveis serão de fraco poder preditivo.

Campanhas Promocionais. As campanhas promocionais consideradas são iguais em todas as regiões. Esta variável é uma variável binária, tomando valor 0 quando não há campanha e valor 1 caso contrário. Por mês, cada promoção considerada não ocorreu mais do que uma vez. Para representar esta variável é feito um gráfico de barras com as frequências relativas de registos com

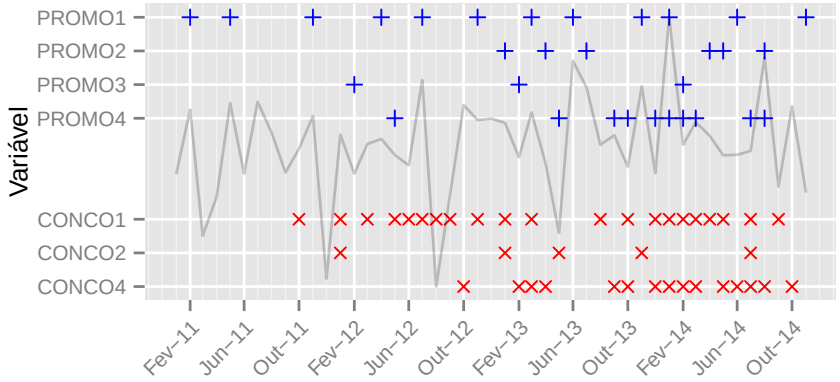
ou sem promoção, o que corresponde a analisar a proporção de meses em que houve ou não a campanha, ver a Figura 2.8:



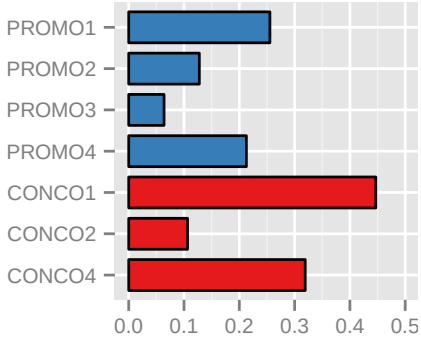
(a) Registo de ocorrência de campanha da unidade 51.



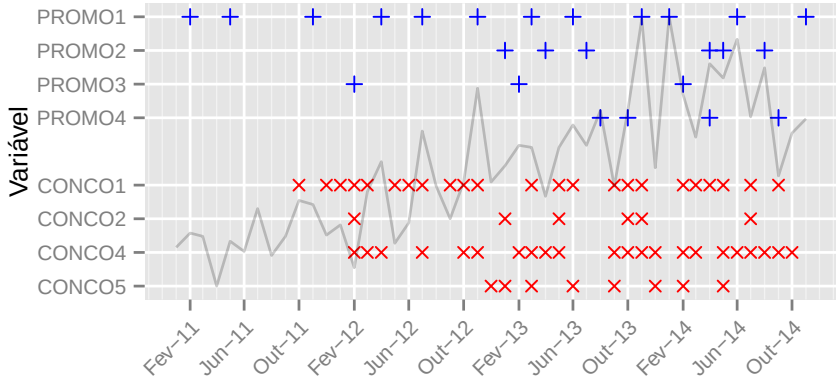
(b) Frequências relativas (47 meses).



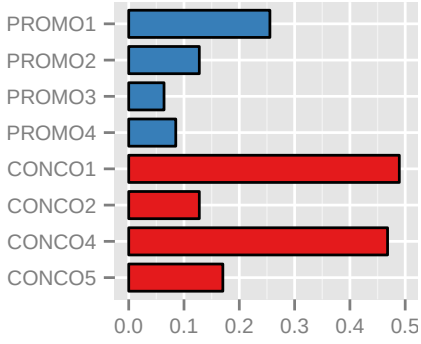
(c) Registo de ocorrência de campanha da unidade 53.



(d) Frequências relativas (47 meses).



(e) Registo de ocorrência de campanha das unidades 54 e 55.



(f) Frequências relativas (47 meses).

Figura 2.8: Frequências absolutas da variável campanha.

De notar que as campanhas PROMOk, com $k \in \{1, \cdots, 4\}$, dizem respeito aos quatro tipos de campanhas promocionais da própria empresa e CONCO1 e CONCO2 às campanhas das duas empresas na competição, no âmbito da unidade de negócios 51. Como se pode ver, ocorreram, quanto muito, duas das quatro campanhas. Neste caso, num dado mês a empresa pode ter mais do que uma campanha promocional, o que de facto acontece.

Observa-se que houve mais campanhas do tipo CONCO1. No entanto, no total, a empresa fez mais campanhas promocionais. Podemos ainda verificar que não há um padrão no histórico de promoções.

2.6 Análise da Correlação entre Variáveis

É importante na análise de um conjunto de variáveis perceber de que forma estas se correlacionam. Um estimador não será estável fazendo uso de variáveis explicativas correlacionadas, além de não ser possível separar os efeitos umas das outras devido a *multicolinearidade* (Haitovsky, 1969).

Os coeficientes de correlação medem a correlação entre duas variáveis. No entanto, diferentes coeficientes deverão ser tomados conforme o tipo das variáveis em causa. De facto, neste trabalho podem considerar-se dois tipos de variáveis: **quantitativas**, como é o caso das quotas, áreas, vendas brutas e número de lojas e **binárias** ou **dicotómicas**, que indicam a existência ou não de campanha promocional.

Entre duas variáveis quantitativas pode usar-se o coeficiente de correlação linear de Pearson. Para as restantes combinações de variáveis as medidas consideradas correspondem a medidas de correlação derivadas deste coeficiente (Lira e Neto, 2006).

Coeficiente de Correlação de Pearson. Este coeficiente de correlação mede o grau de relação linear entre duas variáveis quantitativas. Se existe uma relação linear perfeita entre os valores das variáveis, o coeficiente toma valor 1 ou -1 (o valor -1 indica uma relação linear perfeita mas inversa, isto é, quando uma das variáveis aumenta a outra diminui), quando não existe qualquer relação linear o coeficiente toma valor 0. A sua fórmula é dada por:

$$\rho_{X,Y} = \frac{\text{Cov(X,Y)}}{\sigma_X \sigma_Y}.$$

R A função `cor{stats}` permite calcular o correspondente coeficiente amostral.

Coeficiente de Correlação do Ponto Bisserial. Este coeficiente é utilizado quando uma variável é quantitativa e outra é dicotómica (Lira e Neto, 2006). Tomando X como a variável quantitativa e Y a variável dicotómica, o coeficiente de correlação é dado por

$$\frac{(\bar{X}_1 - \bar{X}_0)\sqrt{\pi(1-\pi)}}{S_x},$$

onde $\bar{X}_0$ e $\bar{X}_1$ são as médias de X quando $Y = 0$ e $Y = 1$, respetivamente, π é a proporção de $Y = 1$ e S_x o desvio padrão da variável X.

R Utilizando o comando `biserial.cor{ltm}` é possível obter o valor do coeficiente ponto bisserial (Rizopoulos, 2013).

Coeficiente de Correlação Phi. Utilizado para determinar a correlação entre variáveis dicotómicas, o coeficiente Phi é determinado através da tabela de frequências (Lira e Neto, 2006).

A título de exemplo, dada a seguinte tabela de contingência:

		Variável *x*		
Variável *y*		1	0	**Total**
	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d$

Então, o coeficiente Phi é dado por

$$\Phi = \frac{a-(a+b)(a+c)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

R A função `phi{psych}` permite aplicar este coeficiente (Revelle, 2015).

Com base nos métodos anteriores, procede-se à reprodução gráfica na Figura 2.9 da matriz de correlações entre as várias variáveis quantitativas e binárias:

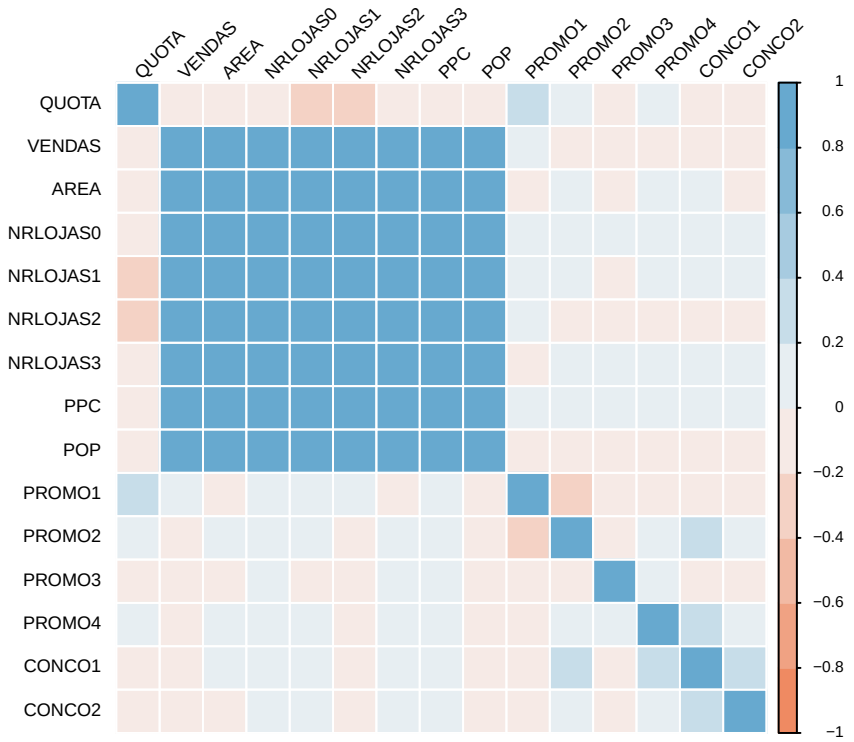


Figura 2.9: Correlações entre as variáveis recolhidas para un51.

A análise das correlações permite verificar um cluster de correlações lineares positivas entre todas as variáveis explicativas quantitativas consideradas. A forte correlação entre as vendas e as áreas será uma consequência da forte correlação entre o número de lojas e as áreas pois a abertura de uma loja leva ao aumento da área de negócio e também ao aumento do volume de vendas; o contrário acontece quando uma loja fecha. A correlação entre o número de lojas da empresa e o valor das áreas já era de esperar que fosse forte, uma vez que parte dos aumentos no valor da área foram resultado da abertura de lojas, bem como algumas reduções foram resultado do fecho.

De entre todas as variáveis fortemente correlacionadas, aquela com maior intervalo, e portanto nos fornecerá mais informação, são as vendas pelo que as restantes deverão ser descartadas.

Regiões. Olhando para cada região, observamos que estas variáveis parecem estar correlacionadas porque variam pouco dentro de cada região, mas muito entre regiões. É possível que sejam todas elas proxies para a variável nominal região. Uma vez que não faz sentido falar de correlações lineares entre uma variável nominal e uma variável quantitativa, vamos testar se existe correlação doutra forma, através do coeficiente de ajustamento duma regressão múltipla.

Para começar vamos considerar o gráfico caixa-bigodes das regiões em relação às vendas da Figura 2.10:

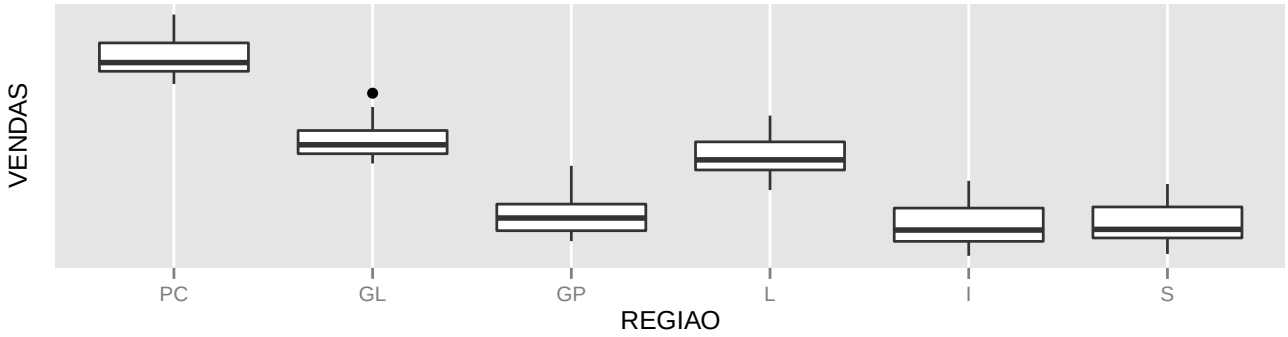


Figura 2.10: Gráfico caixa-bigodes das vendas para as várias regiões, em escala logarítmica (valores suprimidos).

Tomemos a tautologia tal que

$$\text{E[VENDAS|REGIAO]} = \text{E[VENDAS|PC]} \times \text{PC} + \text{E[VENDAS|GL]} \times \text{GL} + ... + \text{E[VENDAS|S]} \times \text{S},$$

sendo PC,..., S, variáveis dictómicas (mutualmente exclusivas) representando 1 ou 0 conforme o valor da região.

Podemos escrever esta equação sobre a forma duma regressão,

$$\text{VENDAS(REGIAO)} = \beta_0 \text{PC} + \beta_1 \text{GL} + ... + \beta_5 S + \varepsilon,$$

em que β_0 = E[VENDAS|PC], β_1 = E[VENDAS|GL], ..., β_1 = E[VENDAS|S]. Temos assim uma forma de exprimir as vendas à custa da variável região (ver Figura 2.11; a linha de regressão é curva porque a escala-y é logarítmica).

O coeficiente de determinação dar-nos-á uma estimativa do quão bem esta regressão aproxima as vendas usando as regiões como variável explicativa. O coeficiente para esta regressão é dado por $R^2 = 1 - SS_{\text{res}}/SS_{\text{tot}} = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, sendo, no nosso caso, $R^2 = 0.89201$.

Uma das interpretações deste coeficiente é que as regiões explicam 89% da variação das vendas, sendo $R = 0.94$. Podemos assim concluir que existe uma forte correlação linear entre as vendas e as regiões e iremos assim utilizar apenas as regiões para modelar as quotas. As regiões serão usadas ou como variável explicativa ou nalguns casos serão modeladas as quotas apenas para os dados ao nível de cada região.

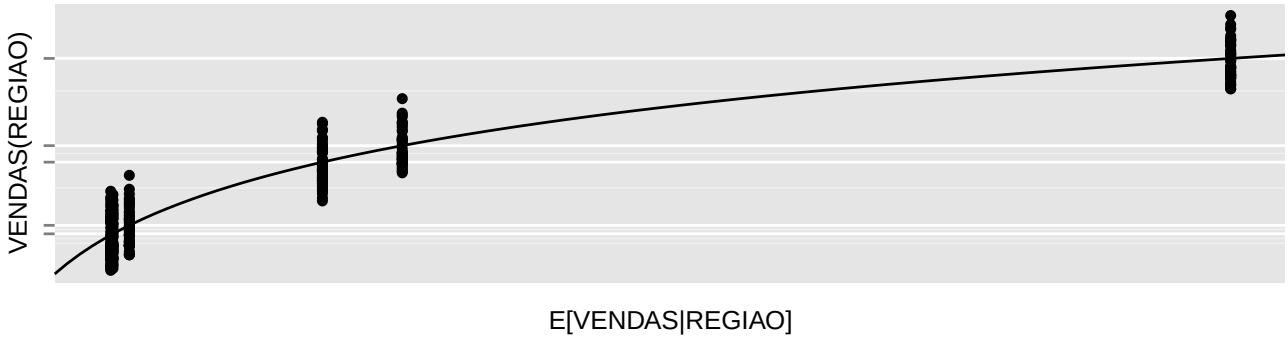


Figura 2.11: Regressão linear das vendas para as várias regiões, com as vendas em escala logoritmica.

2.7 Validação

Será necessário determinar quais os modelos que melhor se ajustam aos dados. Queremos também validar o modelo final.

Dentro da série que dispomos de janeiro de 2011 até novembro de 2014 (47 meses), iremos reservar os últimos 7 meses (cerca de 15% dos dados) para o propósito de validação do modelo final:

- **Treino:** janeiro de 2011 a abril de 2014
- **Teste:** maio de 2014 a novembro de 2014

O treino é por sua vez dividido para a escolha do modelo. Visto termos poucos dados, usamos validação cruzada que consiste em fazer várias amostragens para treino e teste, sendo que o erro dos vários testes é sujeito a uma medida de centralidade (e.g. média ou um quartil) para a escolha do melhor modelo. Uma vez que queremos fazer uso de *lagging* (atraso) na variável dependente (quota) e nas variáveis independentes, uma hipótese será usar validação cruzada temporal que faz amostragens dos dados em janelas, mantendo intacta a ordem dos dados.

Duas técnicas de validação cruzada temporal são frequentemente empregues. **Janela crescente** (*growing window*): em que os dados são divididos em amostras cada vez maiores; a primeira amostra vai da primeira observação até uma i-ésima observação e, a partir daí, da primeira observação até à $(i+k)$-ésima observação, da primeira observação até à $(i+2k)$-ésima, etc, sendo os últimos dados da amostra usados para treino. **Janela deslizante** (*sliding window*): é semelhante à anterior, apenas que o tamanho da janela se mantém constante, de forma que os dados são divididos em amostras que vão da primeira até à k-ésima observação, da segunda até à $(k+1)$-ésima observação, etc, sendo as últimas observações da amostra usadas para teste. Um passo s maior que 1 pode ser usado, de forma que os dados são divididos em $[1, k]$, $[s, k+s]$, $[2s, k+2s]$, etc (Hyndman, 2014, secção 5.1). Para o propósito desta dissertação será utilizada a janela deslizante.

Nem sempre é necessário acedermos a observações consecutivas na base de dados desde que mantenhamos as janelas temporais intactas. Por exemplo, se quisermos usar *lag* numa variável, podemos criar novas variáveis $X_1(t) := X(t-1)$, $X_2(t) := X(t-2)$, etc de forma a que a amostragem duma observação t não perturbe o *lag*. Cada índice reflecte agora uma janela temporal e, portanto, a validação pode ser agnóstica em relação ao tempo. Isto é possível para os modelos do Capítulo 5 de data mining que não seguem um processo estocástico subjacente e para os quais foi utilizado k-

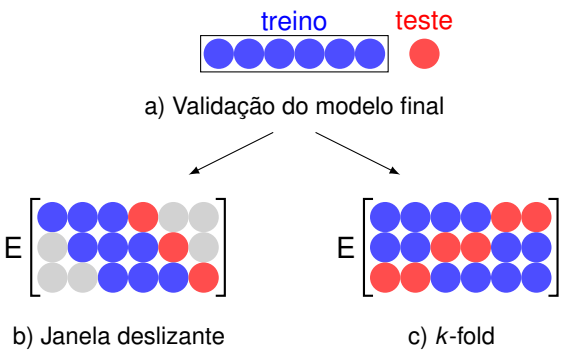


Figura 2.12: Esquemática da validação dos modelos usados.

fold. O k-fold consiste em baralhar a indexação dos dados e, em seguida, dividi-los por k partições: uma das partições é usada para teste e as outras para treino; isto k vezes para as k partições. Foi utilizado $k = 6$, para obter divisões treino-teste de ~85-15, pois é frequentemente observado que um k muito agressivo leva a que a variação dos testes de validação seja muito elevada (Kohavi, 1995).

Outro ponto importante na avaliação do desempenho dum modelo é a **medida de erro** que se usa para tal. As mais comuns são apresentas na Tabela 2.1.

Foi ponderada a utilização de duas destas medidas: a MAE e a MAPE. Suponhamos que a quota real em dois meses é 20% e 30%, respetivamente, e que a previsão foi de 22% e 32%. Utilizando a MAE, ambas as situações têm um erro de 2%, utilizando a MAPE, o erro é de 10% e 6.7%, respetivamente. O MAPE amplia assim os erros e para valores observados mais baixos há uma maior amplificação. Uma vez que errar quando a quota é muito baixa parece mais grave do que errar quando a quota tem valor elevado, optamos por utilizar o MAPE.

Medidas Absolutas
Erro quadrático médio MSE= $\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2$
Média dos desvios absolutos MAE= $\frac{1}{n}\sum_{i=1}^{n}\lvert\hat{x}_i - x_i\rvert$

Medidas Relativas
Estatística de Theil (Greene, 2011) U= $\frac{\sqrt{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - x_{i-1})^2}}$
Erro percentual médio absoluto MAPE= $\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{(\hat{x}_i - x_i)}{x_i}\right\rvert$

Tabela 2.1: Algumas medidas para avaliação do desempenho dos modelos preditivos.

Capítulo 3

Estado da Arte

É aqui feita uma primeira revisão da literatura. Estes modelos serão elaborados ao longo da dissertação, mas alguns serão descartados desde já porque não se adequam aos dados:

Um primeiro determinante na aceitação de um modelo para um dado problema é que a sua estrutura restrinja os seus valores para o domínio que faça sentido para o objeto em estudo. Ora, no caso das quotas, duas restrições são importantes: a soma das quotas ser 1 (restrição da soma), e qualquer quota se encontrar entre 0 e 1 (restrição de domínio) (Leeflang e Reuyl, 1984).

3.1 Modelos Espaciais

Um dos primeiros modelos a considerar competição no espaço foi o modelo de Hotelling (1929) que considerou a localização de duas lojas em competição num mercado linear. Neste modelo, os consumidores optavam pela loja mais próxima. Esta concorrência binária foi melhorada por Huff (1964) através da introdução dum modelo gravítico.

Em analogia à lei da gravitação $F = G\frac{m_1 m_2}{r^2}$, os modelos gravíticos são usados em vários contextos nas ciências sociais para descrever interacções entre variáveis, recorrendo a elementos como massa e distância como forma de metáfora.

No modelo de Huff, a probabilidade do cliente optar por uma loja é proporcional à sua massa (ou atracção) e inversamente porporcional a uma potência da distância. Uma formalização básica do modelo é dada por:

$$P_{ij} = \frac{S_j^{\alpha} D_{ij}^{\beta}}{\sum S_j^{\alpha} D_{ij}^{\beta}},$$

em que P_{ij} é a probabilidade de um consumidor na zona i ir à loja j, S_j é a área da loja j, e D_{ij} é a distância entre a loja e o consumidor, sendo que o coeficiente α é positivo e β é negativo. A gravitação é assim transformada em probabilidade pela divisão de cada combinação (i, j) pelo

somatório (Huff, 2003).

Uma vez que se considera estes modelos como sendo verdade para um indíviduo então também são inferidos para partes da população.

Outros modelos, mais usados por estatísticos de geologia para séries espaciais envolvendo, por exemplo, minérios, são as séries geoespaciais. Numa analogia com as séries temporais, estes são modelos cujos dados são indexados em relação à sua geografia e desta forma são modelos que relacionam a variável dependente com a sua geoespacialidade. Em particular, as chamadas regressões de Krigagem são usadas para interpolar valores em localizações desconhecidas (Sarma, 2009).

Nenhum destes modelos espaciais será considerado neste trabalho uma vez que temos valores de quota para apenas 5 grandes regiões.

3.2 Modelos de Atração

Na literatura de análise da quota de mercado há vários modelos baseados no chamado "Teorema da Quota de Mercado." Antes de proceder à explicação deste teorema vejamos em que consiste o "Teorema Fundamental de Kotler" (Kotler, 1984).

Segundo Kotler, a quota de mercado é proporcional ao esforço de marketing, ou seja,

$$s_i = cM_i \tag{3.1}$$

onde c é uma constante de proporcionalidade, s_i é a quota do produto da marca i e M_i é o esforço de marketing do produto da empresa i.

Uma vez que as quotas têm de somar 1, se considerarmos que o mercado é constituído por m marcas, então $\sum_{i=1}^{m} s_i = 1$ e, portanto, $\sum_{i=1}^{m} cM_i = 1$. Trabalhando esta igualdade resulta que

$$c = \frac{1}{\sum_{i=1}^{m} M_i}. \tag{3.2}$$

Juntando (3.1) e (3.2) obtém-se a fórmula que define o "Teorema Fundamental de Kotler".

Teorema 3.1 (Teorema Fundamental de Kotler). *Considerando m marcas concorrentes, a quota da marca $i \in \{1, \cdots, m\}$, s_i, é dada pelo quociente entre o esforço de marketing da marca, M_i, e a soma de todos os esforços de marketing.*

$$s_i = \frac{M_i}{\sum_{j=1}^{m} M_j}. \tag{3.3}$$

Dada a sua simplicidade, várias variações desta fórmula surgiram. Consideremos o caso em que duas empresas gastam a mesma quantia em marketing, a participação das duas empresas no mercado não é necessariamente a mesma. Nesta situação é mais correto usar a fórmula:

$$s_i = \frac{\alpha_i M_i}{\sum_{i=1}^{m} \alpha_j M_i},$$

onde α_i representa a eficácia do esforço de marketing da empresa *i*.

Em relação ao esforço de marketing, Kotler assumiu que este é função de um conjunto de variáveis de marketing:

$$M_i = f(X_{ki}),$$

onde *k* é o número de variáveis consideradas.

Num estudo paralelo, Bell *et al.* (1975) consideraram que o fator determinante para a escolha de um consumidor, quando realiza uma compra, é a atração que este sente relativamente a cada uma das marcas. Suponhamos que A_i é a atração do produto da marca *i* e s_i a quota correspondente. Então:

1. $A_i \geq 0$ e $\sum_{j=1}^{m} A_j > 0$;
2. Se $A_i = 0$ então $s_i = 0$;
3. Se $A_i = A_j$ então $s_i = s_j$;
4. Independentemente da marca *i* que sofra variação na atração, a quota das restantes marcas é afetada da mesma forma, qualquer que seja *i*.

Teorema 3.2 (Teorema da Quota de Mercado). *A relação entre a quota da marca i, s_i e a atração das m empresas que constituem o mercado é dada por:*

$$s_i = \frac{A_i}{\sum_{j=1}^{m} A_j}. \tag{3.4}$$

A atração é função das variáveis de marketing e será explorada no Capítulo 6. Se olharmos para as equações (3.3) e (3.4) podemos verificar que, de facto, são muito semelhantes. No entanto, partem de ideias diferentes. Na primeira, a quota é vista em função do esforço de marketing da empresa, já na segunda é considerada a atração do consumidor pela marca (Cooper e Nakanishi, 1988, secção 2.4).

Nos modelos de atração, a quota duma dada empresa varia em função explícita da quota das outras empresas. Isto será um problema para os nossos dados uma vez que os relatórios que nos foram disponibilizados apenas contém dados da quota da empresa em estudo, e apenas da empresa em estudo. No entanto o modelo será explorado no Capítulo 6.

3.3 Modelos em Data Mining

Estes métodos consistem em algoritmos de otimização que estimam os parâmetros do modelo em causa de forma a melhor refletir os dados disponíveis, desta forma descobrindo padrões e tendências. Neste trabalho os métodos em data mining usados são: **regressão linear**, **árvores de regressão**, **redes neuronais** e **máquinas de suporte vetorial**.

Além da regressão linear, é usada uma tranformação da variável a prever, devido às restrições no seu domínio. A função *logit* é muito utilizada para estimar probabilidades uma vez que transforma um domínio $[0, 1]$ em $]-\infty, +\infty[$. Podemos então considerar um modelo linear em que o resultado

será uma transformada:

$$\text{logit}(Y) = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \varepsilon.$$

Portanto, usando a função inversa da logit, conhecida como a **função logística**, isto será o equivalente de usar o modelo:

$$Y = \frac{1}{1 + \exp(-(\beta_0 + \sum_{k=1}^{K} \beta_k X_k + \varepsilon))}.$$

3.4 Modelos de Escolha

Modelos de escolha, também conhecidos por modelos de desagregação, são modelos em que se supõe que existe um processo de decisão racional subjacente às escolhas dos consumidores (Ben-Akiva e Bierlaire, 1999). Racional no sentido de economia que define um agente racional como sendo um agente que consegue ordenar as suas preferências sem conflito (se prefere A a B e prefere B a C, então prefere A a C); estas preferências são modeladas por uma função de utilidade. Supondo então que o consumidor segue uma função de utilidade e que a pretende maximizar, a modelação de escolha tenta encontrar os parametros desta função. Estes modelos tentam, portanto, expor parâmetros de decisão subjacentes às decisões que vemos serem feitas como resposta a fatores exógenos ao consumidor, tais como campanhas promocionais.

É sugerida e implementada uma aplicação desta metodologia para quotas na secção 6.2. Foi implementado um método de Monte Carlo usando a abordagem Bayesiana sugerida em Chen e Yang (2007). Este tipo de modelação tem-se tornado mais popular desde o Nobel em Economia de 2000.

Um problema com este modelos, e vários dos anteriores, é que dispomos apenas da quota de mercado da empresa em relação aos concorrentes, mas não das séries de quota de cada uma das empresas no mercado. Isto dificulta a análise de como os consumidores estão a responder às várias campanhas. Vamos considerar, dada a falta de dados e uma vez que a empresa em causa é a líder do mercado, que as campanhas e a resposta da quota é efectivamente entre duas empresas fictícias: SONAE e não-SONAE.

Capítulo 4

Análise de Séries Temporais

Neste capítulo serão abordados os conceitos de séries temporais e sua aplicação aos dados:

4.1 Processos Estocásticos

Uma **série temporal** é um conjunto de observações de uma variável dispostas sequencialmente no tempo (Brockwell e Davis, 1987), pelo que a variável em estudo é uma série temporal. Para garantir a natureza imprevisível de uma observação futura supõe-se que cada observação y_t é a realização de uma variável aleatória Y_t. A série temporal $\{y_t : t \in T_0\}$ é assim a realização de uma família de variáveis aleatórias $\{Y_t : t \in T\}$. Posto isto, a modelação de dados temporais pode ser feita considerando os dados como a realização de um **processo estocástico** $\{Y_t : t \in T\}$.

Daqui em diante, representaremos por Y_t um processo estocástico.

O objectivo da análise de uma série temporal é compreender o mecanismo gerador da série, de forma a identificar padrões não aleatórios no passado, que permitam a previsão do comportamento futuro da série orientando assim a tomada de decisões.

Em contraposição a séries longitudinais, as séries temporais assumem que existe um processo estocástico por detrás. Interessa-nos portanto estudar propriedades estocásticas como a estacionariedade. A nível de amostragens, considerando que os dados são constituídos por n casos medidos em k ocasiões, habitualmente tem-se que nas séries longitudinais, o n é grande e o k pequeno, enquanto, nas séries temporais, o n é pequeno e o k é grande.

4.1.1 Média, Função Autocovariância, Função Autocorrelação e Variância

Dado um processo estocástico, várias medidas podem ser calculadas de forma a compreender o seu comportamento. Nomeadamente:

Medida	**Descrição**
Média – μ_{Y_t}	$E[Y_t]$
Função autocorrelação – $R_X(s,t)$	$E[Y_s Y_t]$
Função de autocovariância – $K_Y(s,t)$	$Cov[Y_s, Y_t] = E[(Y_s - \mu_{Y_s})(Y_t - \mu_{Y_t})]$
Variância – $\sigma^2_{Y_t}$	$Var[Y_t] = E[(Y_t - \mu_{Y_t})^2] = K_Y(t,t)$

Tabela 4.1: Fórmulas da média, função autocorrelação, função autocovariância e variância.

Quando o processo em estudo é uma série temporal, torna-se relevante perceber de que forma a variável se relaciona com os valores que tomou no passado. Assim sendo, é comum utilizar-se as seguintes medidas: a **função autocovariância (ACVF)** de Y_t com lag (atraso) h, dada por $\gamma_X(h) = K_Y(t, t+h)$ e a **função autocorrelação (ACF)** de Y_t com lag h, dada por $\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$. Além destas, destaca-se a **função de autocorrelação parcial (PACF)** que define a correlação entre as observações Y_t e Y_{t-h} removendo o efeito das observações entre Y_{t-h} e Y_t. Esta remoção é feita determinando Y_t e Y_{t-h} como combinação linear das observações $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-h+1}$ (Cryer e Chan, 2008, secção 6.2).

R A função `acf{stats}` permite estimar as funções autocorrelação e autocovariância. Para determinar a função autocorrelação parcial pode usar-se `pacf{stats}`.
Por omissão, os gráficos resultantes da aplicação das função `acf` e `pacf` são correlogramas.

Aplicando a função ACF às séries temporais da Figura 2.5 na página 8 obtivemos os gráficos de autocorrelação da Figura 4.1. As linhas a azul denotam os valores $\frac{-1}{n} \pm \frac{2}{\sqrt{n}}$ (Cowpertwait e Metcalfe, 2009, secção 2.3).

A análise do gráfico das autocorrelações revelou que nas séries observadas para Portugal Continental, as observações não possuem auto-correlação no geral elevada. Para a unidade 51, o maior valor registou-se no lag 12, ou seja, a quota de um mês qualquer está mais auto corelacionada com a quota registada há 12 meses do que com qualquer outro registo; existirá portanto alguma sazonalidade. No entanto, as autocorrelações são baixas. Onde se verifica maior correlação é para as

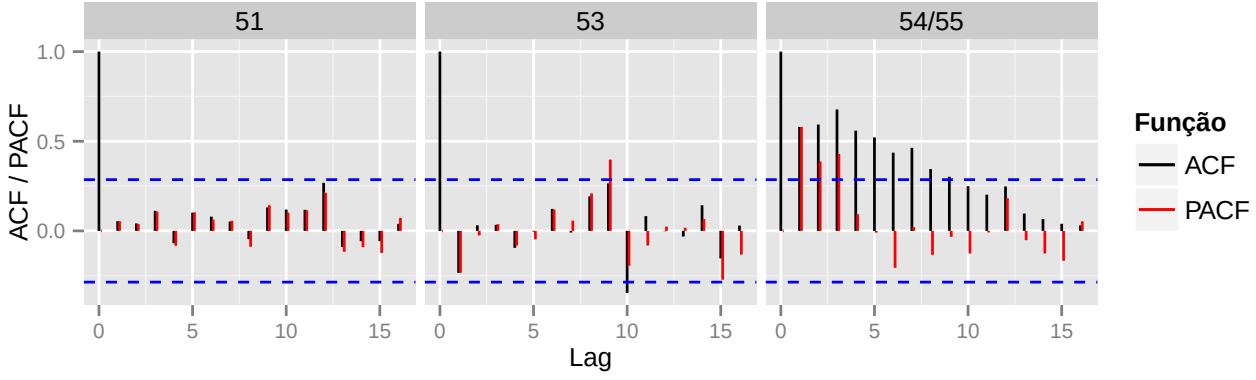


Figura 4.1: Correlogramas da ACF e PACF para as três unidades em Portugal Continental.

unidades 54 e 55, nesta série temporal podemos verificar que até um lag de 7 meses a auto correlação toma valor superior a 0.4 e que o maior valor é obtido para um lag de 3 meses. Isto indica que, para um dado mês, os meses que mais influenciam o seu valor são os 7 meses anteriores. Olhando para o gráfico das autocorrelações parciais podemos verificar que para as duas primeiras unidades a remoção do efeito entre observações não levou a grandes alterações no valor da sua autocorrelação. Para as unidades 54 e 55, a alteração de valores é mais notável havendo uma redução nos valores de autocorrelação entre duas observações.

No resto do capítulo, faremos uma análise do comportamento das série de forma a tentar retirar alguma informação sobre o processo gerador dos seus valores, as quotas.

4.2 Modelo Clássico de Séries Temporais

Além do estudo da autocorrelação com lag k é de especial interesse perceber se o valor de uma série tende a crescer ou decrescer e se existem efeitos periódicos na série. Posto isto, de seguida é feita uma revisão do modelo clássico de decomposição de séries temporais, bem como de dois algoritmos para obtenção de cada uma das suas componentes.

4.2.1 Decomposição em Componentes Básicas

Segundo o modelo clássico, uma série temporal é composta por três componentes básicas a partir das quais poderão ser feitas previsões:

Componente		**Descrição**
T_t	Tendência	Comportamento crescente ou decrescente ao longo do tempo.
S_t	Componente sazonal ou Sazonalidade	Flutuações cíclicas relacionadas com calendário, ocorrem em séries de dados relativas a períodos inferiores a um ano.
I_t	Componente irregular ou erro	Flutuações aleatórias.

Tabela 4.2: Componentes de uma série temporal segundo o modelo clássico.

O objectivo do modelo clássico é decompor uma série analisando cada uma das suas componentes separadamente. O processo de decomposição requere a remoção sistemática de cada uma das componentes.

Salienta-se que, mesmo que o modelo clássico seja o mais adequado, nem todas as séries temporais terão as três componentes acima referidas. Obviamente, as flutuações aleatórias estarão sempre presentes. Esta decomposição pode ser consultada em Brockwell e Richard A. Davis (2002).

Modelo Aditivo vs Modelo Multiplicativo

Dada a decomposição anterior, coloca-se a questão de como combinar as componentes. Esta combinação pode ser feita somando ou multiplicando as componentes não observáveis resultando assim num modelo aditivo ou multiplicativo, respetivamente.

Modelo Aditivo	**Modelo Multiplicativo**
$Y_t = T_t + S_t + I_t$	$Y_t = T_t \cdot S_t \cdot I_t$

Tabela 4.3: Modelo clássico aditivo e multiplicativo.

O modelo multiplicativo deverá ser usado quando a magnitude do padrão de sazonalidade nos dados depende da magnitude dos mesmos, ou seja, a magnitude da sazonalidade cresce quando os valores dos dados crescem e decresce quando estes decrescem.

Por outro lado, o modelo aditivo deverá ser escolhido quando a magnitude da sazonalidade não depende do comportamento dos dados. Neste sentido, um teste simples é a construção de um gráfico da amplitude sazonal em função da tendência.

O modelo multiplicativo pode ser transformado num modelo aditivo através da aplicação da função logaritmo. Para a variável em estudo, o modelo aditivo será o mais adequado.

4.2.2 Estimação da Tendência, Componente Sazonal e Aleatória

A tendência descreve o movimento dos dados ao longo do tempo e, em geral, é a componente mais importante de uma série temporal. Como referido por Granger (1979), a sazonalidade é causada por movimentos oscilatórios periódicos que ocorrem em períodos inferiores a um ano, como feriados, por exemplo. Para a análise de uma série temporal, determinar a diferença entre o que ocorre normalmente e o que ocorre em períodos específicos é bastante importante. Posto isso, a remoção da tendência e da componente sazonal é necessária. À remoção da componente sazonal dá-se a designação de **ajuste sazonal** ou **dessazonalização**. Existem vários métodos para obter estas componentes.

Supondo que existem n observações $\{y_1, ..., y_n\}$, segue-se a explicação de dois deles.

Método 1

Passo 1 Estimar a tendência aplicando um filtro de média móvel especialmente escolhido para eliminar o componente sazonal e amortecer o ruído (Brockwell e Richard A. Davis, 2002, secção 1.5). Seja d o período da série, usa-se:

$$\hat{T}_t = \begin{cases} (0.5y_{t-q} + y_{t-q+1} + ... + y_{t+q-1} + 0.5y_{t+q})/d, & \text{se } d = 2q \\ d^{-1}\sum_{i=1}^{q} Y_{t-j}\,, & \text{se } d = 2q+1, \end{cases}$$

com $q + 1 \leq t \leq n - q$.

Passo 2 Estimar a sazonalidade. Para cada $k = 1, ..., d$ calcular a média w_k dos desvios $\{(y_{k+jd} - \hat{T}_{k+jd}) : q < k + jd \leq n - q\}$.
A sazonalidade será estimada por

$$\hat{S}_k = w_k - \frac{1}{d}\sum_{i=1}^{d} w_i,$$

para $k > d$, $\hat{S}_k = \hat{S}_{k-d}$.

R A função `decompose{stats}` decompõe séries temporais usando médias móveis. Primeiro estima a tendência usando médias móveis (se o filtro é `NULL`, uma janela simétrica com pesos iguais é usada), e de seguida remove-a da série. Depois, a sazonalidade é calculada pela média, para cada unidade de tempo, em todos os períodos. Finalmente, a componente erro é determinada através da remoção de tendência e sazonalidade da série original.

Método 2

Este algoritmo é apresentado em Cleveland *et al.* (1990) e consiste na decomposição de séries temporais baseada no método de regressão LOESS (Cleveland, 1979). O algoritmo conta com dois processos recursivos estando um inserido no outro. Portanto, os processos serão designados por processo interno e externo, respetivamente.

```
1 Para i = 1 até n_(o)
2     Para j = 1 até n_(i)
3         Estimação da tendência e sazonalidade
4     Fim de Para j
5     Estimação da componente aleatória
6     Cálculo dos pesos para estimação robusta
7 Fim de Para i
```

Algoritmo 1: Pseudo-algoritmo STL.

O processo interno conta $n_{(i)}$ iterações sendo feita uma atualização da tendência e sazonalidade.

No processo externo cada iteração consiste numa passagem pelo processo interno seguido do cálculo robusto de pesos que serão usados na próxima execução do ciclo interno de forma a suavizar a influência de pontos com comportamento distinto. Na primeira os pesos são iguais a 1.

Suponha-se que o número de observações em cada período ou ciclo da componente sazonal é $n_{(p)}$. Para dados mensais com periodicidade anual, toma-se $n_{(p)} = 12$. Serão consideradas as sub-séries constituídas pelas observações correspondentes a cada posição do ciclo sazonal. Para o exemplo anterior, existem 12 sub-séries sendo que a primeira é formada por todas as observações de janeiro, a segunda pelas observações de fevereiro e assim sucessivamente.

Processo interno. Suponha-se que $S_v^{(k)}$ e $T_v^{(k)}$, para $v = 1, \ldots, n$, são as componentes sazonal e tendência no final da k-ésima iteração. Veja-se como calcular $S_v^{(k+1)}$ e $T_v^{(k+1)}$:

1. *Remoção da tendência*. Cálculo da série $Y_t - T_v^{(k)}$.

2. *Suavização das subséries*. Todos os pontos de cada sub-série da tendência são suavizados com o método LOESS usando $q = n_{(s)}$ e $d = 1$. O conjunto dos $n + 2n_{(p)}$ valores resultantes da suavização em todas as sub-séries forma a série sazonal $C_v^{(k+1)}$ com $v = -n_{(p)} + 1, \ldots, n + n_{(p)}$.
3. *Aplicação do filtro passa-baixo*. Cálculo de $L_v^{(k+1)}$, $v = 1, \ldots, n$, aplicando a $C_v^{(k+1)}$ um filtro de médias móveis de tamanho $n_{(p)}$, seguido de outro filtro de médias móveis de tamanho $n_{(p)}$, um filtro de médias móveis de tamanho 3 e, finalmente, aplicando uma suavização de LOESS com $d = 1$ e $q = n_{(l)}$.
4. *Cálculo da sazonalidade*. $S_v^{(k+1)} = C_v^{(k+1)} - L_v^{(k+1)}$.
5. *Dessazonalização*. Cálculo da série $Y_v - S_v^{(k+1)}$.
6. *Suavização da tendência*. A série dessazonalizada é suavizada pelo método LOESS com $q = n_{(t)}$ e $d = 1$. A tendência no $(k + 1)$-ésimo passo, $T_v^{(k+1)}$ é dada pelos valores obtidos.

Processo Externo. Suponha-se que da primeira passagem pelo ciclo interno resultam estimativas da tendência e sazonalidade, T_v e S_v. Então a componente aleatória é estimada por $R_v = Y_v - T_v - S_v$.

O próximo passo é determinar pesos para cada observação, pesos que refletem o quão elevado é R_v. Valores pequenos ou até mesmo nulos serão atribuídos a outliers, que resultam em valores elevados de $|R_v|$.

Seja $h = 6mediana(|R_v|)$, a fórmula para o cálculo dos pesos em v é $\rho_v = B(|R_v|/h)$ com

$$B(u) = \begin{cases} (1-u^2)^2, & \text{se } 0 \leq u \leq 1 \\ 0, & \text{caso contrario.} \end{cases}$$

O processo interno é repetido usando na suavização das subséries e da tendência o método de LOESS com uma atualização de pesos obtida multiplicando o peso das vizinhanças de v por ρ_v (Cleveland *et al.*, 1990).

Em relação à decomposição usando médias móveis, este método tem a vantagem de manter o número de observações.

Estimação de Parâmetros

Parâmetro	Descrição	Estimação
$n_{(p)}$	Número de observações por sub ciclo sazonal.	Para dados mensais com periodicidade anual $n_{(p)} = 12$.
$n_{(i)}$, $n_{(o)}$	Número de iterações por cada passagem no ciclo interno e externo, respectivamente.	A estimação robusta é necessária quando o comportamento não Gaussiano da série lida com valores extremos (existência de outliers, por exemplo, são uma evidência). Se tal não acontecer, tomar $n_{(o)} = 0$ e $n_{(i)} = 2$; caso contrário, tomar $n_{(i)} = 2$ e $n_{(o)} = 5$ ou $n_{(o)} = 10$. O valor $n_{(o)}$ pode também ser determinado segundo um critério de convergência que termina quando o número de iterações satisfaz o critério.

Parâmetro	Descrição	Estimação
$n_{(l)}$	Parâmetro da suavização no filtro passa-baixo.	Menor inteiro ímpar tal que $n_{(l)} \geq n_{(p)}$.
$n_{(t)}$	Parâmetro da suavização da tendência.	Menor inteiro ímpar tal que $n_{(t)} \geq \frac{1.5 n_{(p)}}{1-1.5 n_{(s)}^{-1}}$.
$n_{(s)}$	Parâmetro da suavização da sazonalidade.	Inteiro ímpar maior ou igual a 7 de forma a minimizar a distância entre a reta resultante da aplicação da suavização e os valores de s_k.

Tabela 4.4: Parâmetros do algoritmo STL.

R A função `stl{stats}` decompõe séries temporais usando o método STL.

Aplicando os métodos Decompose e STL à série temporal das unidades 54/55, obtêm-se as componentes da Figura 4.2:

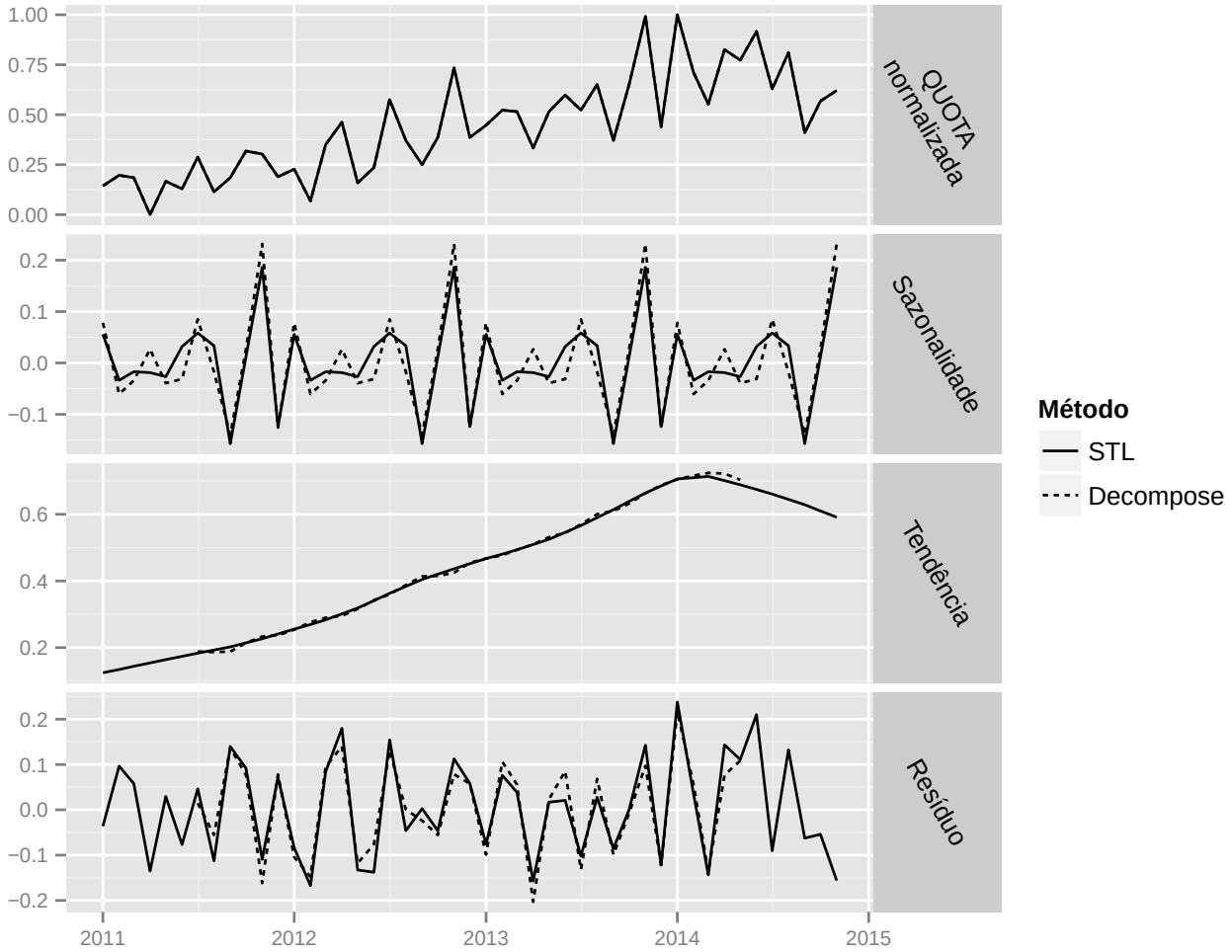


Figura 4.2: Decomposição da série temporal usando os métodos STL e Decompose para as unidades 54 e 55.

De facto, comparando os gráficos da tendência, pode verificar-se que a estimação feita usando o método de LOESS resulta numa curva mais suave do que a determinada usando o método das médias móveis. No entanto, a estimação das variáveis resulta em valores muito semelhantes.

4.3 Processos Estacionários

Para começar, é importante rever o conceito de estacionariedade de um processo visto que existem duas definições que não são equivalentes.

Processo Estritamente Estacionário: O processo Y_t diz-se estritamente estacionário se para todo o n e todo o $t_1, \ldots, t_n$, a distribuição conjunta de $y_{t_1}, \ldots, y_{t_n}$ é igual à distribuição conjunta de $y_{t_1+k}, \ldots, y_{t_n+k}$ qualquer que seja $k \in \mathbb{N}$.

Processo Fracamente Estacionário ou Estacionário de Segunda Ordem: Processo Y_t com variância finita tal que:

- $E[Y_t] = \mu$, μ constante;
- $K_Y(s,t) = K_Y(s+r, t+r), \forall r, s, t$.

Um processo estritamente estacionário é também estacionário de segunda ordem mas o contrário nem sempre se verifica. Daqui em diante, entende-se por processo estacionário, o processo que for fracamente estacionário.

Até agora foram vistas ferramentas para a análise do comportamento de uma série temporal. De seguida será feita a revisão teórica dos modelos mais comuns para modelação e previsão de séries temporais como realização de um processo estocástico, sendo que estes modelos podem ser categorizados nos seguintes grupos:

- AR
- MA
- ARMA

} Processos Estacionários

- AR
- ARMA
- ARIMA

} Processos Não-Estacionários

4.3.1 Processos Auto-Regressivos

A série temporal y_t diz-se ser um processo auto-regressivo de ordem p, denotado por AR(p), se

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-r} + \varepsilon_t, \tag{4.1}$$

onde $\alpha_1, \ldots, \alpha_p$ são constantes fixas ($\alpha_p \neq 0$), designadas por parâmetros de auto-regressão e $\{\varepsilon_t\}$ é uma sequência de variáveis aleatórias independentes com média 0 e variância σ^2, também denominada por ruído branco. A última equação pode ser escrita como um polinómio de ordem p definido à custa do operador B de "backward shift":

$$\theta_p(B) y_t = (1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^p) y_t = \varepsilon_t. \tag{4.2}$$

Os parâmetros deste modelo poderão ser estimados minimizando a soma do quadrado dos erros.

No entanto, estes modelos também podem ser utilizados para modelar processos não estacionários. Por definição, a equação característica é dada pela equação $\theta_p(B) = 0$ onde B é formalmente tratado como um número (real ou complexo). Todas as raízes da equação característica terão de ter valor absoluto superior a 1 para que o processo seja estacionário (Cowpertwait e Metcalfe, 2009, secção 4.5).

4.3.2 Processos de Médias Móveis

A série temporal y_t diz-se um processo de médias móveis de ordem q, denotado por MA(q), se:

$$y_t = \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}, \tag{4.3}$$

onde $\beta_0, ..., \beta_q$ são constantes fixas ($\beta_q \neq 0$), designadas por parâmetros de médias móveis, $\beta_0 = 1$ e $\{\varepsilon_t\}$ é uma sequência de variáveis aleatórias independentes com média 0 e variância σ^2. Tal como no processo auto-regressivo, é possível escrever a equação anterior como um polinómio de ordem q em termos do operador B de "backward shift", obtendo-se:

$$y_t = (1 + \beta_1 B + ... + \beta_q B^q)\varepsilon_t = \phi_q(B)\varepsilon_t. \tag{4.4}$$

Uma vez que os processos de médias móveis são uma soma finita de termos de um ruído branco então são estacionários e, consequentemente, possuem média e auto-covariância independentes do tempo. Facilmente se conclui que a média do processo MA(q) é zero e, visto que os termos são mutuamente independentes, a variância é $\sigma^2(1 + \beta_1^2 + ... + \beta_q^2)$ (Cowpertwait e Metcalfe, 2009, secção 6.3).

4.3.3 Processos Auto-regressivos com Médias Móveis

Uma série temporal $\{y_t\}$ é dita um processo auto-regressivo de ordem p com médias móveis de ordem q, denotado por ARMA(p, q), se

$$y_t - \sum_{i=1}^{p} \alpha_i y_{t-r} = \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}, \tag{4.5}$$

com $\alpha_1, ..., \alpha_p$, $\beta_0, ..., \beta_q$ como constantes ($\alpha_p \neq 0$, $\beta_q \neq 0$), $\beta_0 = 1$ e $\{\varepsilon_t\}$ ruído branco como nos processos anteriormente definidos. Reescrevendo a expressão em termos do operado "backward shift" obtém-se

$$\theta_p(B)y_t = \phi_q(B)\varepsilon_t. \tag{4.6}$$

À semelhança dos processos auto-regressivos, os processos ARMA, também podem ser usados para modelar séries não estacionárias (Cowpertwait e Metcalfe, 2009, secção 6.4). Note-se que:

- O processo é estacionário se todas as raízes de θ têm valor absoluto superior a 1;
- O processo é invertível se todas as raíxes de ϕ têm valor absoluto superior a 1;
- Se $p = 0$ então o processo é MA(q);
- Se $q = 0$ então o processo é AR(p).

Comportamento das funções ACF e PACF de um modelo ARMA

A análise do comportamento das funções ACF e PACF, referidas na secção 4.1.1, sugere quais os parâmetros p e q do modelo ARMA (Shumway e Stoffer, 2011, Tabela 3.1). A tabela seguinte

sumariza como deve ser feita a interpretação dos valores destas funções:

	AR(p)	**MA(q)**	**ARMA(p, q)**
ACF	Decrescimento gradual	Trunca para lag $> q$	Decrescimento gradual
PACF	Trunca para lag $> p$	Decrescimento gradual	Decrescimento gradual

Tabela 4.5: Comportamento das funções ACF e PACF para modelos ARMA.

Considerando o gráfico da Figura 4.1 na página 22, para a unidade 54/55, é notável um decrescimento gradual da magnitude da cauda quer no ACF, quer no PACF. Este comportamento sugere que esta série seja a realização de um processo ARMA. A interpretação das outras unidades é inconclusiva.

4.4 Testes de Hipóteses para Avaliação da Estacionariedade

Na secção anterior foram vistos os três métodos mais comuns para modelar séries temporais consideradas como realização de um processo estacionário. No entanto, na prática nem sempre é fácil perceber se a série observada é ou não estacionária. Posto isto, são apresentados nesta secção os testes mais comuns para testar a estacionariedade de um conjunto de observações. São eles: teste Dickey-Fuller Aumentado (ADF), teste de Phillips-Perron (PP) e teste Kwiatkowski-Phillips-Schmidt-Shin (KPSS). A ideia subjacente a estes testes é a verificação da existência de raíz unitária (se a série tem raíz unitária então não é estacionária).

Para começar, considere-se a série temporal y_t escrita como $y_t = \phi y_{t-1} + \varepsilon_t$ (correspondente a um modelo AR(1)) sujeito a $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$. Neste caso, se $|\phi| < 1$ então y_t é estacionário. A hipótese a testar é então $H_0 : \phi = 1$, ou seja, a série não é estacionária contra a hipótese alternativa $H_1 : |\phi| < 1$.

O mais comum é transformar o modelo de forma a testar se os coeficientes são nulos. Para isso, basta tomar $\Delta y_t = (\phi - 1) y_{t-1} + \varepsilon_t = \pi y_{t-1} + \varepsilon_t$ e a hipótese nula passa a ser $H_0 : \pi = 0$. Esta técnica é usada no modelo Dickey-Fuller modificado. No entanto, algumas séries possuem uma estrutura mais complexa, pelo que um modelo AR(1) nem sempre é suficiente para capturar a existência de raíz unitária.

Teste de Dickey-Fuller Aumentado. É conhecido na literatura como teste ADF (Augmented Dickley-Fuller) e requer o estudo sobre a seguinte regressão:

$$\Delta y_t = \beta_0 + \beta_1 t + \pi y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + \varepsilon_t$$

onde β_0 é a constante de interceção, também denominada como drift da série; β_1 é o coeficiente de tendência; π é o coeficiente de presença de raiz unitária e p é o número de *lags* tomados na série (Gujarati, 2004, secção 21.9).

A estatística de teste é

$$T = \frac{\hat{\pi}}{\text{sd}(\hat{\pi})},$$

onde $\hat{\pi}$ é um estimador para π e, $\text{sd}(\hat{\pi})$ é um estimador para o desvio padrão do erro de π. Os valores

críticos da estatística T foram tabelados por Dickley e Fuller (Gujarati, 2004) através de simulação Monte Carlo e variam nos casos de presença somente de constante de interceção, presença somente de tendência e presença de ambos.

Teste de Phillips-Perron. Este teste é uma generalização do teste de Dickley-Fuller para os casos em que os erros $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ são correlacionados e, possivelmente, heterocedásticos (Philips e Perron, 1988). Neste caso, a estatística Z é calculada por

$$Z = n\hat{\pi} - \frac{n^2 \text{sd}(\hat{\pi})}{2\hat{\sigma}^2}\left(\hat{\lambda}^2 - \hat{\sigma}^2\right),$$

onde $\hat{\sigma}^2$ pode ser estimado pela variância da soma dos quadrados dos resíduos $\hat{u}_t$ e $\hat{\lambda}^2$ pode ser obtida utilizando a estimativa $\frac{1}{n^2}\sum_{t=1}^{n}(\hat{u}_t)^2$.

Teste KPSS. As hipóteses do teste são H_0: a série é estacionária contra H_1: a série apresenta raiz unitária (Kwiatkowski *et al.*, 1992). Suponha-se que é possível decompor a série nas componentes tendência, passeio aleatório e erro:

$$Y_t = T_t + r_t + \varepsilon_t,$$

onde r_t é o passeio aleatório $r_t = r_{t-1} + \mu_t$ com μ_t i.i.d com média zero e variância σ^2_μ.

A estatística do teste é dada por

$$\text{KSPP} = \sum_{t=1}^{n} \frac{\hat{S}_t^2}{n^2\hat{\lambda}^2},$$

onde $\hat{S}_t = \sum_{i=1}^{t}\hat{u}_i$, $\hat{u}_t$ são os resíduos da regressão em x em ordem à tendência e $\hat{\lambda}^2$ é um estimador da variância a longo prazo de u_t usando $\hat{u}_t$. A estatística KSPP tem distribuição que converge assintoticamente para um movimento Browniano cujos valores críticos são tabelados. O movimento Browniano (também conhecido por processo de Weiner) é um processo estocástico W_t em tempo contínuo muito usado para descrever movimentos aleatórios em que: $W_0 = 0$, $\{W_{t_2} - W_{t_1}, \ldots, W_{t_n} - W_{t_{n-1}}\}$ são independentes, $0 \leq t_1 < \ldots < t_n$, $n > 2$, e $W_t - W_s \sim \mathcal{N}(0, t-s)$ para $0 \leq s \leq t$ (Shumway e Stoffer, 2011, secção 5.3).

R As funções `adf.test{tseries}`, `PP.test{stats}` e `kpss.teste{tseries}` permitem relizar os testes de Dickey-Fuller Aumentado, Phillips-Perron e KSPP, respetivamente (Trapletti e Hornik, 2013).

A aplicação destes testes à série temporal da quota em Portugal Continental resultou nos seguintes valor-p:

Teste	**Valor-*p***		
	UN51	UN53	UN54/55
ADF	0.01	< 0.01	0.64
PP	0.01	0.01	0.01
KPSS	< 0.01	0.09	< 0.01

Tabela 4.6: Aplicação dos testes de estacionariedade.

Os testes de Dickey-Fuller Aumentado e de Phillips-Perron testam a hipótese nula de que a série é não estacionária. A análise do valor-p do teste de Phillips-Perron leva à rejeição da hipótese nula,

a um nível de significância de 5%. Por outro lado, no que se refere ao teste ADF podemos verificar que para a unidade 54/55 não há rejeição da hipótese nula. O teste KPSS testa a hipótese nula de estacionariedade, havendo rejeição da mesma, a um nível de significância de 1%, quer para a unidade 51, quer para a unidade 54/55. Neste teste, a exceção é a unidade 53 que mesmo a um nível de 5% não permite rejeitar a hipótese nula de estacionariedade. Com a realização destes testes podemos inferir que as séries poderão não ser estacionárias.

4.5 Processos Não-Estacionários

Para a modelação de processos não estacionários outros modelos podem ser adotados além dos processos auto-regressivos apenas e auto-regressivos de médias móveis. Na literatura de séries temporais são referidos vários mas uma vez que os nossos dados não apresentam uma componente sazonal significativa apenas consideramos um deles.

4.5.1 Processos ARIMA

Um processo auto-regressivo integrado de média móvel é uma generalização dos processos ARMA que incorpora séries não estacionárias (Brockwell e Richard A. Davis, 2002, secção 6.1). Dada uma série temporal y_t, a primeira diferença desta é dada pela diferença entre os valores y_t e y_{t-1}. Deste procedimento resulta a série $\nabla y_t = y_t - y_{t-1}$. Da mesma forma, a segunda diferenciação de y_t é a série temporal dada por $\nabla^2 y_t = \nabla y_t - \nabla y_{t-1}$.

Posto isto, um processo ARIMA é tal que ao fim de um número finito de diferenciações (no sentido exposto anteriormente) é reduzido a um modelo ARMA.

Por definição, a série $\{y_t\}$ é um processo ARIMA(p, d, q) se existe algum d inteiro não negativo tal que o processo

$$\nabla^d y_t = (1 - B)^d y_t$$

é um processo ARMA(p, q). Tal como nos casos anteriores, uma forma mais sucinta é dada por

$$\theta_p(B)(1 - B)^d y_t = \phi_q(B)\varepsilon_t.$$

Se $p = 0$, $d > 0$ e $q > 0$ o processo pode ainda ser designado IMA(d, q). Da mesma forma, se $p > 0$, $d > 0$ e $q = 0$, o processo tem a designação ARI(p, d) (Cryer e Chan, 2008, secção 5.2).

R A função `arima{stats}` permite determinar os coeficientes associados a um processo ARIMA(p, d, q).

Tomando a primeira diferença da série da quota verificou-se que, em quaquer uma das regiões, os testes ADF e PP rejeitam a hipótese nula de não estacionariedade a um nível de significância de 1%. Por outro lado, o teste KPSS revelou valores que não permitiram rejeitar a hipótese de que a série é estacionária. A coerência entre os testes poderá ser um forte índicio de que a primeira diferença da série é estacionária.

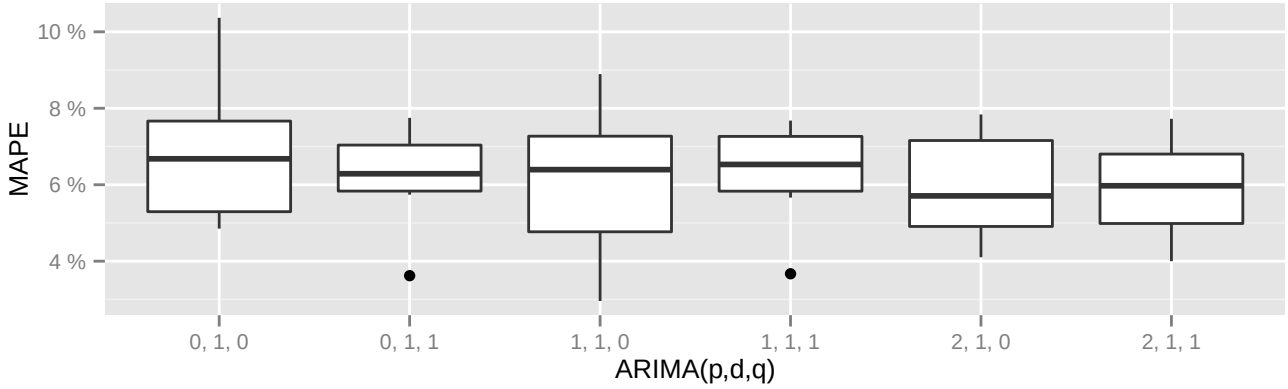


Figura 4.3: Avaliação do erro ARIMA(p, d, q) para as unidades 54/55.

4.6 Escolha do Modelo

A escolha do modelo deve ser feita tendo em conta o teste de estacionariedade e as funções ACF e PACF. Uma série não estacionária deve ser diferenciada até que tal se verifique. As funções ACF e PACF devem ser interpretadas para a série estacionária usando a Tabela 4.5. Para a série em estudo, deverá ser considerada a hipótese de que a série possa não ser estacionária.

Tomando $d = 0$ e analisando as funções ACF e PACF podemos inferir que o valor de p será no máximo 3 e que o valor de q será no máximo 8. Por outro lado, tomando $d = 1$, o valor de p deverá ser no máximo 2 e o de q no máximo 1.

A título ilustrativo mostramos os boxplots com os erros da validação cruzada correspondentes às ordens resultantes das combinações de $d = 1$ com $p \in \{0, 1, 2\}$ e $q = \{0, 1\}$. A validação de janela deslizante foi feita com uma janela de 19 meses com 7 meses de teste e um salto de 4 meses. Foram utilizados os primeiros 40 meses, como anteriormente descrito na secção 2.7 na página 14.

Inicialmente escolhemos a ordem com menor mediana para os erros de validação cruzada. No entanto, interessa-nos que o método escolhido não tenha valores outlier. Assim sendo, decidimos usar a média como medida de escolha do melhor modelo. Posto isto, se as ordens testadas fossem apenas as da Figura 4.3, tomaríamos a ordem $(2, 1, 1)$.

Como veremos adiante, as campanhas promocionais têm impacto no valor da quota mas uma vez que não ocorrem periodicamente estes modelos não conseguem captar o seu impacto. Se as campanhas ocorressem em períodos fixos muito provavelmente a série temporal teria uma componente sazonal mais notável que refletiria esse mesmo impacto.

Capítulo 5

Modelos em Data Mining

Neste capítulo serão aplicados vários modelos tradicionais de estatística e data mining:

Estes modelos serão depois comparados entre si, assim como com o resto dos modelos da dissertação, no Capítulo 7.

5.1 Regressão Linear

Os modelos de regressão linear modelam a relação entre uma variável contínua Y, designada **resposta** ou **variável dependente**, e um conjunto de variáveis $\mathbf{X} = (X_1, \ldots, X_k)$ que podem ser de qualquer tipo e são designadas por **variáveis independentes** ou **explicativas**.

No modelo de regressão linear clássico assume-se que a distribuição condicionada $Y|\mathbf{X} = \mathbf{x}$ tem distribuição normal com média dependente de $\mathbf{x}$:

$$Y|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x})),$$

tal que

$$\mu(\mathbf{x}) = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \beta_0 + \sum_{i=1}^{k} \beta_i x_i.$$

As constantes $\beta_0, \ldots, \beta_k$ são os **parâmetros** ou **coeficientes de regressão**. Posto isto,

$$Y(\mathbf{X} = \mathbf{x}) = \mu(\mathbf{x}) + \varepsilon = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon,$$

onde $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Se $k = 1$, trata-se de uma **regressão linear simples**; caso contrário, **regressão linear múltipla**.

A regressão linear assume que dadas n observações, $\{y_i : x_{i1}, \ldots, x_{ik}\}_{i=1}^{n}$, a relação entre y_i e as k variáveis correspondentes é linear. Pode representar-se de forma matricial como:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

ou de forma mais abreviada:

$$\mathbf{y} = X\beta + \varepsilon.$$

Cada observação y_i deve ser considerada como a realização de uma variável aleatória $Y_i \sim \mathcal{N}(\mu(x_i), \sigma^2(x_i))$. Tal como nos **modelos de probabilidade linear**, na nossa variável de resposta, a quota, existe a restrição **implícita** $0 \leq Y \leq 1$, embora não haja nenhuma restrição vinculativa no contradomínio do modelo.

Aplicação. Tendo em conta os dados apresentados no Capítulo 2, iremos começar por considerar o seguinte modelo completo com interacções entre todas as variáveis que não são exclusivas e considerando ainda as campanhas com lag (atraso) de um mês. O modelo foi construído portanto pelas seguintes variáveis e os seus efeitos de segunda ordem: PROMOk(t), CONCOj(t), PROMOk(t-1), CONCOj(t-1) e REGIAO, $\forall j, k$.

Como explorado no Capítulo 2, as várias variáveis demográficas (como população e poder de compra), assim como as vendas, estão linearmente correlacionadas entre si, e com as regiões. Uma vez que um estimador não será estável ao fazer uso de variáveis explicativas correlacionadas (Gujarati, 2004, Capítulo 10), foi escolhida a variável categórica das regiões. Esta variável diz respeito à região em que a observação foi feita, capturando os efeitos das anteriores sobre a quota e funcionando como um controlo sobre as diferentes médias da quota nas várias regiões.

Obtemos um modelo reduzido a partir deste através dum algoritmo stepwise, em que as variáveis foram sendo retiradas, uma a uma, pela ordem da remoção que mais minimizava o cálculo do AIC, um critério de qualidade de modelos estatísticos (Akaike, 1998). Na Figura 5.3 estão representados os coeficientes com maior valor acima de um certo valor de corte.

Existe ainda um teste de hipóteses para avaliar o quão relevante é uma variável para o modelo de regressão, o **teste-*t*** (Seltman, 2015, Capítulo 9), descrito de seguida: Suponhamos que se pretende testar a hipótese $H_0 : \beta_j = 0, j \in \{0, \cdots, p\}$, o que significa que o coeficiente β_0 não é significativo e que a variável X_j não deve constar no modelo de regressão, então esta hipótese pode ser testada usando a estatística de teste

$$T_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}.$$

Sob H_0, $T_j \overset{a}{\sim} t(n - (p+1))$.

R A função `lm{stats}` permite estimar um modelo linear usando um estimador OLS (mínimos quadrados). Os valor-p associados ao teste-t podem ser obtidos usando o comando `summary.lm{stats}`.

Uma vez que dispomos de ferramentas estatísticas para escolher o modelo de regressão linear mais significativo (algoritmo stepwise e teste-t) não necessitámos de recorrer ao método k-fold, um modelo de regressão linear ilustrativo pode ser construído usando cerca de 70% da amostra e o seu erro pode ser estimado recorrendo aos restantes 30%.

A análise gráfica dos erros (Figura 5.1) permite verificar que a unidade 54 e 55 tem, em geral, um erro inferior às restantes. Apesar da previsão usando este modelo ter dado valores entre 0 e 1 isto podia não ter acontecido pois não existe garantia que o contra-domínio do modelo seja $[0, 1]$. O facto das previsões não terem saído do intervalo está relacionado com os valores que a variável tomou no histórico pois estes nunca foram muito próximos de 0 ou 1.

Outra possível desvantagem destes modelos é que os coeficientes representam efeitos absolutos na variável dependente, o que pode não ser desejável. Um incremento em 1 duma variável dependente X_i tem sempre um efeito β_i, independentemente do valor de X_i. Se houver retornos diminuídos (uma campanha tem um efeito maior na quota quando a quota é baixa do que quando esta é alta), pode fazer mais sentido observar o efeito de um aumento duma variável independente em unidades relativas como na função logística que será vista de seguida, e não em absolutas como no modelo linear (Faraway, 2006).

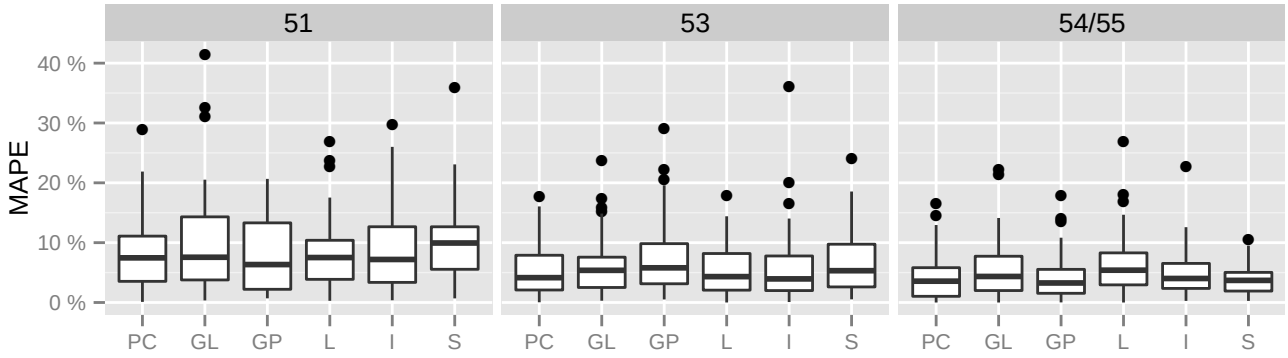


Figura 5.1: Comparação dos erros estimados para a regressão linear, por região e unidade.

5.2 Modelo Logístico

De forma a garantir a restrição no domínio da variável a prever, consideramos a modelação da quota através de um modelo não linear, de forma que

$$Y = \frac{1}{1 + \exp^{-(\beta_0 + ... + \beta_k X_k)}}.$$

Esta função é a função logística e apresenta a forma de S. Adotando este modelo, estamos a assumir que a quota segue uma função logística, ou seja, assumimos que para valores baixos da função preditora, a quota será próxima de zero e que aumenta à medida que o preditor aumenta.

Relativamente à estimação dos parâmetros, uma vez que se trata da função logística, podemos inverter a expressão de forma a poder utilizar o método OLS para estimar os parâmetros:

$$\text{logit}(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k,$$

sendo logit(Y) = log $\frac{Y}{1-Y}$.

Neste caso, os parâmetros têm um significado diferente. O coeficiente β_i é interpretado como a alteração que o aumento de uma unidade na variável *i* produz no logaritmo dos odds da quota. No caso das nossas variáveis, que tomam valor 0 ou 1 conforme haja ou não campanha, o coeficiente β_i é visto como o impacto que a existência da campanha tem sobre o logit da quota.

Procedendo de forma análoga à regressão linear, estimamos os erros médios cometidos pelo modelo de cada região:

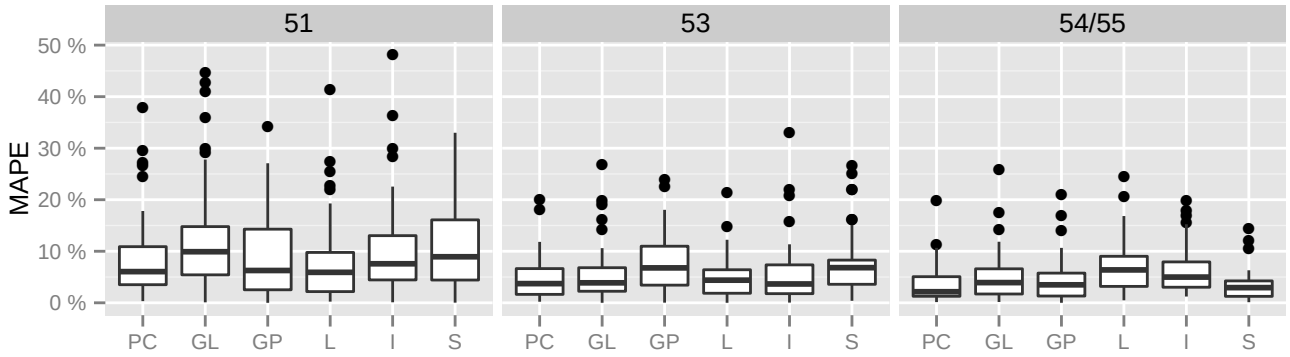


Figura 5.2: Comparação dos erros estimados para a regressão logistica, por região e unidade.

5.2.1 Análise dos coeficientes para regressão linear e modelo logit

Consideremos os modelos construídos utilizando os dados de todas as regiões com todas as variáveis e os seus efeitos de segunda ordem: PROMOk(t), CONCOj(t), PROMOk(t-1), CONCOj(t-1) e REGIAO, $\forall j, k$. Vamos assumir que os modelos resultantes do algoritmo stepwise têm todos os coeficientes significativos. Apenas são representados coeficientes com valor absolutos superior a um dado valor de corte que entendemos ser interessante.

Comecemos pela análise da Figura 5.3 na próxima página. Para a unidade 51, a interação entre a PROMO4(t) e a CONCO2(t) tem um coeficiente muito semelhante ao da interação entre a CONCO2(t) e a PROMO1($t-1$) mas com sinal oposto, ou seja, o modelo estima que se todas as outras variáveis se mantiverem fixas, a ocorrência em simultâneo da PROMO4 e CONCO2 leva a uma subida na quota e que se houver uma PROMO1 no mês t-1 seguida de uma CONCO2 no mês t então a quota vai sofrer uma descida. Pode observar-se ainda que os valores absolutos dos coeficientes têm valores bastantes próximos. Algo que não seria de esperar era um sinal negativo para a PROMO4; uma razão para este valor poderá ser o facto desta campanha ocorrer quase sempre em simultâneo com alguma da concorrência levando a que o modelo estime uma descida quando esta ocorre.

Para a unidade de negócio 53, verifica-se uma maior variação entre os valores absolutos dos coeficientes. De observar que o coeficiente com maior valor está associado à ocorrência, quando ocorrem as campanhas PROMO3 e CONCO1 no mesmo mês, logo seguido da ocorrência de CONCO2 tendo havido no mês anterior PROMO1. Nesta unidade de negócio, a PROMO4 surge com coeficiente positivo; no entanto, a PROMO3 surge com valor associado negativo. Por último, na unidade 54 e 55, os coeficientes tomam valores inferiores pelo que, para o valor de corte escolhido, apenas três são representados. Os três coeficientes têm valores muito semelhantes. Uma nota a esta análise é

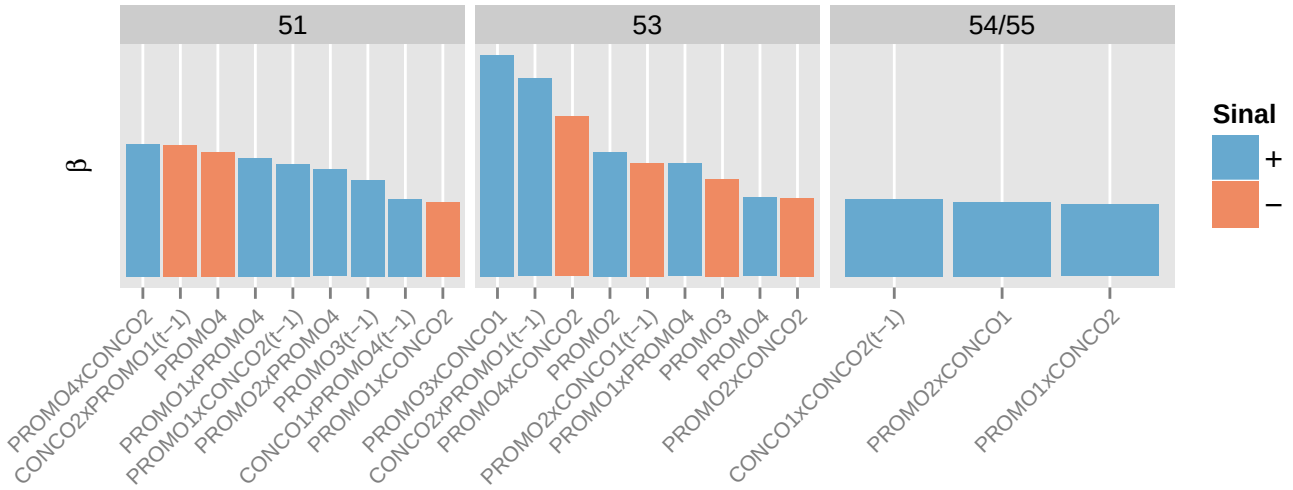


Figura 5.3: Principais betas para a regressão linear.

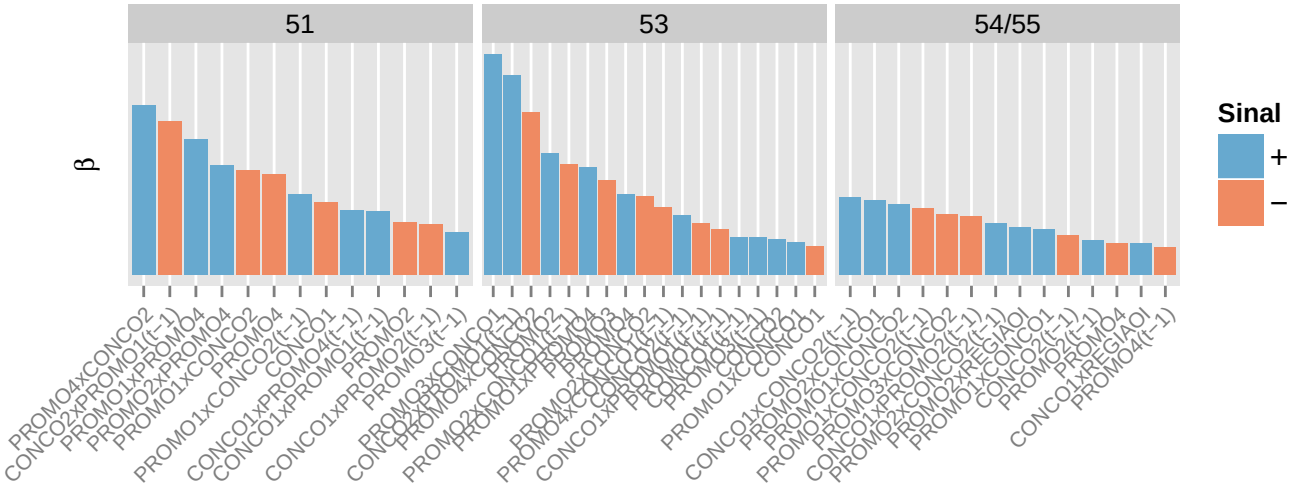


Figura 5.4: Principais betas para o modelo.

o facto de que a mesma campanha promocional ou interação entre campanhas tem diferentes impactos nas diferentes unidades de negócio, não se mantendo a ordem dos coeficientes. Um exemplo evidente é a ocorrência num mesmo mês das campanhas promocionais PROMO4 e CONCO2 que para a unidade 51 tem coeficiente associado positivo e para a unidade 53 tem valor ligeiramente superior es com sinal negativo.

Passando para a análise da Figura 5.4, podemos verificar um decréscimo mais acentuado entre os valores dos coeficientes. Isto deve-se ao facto dos coeficientes refletirem o incremento que a existência de uma campanha promocional provoca no logit da quota e não na quota. Numa análise muito geral podemos verificar que, nas unidades 53 e 54/55, a ordem dos coeficientes apresentados na regressão linear se manteve. Para a unidade 51, os dois primeiros coeficientes correspondem aos da regressão linear, embora depois a ordem seja diferente.

Relação Causa-Efeito. É importante fazer notar que fazer estatística com dados humanos (um agente racional) é algo muito diferente da maioria das aplicações estastícas que observam fenómenos não-humanos (agentes passivos). O problema é que, quando se trata de seres humanos, o observado é também o observador, de forma que estabelecer uma relação causa-efeito pode ser complicado, e uma regressão assume que a variável independente é, de facto, independente. É

perfeitamente plausível que a decisão de fazer promoções advenha duma redução da quota baixa. De facto, se este estudo tiver algum efeito a nivel de decisão, então o estudo será automaticamente invalidado a partir desse ponto no tempo.

5.3 Árvores de Regressão

Árvores de classificação e regressão são mais um método de aprendizagem para construção de modelos de previsão a partir de dados. Estes modelos são obtidos através da partição sucessiva do espaço de dados com estimação de modelos simples em cada partição. O resultado pode, desta forma, ser representado como uma árvore.

As árvores de classificação foram desenvolvidas para variáveis dependentes que tomam um número finito de valores e cujo erro de previsão é medido através de uma função custo. Por outro lado, as árvores de regressão foram pensadas para variáveis dependentes que tomam valor contínuo ou discreto ordinal. Neste caso, a medida mais usada é o quadrado das diferenças entre o valor previsto e o valor real.

O algoritmo CART – Classification and Regression Trees (Hand *et al.*, 2000, secção 5.2) – é o mais comum como metodologia de regressão não paramétrica para a previsão do desempenho de uma variável. As Árvores de Regressão CART são essencialmente usadas para explicar e prever uma determinada variável a partir de valores observados de variáveis explicativas da mesma. Este método permite ainda construir grupos homogéneos de indivíduos que são caracterizados pelos mesmos valores dos atributos. A metodologia de regressão CART pressupõe três etapas:

- Crescimento da árvore procedendo a diversas ramificações binárias no sentido de diminuir a diversidade da variável em estudo;
- Validação da árvore;
- Interpretação da árvore resultante.

Crescimento da Árvore. A árvore de regressão CART é obtida a partir de sucessivas divisões binárias do conjunto de dados (usando a amostra de treino) através de uma medida de homogeneidade que é usada para decidir qual a melhor variável de corte e valor de corte associados a cada nó. Cada nó é dividido em dois nós descendentes, de forma a reduzir a heterogeneidade dos valores da variável dependente nestes nós relativamente ao nó ascendente. Em cada divisão é avaliada a redução da variância da variável a prever de forma a definir a melhor variável de corte.

Todo este processo é recursivo, cada nova ramificação origina uma árvore com menor variabilidade do que a árvore que a antecedia. No entanto, o crescimento da árvore pode ajustar-se demasiado aos valores da amostra de treino, o que pode causar algumas dificuldades na generalização do modelo obtido. Assim sendo, é comum definirem-se regras de paragem de crescimento da árvore. Algumas regras de paragem do crescimento de uma árvore são a consideração de um número máximo de níveis, a definição dos números mínimos de observações para nós a ramificar ou para nós descendentes e a imposição de um decréscimo mínimo da diversidade.

Terminada a construção da árvore, a previsão associada a um elemento que foi encaminhado para determinado nó folha será dada pela média no nó-folha onde esse elemento se enquadra (uma previsão que é igual para todos os elementos que pertençam ao mesmo nó- folha).

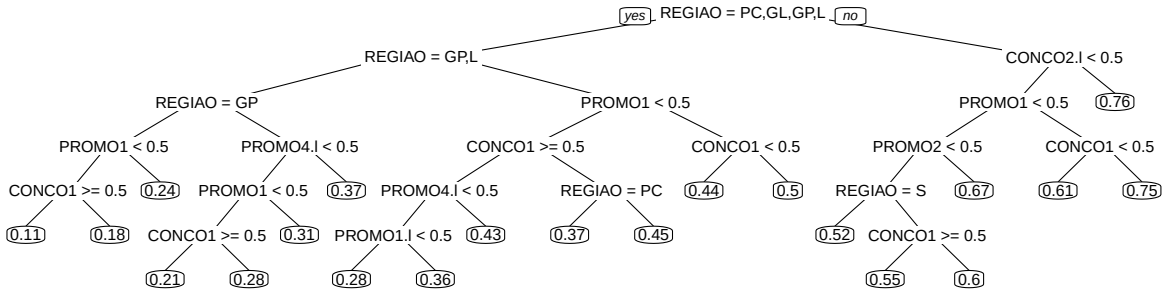


Figura 5.5: Árvore de regressão da unidade 51 para a quota normalizada, usando os dados de todas as regiões.

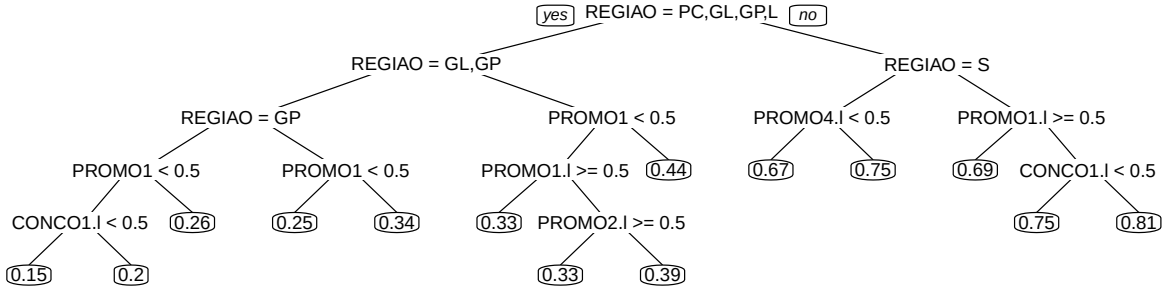


Figura 5.6: Árvore de regressão da unidade 53 para a quota normalizada, usando os dados de todas as regiões.

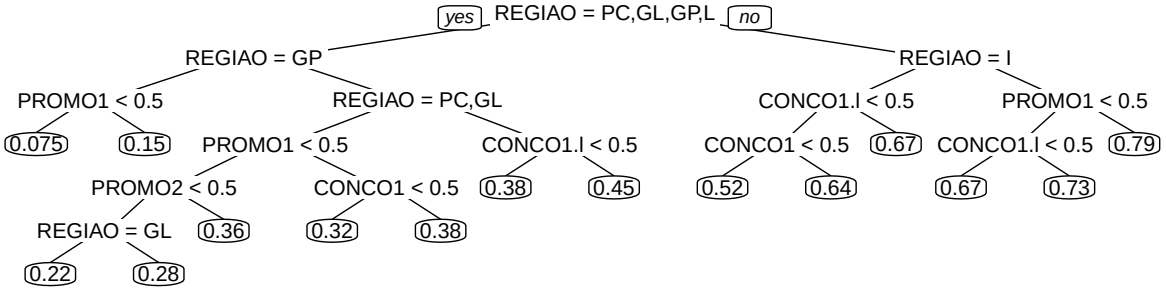


Figura 5.7: Árvore de regressão da unidade 54 e 55 para a quota normalizada, usando os dados de todas as regiões.

Aplicando este método ao conjunto de dados de todas as regiões e considerando as promoções PROMOk(t), PROMOk($t-1$), CONCO1(t) e CONCO2(t), como variáveis explicativas obtivemos as seguintes árvores nas Figuras 5.5 - 5.7.

Uma das grandes vantagens das árvores de regressão é a sua fácil interpretabilidade, algo particularmente útil num ambiente empresarial como a SONAE em que métodos de data mining são em geral desconhecidos. Claro que tomar decisões com base nestes valores sofre dos problemas referidos na primeira secção do capítulo. A análise das árvores obtidas permite mais uma vez concluir que as variáveis terão diferente impacto em diferentes regiões. Se fixarmos, por exemplo, a região Sul podemos verificar que para a unidade 51 as variáveis tidas em consideração são CONCO2($t-1$), PROMO1(t) e a PROMO2(t); para a unidade 53 apenas se considera a PROMO4($t-1$) e para as unidades 54 e 55 a mesma análise revela que as variáveis utilizadas são a PROMO1(t) e CONCO1($t-1$).

Para cada região foi construído um modelo usando as promoções e promoções com lag, para cada unidade de negócio. Tal como nos modelos de regressão linear, uma vez que apenas temos um modelo a avaliar por região, 70% da amostra foi usada para atreinar o modelo e 30% para estimar o erro do mesmo. O resultados encontram-se na Figura 5.8.

Como critério de paragem de crescimento da árvore utilizamos um valor de corte de 0.001, ou

seja, se a média da quota numa nova folha não decresce pelo menos 0.1% em relação ao nó que a origina então a árvore é podada nesse nó.

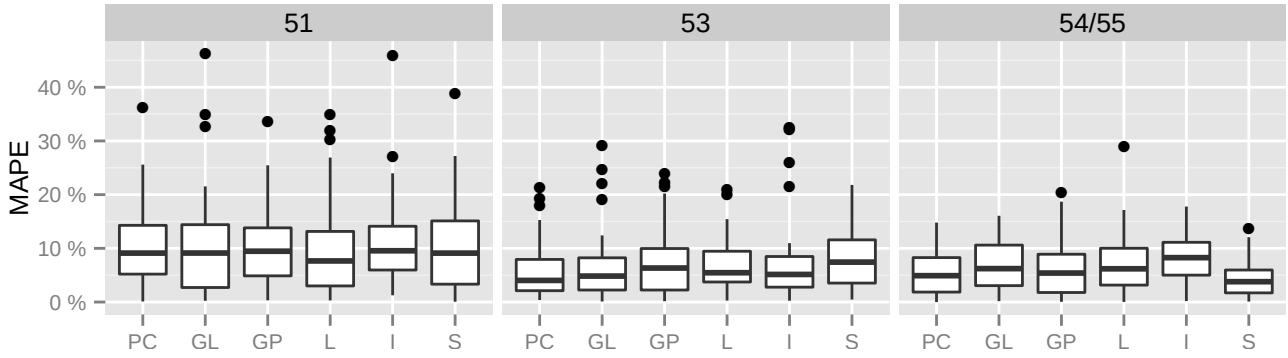


Figura 5.8: Comparação dos erros estimados pelas árvores de regressão, por região e unidade.

5.4 Redes Neuronais

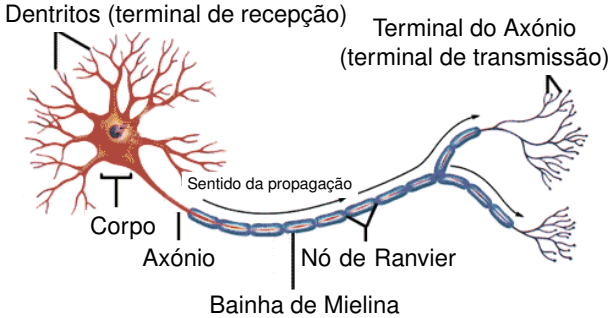
As redes neuronais consistem num método de aprendizagem inspirado pela capacidade de aprendizagem, bem como reconhecimento de padrões, por parte do sistema nervoso central de um ser vivo.

Motivação Biológica – Funcionamento do Sistema Nervoso

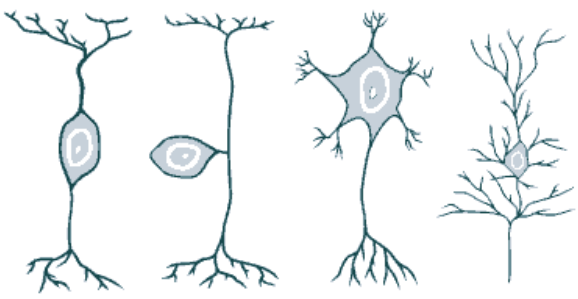
O sistema nervoso detecta estímulos internos e externos desencadeando repostas quer a nível dos músculos, quer a nível das glândulas e é formado, essencialmente, por células nervosas que se comunicam através de **sinapses**, formando as redes neuronais.

O **neurónio** (ou célula nervosa) é o principal componente do sistema nervoso estimando-se que o cérebro humano seja constituído por cerca de 86 bilhões de neurónios. Existem vários tipos de neurónios com diferentes funções e estruturas morfológicas.

As sinapses são regiões através das quais ocorrem os processos de comunicação entre neurónios (a transmissão do sinal neural ocorre devido a processos electroquímicos específicos). As sinapses estão presentes não apenas entre uma terminação nervosa e um neurónio mas também entre a terminação nervosa e células musculares ou glandulares.



(a) Componentes de um neurónio.



(b) Morfologias básicas de neurónios.

Figura 5.9: Estrutura de um neurónio.

Durante este processo de transmissão de informação os neurónios não se tocam, permanecendo um espaço entre eles denominado de fenda sináptica, onde um neurónio pré-sináptico se liga a outro (neurónio pós-sináptico). O sinal ou impulso nervoso, transmitido pelo axónio da célula pré-sináptica chega à extremidade e estimula a libertação de neurotransmissores na fenda (estes neurotransmissores encontram-se armazenados em bolsas chamadas vesículas sinápticas). Este elemento químico liga-se a receptores específicos no neurónio pós-sináptico, dando continuidade à propagação do sinal. Em termos funcionais, o que acontece é que os neurónios sensoriais recebem informação e enviam-na para o sistema nervoso central onde os neurónios de associação a codificam, compararam, guardam e tomam uma decisão que é depois enviada, desencadeando alguma reacção por parte dos músculos ou glândulas.

Redes Neuronais Artificiais

Numa rede neuronal artificial várias camadas de unidades de processamento estão ligadas às anteriores na forma de uma rede, tal como ilustrado na Figura 5.10.

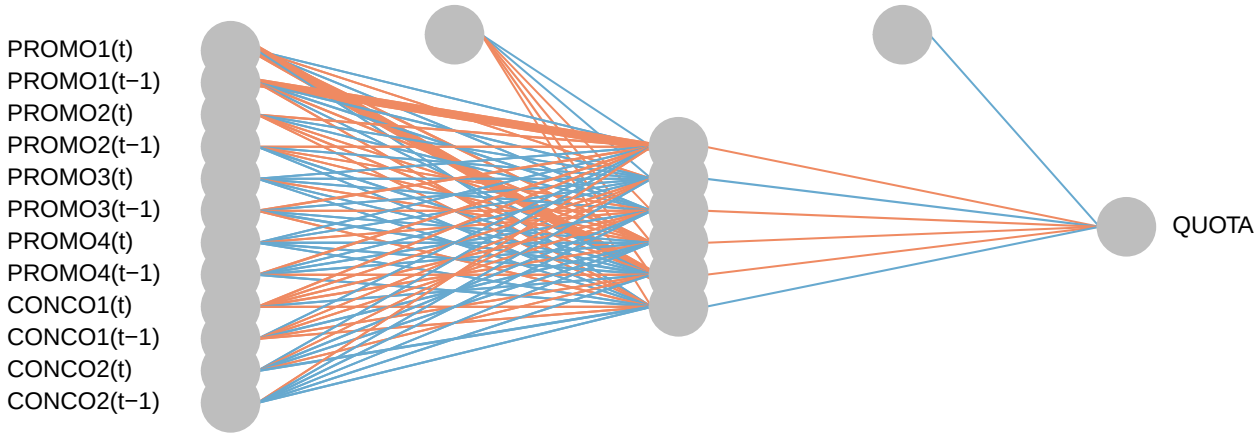


Figura 5.10: Esquema duma rede neuronal com uma camada escondida.

A camada inicial é o vetor **x** de valores de entrada x_i a partir do qual queremos construir a nossa previsão y. Cada camada é ligada à anterior por um vetor de pesos que é calculado a partir da camada anterior a por uma função $f(a, \mathbf{W})$, onde **W** é a matriz de pesos e f é designada função de ativação.

Seguem dois exemplos de funções de ativação e respetivos contradomínios:

Função Ativação	Fórmula	Contradomínio
linear	x	$]-\infty, +\infty[$
sigmóide	$\frac{1}{1+e^{(-x)}}$	$[0, 1]$

Tabela 5.1: Duas funções comuns de ativação de redes neuronais.

Para um determinado problema, o número de camadas escondidas (camadas de neurónios intermédios), o número de unidades em cada uma dessas camadas e os pesos de conexão dependem do conjunto de treino. O algoritmo de aprendizagem supervisionada mais popular é o algoritmo de retropropagação (ou *back-propapagation*) do erro (Almeida, 1997). Neste algoritmo, os dados

recolhidos pela camada de entrada são propagados até à camada de saída. A saída resultante é comparada com a saída desejada para avaliar o erro, que é retro-propagado ao longo das várias camadas. Cada unidade na camada escondida recebe uma proporção do erro, proporcional à contribuição relativa dessa unidade na construção da saída. Os pesos e polarizações são atualizados progressivamente com o intuito de minimizar a soma do quadrado dos erros, podendo ser usado o método do gradiente nessa minimização. No entanto, existem algumas limitações pois o método do gradiente possui uma velocidade de aprendizagem reduzida o que, consequentemente, exige um tempo computacional significativo. Além disso, a convergência para o mínimo global não é garantida.

Como mostra o teorema de Hornik (1991), uma camada intermédia é suficiente para modelar qualquer função contínua por partes, pelo que nomalmente apenas se considera uma camada intermédia.

Rede Neuronal *Feed-Forward* com uma Camada Escondida

Como refere Faraway (2006), este algoritmo usa a seguinte forma:

$$y = f_o\left(\sum_h w_{ho} f_h\left(\sum_i w_{ih} x_i\right)\right),$$

onde f_o é a função activação da saída e f_h a função de activação na camada escondida.

O mais comum é tomar f_h como a função logística. Por outro lado, a escolha da função de activação da saída depende da natureza da variável resposta. Para uma resposta contínua sem restrições, a função identidade é apropriada, mas se por exemplo a resposta varia entre 0 e 1, a função logística pode ser usada. Os pesos W são obtidos de forma a minimizar algum critério, como por exemplo:

$$\text{Erro} = \sum_i (y_i - \hat{y}_i)^2,$$

onde y é o valor observado e $\hat{y}$ o valor estimado.

Uma observação ao modelo das redes neuronais artificiais é o facto de poderem ser vistas como uma extensão do modelo linear generalizado, permitindo interações mais sofisticadas entre as variáveis através do uso de uma camada intermédia. A principal desvantagem no uso de uma rede neuronal é a dificuldade na interpretação do modelo resultante, daí ser muitas vezes designado por **caixa negra**.

R A função `neuralnet{neuralnet}` permite a construção de uma rede neuronal com uma (ou mais) camada(s) escondida(s). A otimização é feita recorrendo ao método quasi-Newton "BFGS" (Broyden–Fletcher–Goldfarb–Shanno). Este pacote é mais recente e fornece mais funcionalidades que o antigo `nnet` (Fritsch *et al.*, 2012).

Na sua aplicação, o primeiro passo é determinar qual o número de neurónios a usar na camada intermédia. Se usarmos poucos neurónios, corremos o risco da rede ser incapaz de modelar dados mais complexos, resultando numa baixa capacidade de generalização ("underfitting"). Por outro lado, se usarmos um número elevado de neurónios, o treino da rede pode tornar-se muito longo e a rede pode sofrer problemas de sobreajuste ("overfitting") resultando na perda da capacidade preditiva da

rede, pois observa-se pequenos desvios de previsão para os dados usados na fase de treino, mas grandes desvios quando novos dados de entrada são utilizados.

A determinação do melhor número de neurónios na camada intermédia para modelar cada combinação região/unidade é feita recorrendo ao método de validação cruzada $k-$ fold. Os erros obtidos no $k-$fold, pelo melhor número de neurónios na camada intermédia, são representados na Figura 5.11.

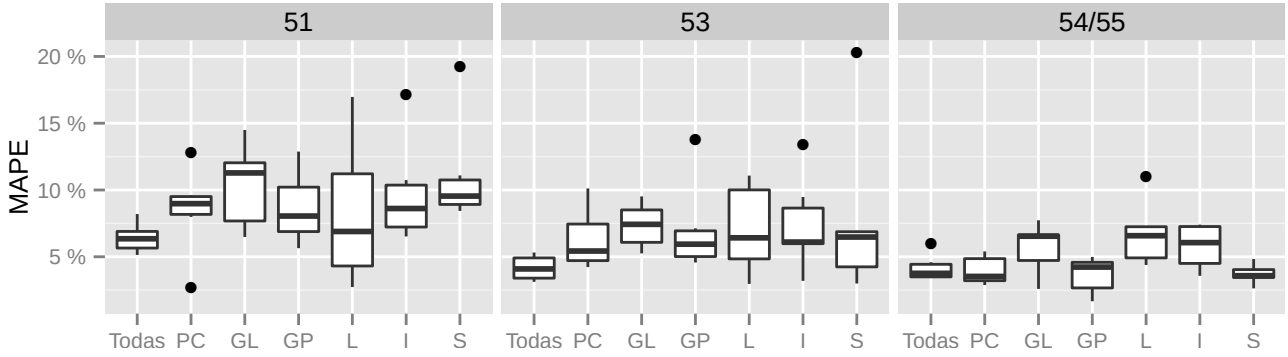


Figura 5.11: Comparação dos erros estimados para redes neuronais, usando k-fold.

Em relação ao número de neurónios a usar para modelar a unidade 51 no Grande Porto, por exemplo, obtivemos o valor 7.

5.5 Máquinas de Suporte Vectorial

As primeiras máquinas de suporte vectorial tinham como objetivo a separação de duas classes. A abordagem consiste em determinar o hiperplano que separa de forma óptima as duas classes maximizando a margem entre os pontos mais próximos das duas classes como mostra a Figura 5.12.

Os pontos que estão sobre os limites são denominados vectores de suporte e o meio da margem resultante é o hiperplano de separação ótimo. Várias extensões deste modelo surgiram. Para este trabalho interessa a extensão das máquinas de suporte vetorial a problemas de regressão.

Considere-se o conjunto de dados de treino $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\} \subset \mathbb{R}^k \times \mathbb{R}$, onde k é o número de variáveis de entrada. O objetivo será determinar uma função $f(x)$ cujo desvio, em relação a todas

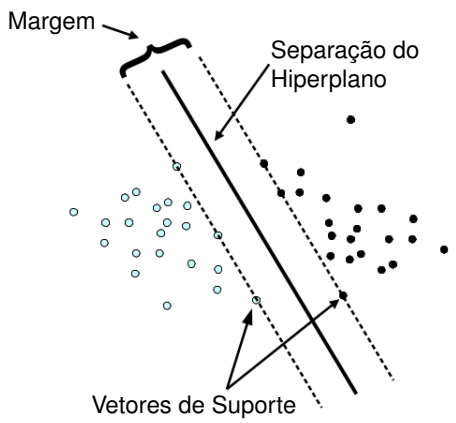


Figura 5.12: Classificação (separação linear).

as variáveis de output y_i no treino, seja no máximo ε.

Começando pelo caso mais simples, considere-se a função f como sendo a função linear dada por

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \ \mathbf{w} \in \mathbb{R}^d, \ b \in \mathbb{R},$$

onde $\langle \cdot, \cdot \rangle$ representa o produto escalar.

Neste caso procura-se determinar um pequeno valor de w. Isto pode ser feito minimizando a norma, $\|\mathbf{w}\|^2 = \langle \mathbf{w}, \mathbf{w} \rangle$ (Smola *et al.*, 2004). Este problema pode ser escrito à custa do problema de otimização convexa dado por:

$$\begin{array}{ll} \min\limits_{\mathbf{w}} & \frac{1}{2} \|\mathbf{w}\|^2, \\ \text{sujeito a} & y_i - \langle \mathbf{w}, \mathbf{x_i} \rangle - b \leq \varepsilon, \\ & \langle \mathbf{w}, \mathbf{x_i} \rangle + b - y_i \leq \varepsilon. \end{array}$$

Analogamente, se forem consideradas margens suaves, são introduzidas variáveis de folga $\boldsymbol{\xi}$ e $\boldsymbol{\xi}^*$. Neste caso, o problema de otimização é dado por:

$$\begin{array}{ll} \min\limits_{\mathbf{w}} & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \\ \text{sujeito a} & y_i - \langle \mathbf{w}, \mathbf{x_i} \rangle - b \leq \varepsilon + \xi_i, \\ & \langle \mathbf{w}, \mathbf{x_i} \rangle + b - y_i \leq \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0. \end{array}$$

Têm-se assim as funções a otimizar bem como as respetivas restrições, quer para o caso em que se consideram fronteiras rígidas, quer para o caso em que se consideram fronteiras suaves. No entanto, na maioria dos casos, estes problemas são resolvidos mais facilmente considerando a sua representação dual.

O dual correspondente ao último problema é dado por:

$$\begin{array}{ll} \max\limits_{\alpha, \alpha^*} & -\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)^\intercal (\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle - \varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \\ \text{sujeito a} & 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, l, \\ & \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0. \end{array}$$

Daqui resulta que $w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i$ e, portanto, $f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b$.

Este processo pode ser generalizado, permitindo o uso de funções f não lineares (Smola *et al.*, 2004). Considere-se C o limite superior e Q uma matriz semi-definida positiva, $Q_{ij} = y_i y_j K(x_i, x_j)$. K é designada a função núcleo. Exemplificam-se algumas funções núcleo na Tabela 5.2.

Função Núcleo	**Fórmula**	**Parâmetros**
Linear	$\mathbf{u}^\intercal \mathbf{v}$	–
Polinomial	$\gamma(\mathbf{u}^\intercal \mathbf{v} + c_0)^d$	γ, d, c_0
Radial	$\exp\{\gamma(\lvert \mathbf{u} - \mathbf{v} \rvert^2\}$	γ, d, c_0
Sigmóide	$\tanh\{\gamma \mathbf{u}^\intercal \mathbf{v} + c_0\}$	γ, c_0

Tabela 5.2: Funções núcleo K de máquinas de suporte vetorial.

Uma vez que a variável a prever é a quota, que pode ser representada como uma proporção, a função núcleo a usar será a sigmóide.

Conforme se tomem fronteiras rígidas ou suaves, a representação dual é, respetivamente uma das descritas de seguida: regressão ε ou ϑ.

Regressão – ε

A representação dual do modelo é:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^\intercal \mathbf{Q}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i-\alpha_i^*) \\ \text{sujeito a} & 0 \leq \alpha_i,\ \alpha_i^* \leq C,\ i = 1,\ldots,l, \\ & \sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0. \end{array}$$

Regressão – ϑ

A representação dual do modelo é:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^\intercal \mathbf{Q}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + \mathbf{z}^\intercal(\alpha_i-\alpha_i^*) \\ \text{sujeito a} & 0 \leq \alpha_i,\ \alpha_i^* \leq C,\ i = 1,\ldots,l, \\ & \sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0, \\ & \sum_{i=1}^{l}(\alpha_i+\alpha_i^*) = C\vartheta. \end{array}$$

R A função `svm{e1071}` permite estimar os parâmetros do modelo (Meyer, 2014). Pode ser utilizada quer para problemas de classificação, quer para problemas de regressão (ε e ϑ) e tem disponíveis as quatro funções núcleo descritas anteriormente.

Para modelar cada um dos conjuntos de dados, correspondentes às combinações região/unidade, utilizamos máquinas de suporte vetorial de margem rígida. A função núcleo utilizada foi a sigmóide pelo que foi necessário recorrer a validação cruzada para escolher os melhores parâmetros.

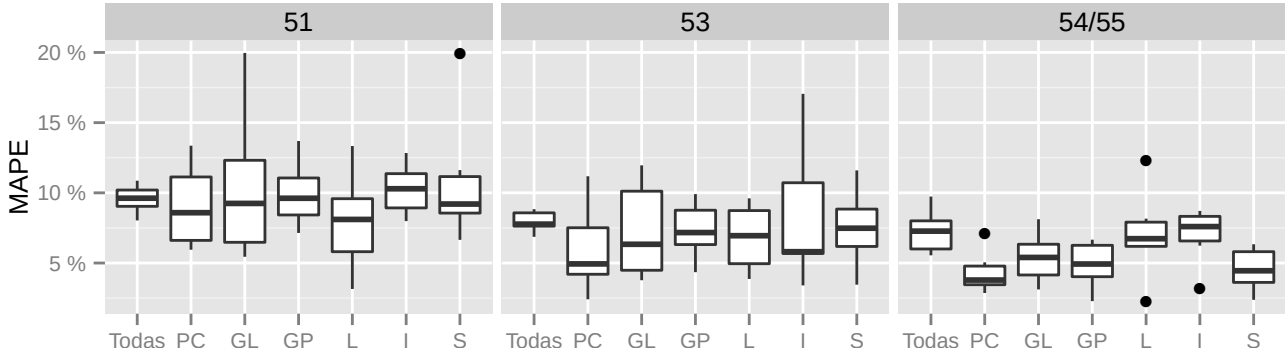


Figura 5.13: Comparação dos erros estimados para máquinas de suporte vetorial, usando k-fold.

Capítulo 6

Outros Modelos

Anteriormente estudamos a aplicação de modelos mais tradicionais de estatística e data mining. Entraremos em detalhe em algumas abordagens mais esotéricas discutidas no Estado da Arte (Capítulo 3):

6.1 Modelo de Atração

Como visto no Capítulo 3, o modelo de atração pressupõe que

$$s_{it} = \frac{A_{it}}{\sum_{j=1}^{m} A_{jt}},$$

onde s_{it} e A_{it} são, respetivamente, a quota e a atração da marca $i \in \{1, \ldots, m\}$, no instante t.

Comecemos com o caso mais simples: assumir que a atração da marca i depende apenas das suas ações, ou seja, o marketing por parte das outras marcas não tem efeito na atração da marca i, o que nem sempre é razoável.

Um modelo MCI (Multiplicative Competitive Interaction) é tal que:

$$A_{it} = \prod_{k=1}^{K} f_k(x_{kit})^{\beta_k},$$

onde x_{kit} é o valor da k-ésima variável explicativa da marca i no instante t, f_k é uma transformação monótona em x_{kt} e β_k são os parâmetros a estimar. Este modelo pressupõe que os parâmetros β dependem apenas da variável k e não da marca i (Cooper e Nakanishi, 1988, secção 2.5).

Um exemplo deste modelo é o caso em que a atração de um produto da marca i é dada por

$$A_{it} = \exp(\alpha_i) \prod_{k=1}^{K} x_{kit}^{\beta_k} \cdot \varepsilon_{it}.$$

A estimação de parâmetros utilizando diretamente estes modelos pode ser muito trabalhosa. No entanto, o modelo pode ser reduzido a um modelo linear de forma a que se possam usar técnicas de regressão linear (Cooper e Nakanishi, 1988, secção 2.5). Se, por exemplo, aplicarmos o logaritmo em ambos os membros obtemos a expressão

$$\log s_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k \log x_{kit} + \log \varepsilon_{it} - \log \Big(\sum_{j=1}^{m} (\alpha_j \prod_{k=1}^{K} x_{kjt}^{\beta_k} \varepsilon_{jt}) \Big).$$

Por outro lado, se somarmos a equação anterior para todo o i e dividirmos por m obtemos:

$$\frac{1}{m} \sum_{i=1}^{m} \log s_{it} = \frac{1}{m} \sum_{i=1}^{m} \Big(\alpha_i + \sum_{k=1}^{K} \beta_k \log x_{kit} + \log \varepsilon_{it} - \log \Big(\sum_{j=1}^{m} (\alpha_j \prod_{k=1}^{K} x_{kjt}^{\beta_k} \varepsilon_{jt}) \Big) \Big).$$

Designando por $\tilde{s}_t$, $\tilde{x}_{kt}$ e $\tilde{\varepsilon}_t$, a média geométrica de s_t, x_{kt} e ε_t respetivamente, podemos re-escrever a equação anterior como:

$$\log \tilde{s}_t = \bar{\alpha} + \sum_{k=1}^{K} \beta_k \log \tilde{x}_{kt} + \log \tilde{\varepsilon}_t - \log \Big(\sum_{j=1}^{m} (\alpha_j \prod_{k=1}^{K} x_{kjt}^{\beta_k} \varepsilon_{jt}) \Big).$$

Finalmente, subtraindo as duas expressões obtemos um modelo linear:

$$\log \frac{s_{it}}{\tilde{s}_t} = (\alpha_i - \bar{\alpha}) + \sum_{k=1}^{K} \beta_k \log \frac{x_{ki}}{\tilde{x_{kt}}} + \log \frac{\varepsilon_{it}}{\tilde{s}_t}.$$

Considerando ainda que não há efeito competitivo na atração da empresa, pode considerar-se um outro tipo de modelação.

O modelo MNL (MultiNomial Logit) pressupõe que:

Modelo MNL: $A_{it} = \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it})$

Tal como no modelo MCI a expressão pode ser manipulada de forma a obter um problema de estimação linear.

Como visto, outros modelos comuns na modelação da quota de mercado são:

Modelo Linear: $s_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it}$

Modelo Multiplicativo: $s_{it} = \exp(\alpha_i) \prod_{k=1}^{K} x_{kit}^{\beta_k} \varepsilon_{it}$

Modelo Exponencial: $s_{it} = \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it})$

As relações mais interessantes entre estes modelos e os modelos de atração são a relação entre o modelo multiplicativo e o modelo MCI e ainda entre o MNL e o modelo exponencial. Relativamente ao modelo multiplicativo, este assume que a quota é uma função multiplicativa das variáveis explicativas. Por sua vez, o modelo MCI assume que as atrações são uma função multiplicativa das variáveis, sendo a quota obtida através da normalização das atrações. A principal diferença é então a normalização (Cooper e Nakanishi, 1988, secção 2.5).

Resumidamente, as formas simplificadas para estes modelos são:

Modelo Linear: $s_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it}$

Modelo Multiplicativo: $\log s_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k \log x_{kit} + \log \varepsilon_{it}$

Modelo Exponencial: $\log s_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it}$

Modelo MCI $\log \frac{s_{it}}{\bar{s}_t} = \alpha_i^* + \sum_{k=1}^{K} \beta_k \log \frac{x_{kit}}{\bar{x_{kt}}} + \varepsilon_{it}^*$

Modelo MNL $\log \frac{s_{it}}{\bar{s}_t} = \alpha_i^* + \sum_{k=1}^{K} \beta_k (x_{kit} - \bar{x_{kt}}) + \varepsilon_i^*$

Continuando com os modelos de atração, podemos introduzir uma maior complexidade ao modelo considerando que os parâmetros β dependem não só de k mas também da marca i. Esta mudança resulta na seguinte fórmula para a atração A_{it} :

Modelo MCI: $A_{it} = \exp(\alpha_i + \varepsilon_{it}) \prod_{j=1}^{m} \prod_{k=1}^{K} f_k(x_{kit})^{\beta_{ki}}$

Modelo MNL: $A_{it} = \exp(\alpha_i + \sum_{k=1}^{K} \sum_{j=1}^{m} \beta_{ki} x_{kit} + \varepsilon_{it})$

Esta modelação é também designada por *differentials effects model* (Cooper e Nakanishi, 1988, secção 3.3). No entanto, esta modelação ainda não é totalmente adequada na maioria das aplicações. Porquê? Não são considerados efeitos entre marcas. A atração da marca i é escrita em função apenas das suas ações/variáveis de marketing.

De forma a contemplar uma estrutura que tenha em conta os efeitos entre marcas, podemos definir o modelo de atração extendido (Cooper e Nakanishi, 1988, secção 3.4):

Modelo MCI extendido: $A_{it} = \exp(\alpha_i + \varepsilon_{it}) \prod_{j=1}^{m} \prod_{k=1}^{K} f_k(x_{kjt})^{\beta_{kij}}$

Modelo MNL extendido: $A_{it} = \exp(\alpha_i + \sum_{k=1}^{K} \sum_{j=1}^{m} \beta_{kij} x_{kjt} + \varepsilon_{it})$

Relativamente ao desempenho destes modelos, Naert e Weverbergh (1981) num estudo compreendendo dados empíricos no mercado de consumo, tanto no mercado da gasolina como em lâminas de barbear, concluiram que os modelos de atração resultam em melhores previsões que os modelos linear e multiplicativo. Um estimador OLS poderá estimar os β deste modelo ou com GLS para tomar em conta a correlação dos erros. Nalbantov *et al.* (2010) estima os parâmetros usando SVMs.

Nos modelos de atração, a quota duma dada empresa varia em função explicita da quota das outras empresas. Considerando o caso em estudo, isto revela-se um problema para os nossos dados uma vez que os relatórios que nos foram disponibilizados *apenas* contêm dados da quota da empresa em estudo. Posto isto, temos informação para a quota SONAE e para a quota não-SONAE (visto que a soma destas é 1).

De forma a utilizar este modelo consideramos então a existência de duas "empresas": a SONAE e a não-SONAE. Trata-se de uma aproximação muito grosseira mas os dados não permitem fazê-lo de outra forma. Tomando a SONAE como empresa 1 e a não-SONAE como empresa 2, as quotas são, respetivamente, dadas por: $s_{1t} = A_{1t}/(A_{1t} + A_{2t})$ e $s_{2t} = A_{2t}/(A_{1t} + A_{2t})$.

Tendo em conta que apenas consideramos duas empresas, então a estimação deste modelo pode reduzir-se a um modelo linear que determina o logit da quota de uma empresa como uma combinação linear de uma transformação das variáveis. Para verificar este raciocínio, basta aplicar o logaritmo a ambos os membros das expressões da quota das duas empresas e subtrair:

$$\log(s_{1t}) - \log(s_{2t}) = \log(A_{1t}) - \log(A_{1t} + A_{2t}) - \log(A_{2t}) + \log(A_{1t} + A_{2t})$$

$$\Leftrightarrow \log(s_{1t}) - \log(s_{2t}) = \log(A_{1t}) - \log(A_{2t})$$

Uma vez que existem duas empresas então $s_{2t} = 1 - s_{1t}$ e a expressão anterior equivale a:

$$\log\left(\frac{s_{1t}}{1 - s_{1t}}\right) = \log(A_{1t}) - \log(A_{2t})$$

Para o caso do modelo MCI extendido,

$$A_{1t} = \exp(\alpha_1 + \varepsilon_{1t}) \prod_{j=1}^{m} \prod_{k=1}^{K} f_k(x_{kjt})^{\beta_{k1j}} \text{ e}$$

$$A_{2t} = \exp(\alpha_2 + \varepsilon_{2t}) \prod_{j=1}^{m} \prod_{k=1}^{K} f_k(x_{kjt})^{\beta_{k2j}},$$

de onde resulta a forma reduzida:

$$\text{logit}(s_{1t}) = (\alpha_1 - \alpha_2) + (\varepsilon_{1t} - \varepsilon_{2t}) + \sum_{k=1}^{K} (\beta_{k1j} - \beta_{k2j}) \log(f_k(x_{kjt})).$$

Para o modelo MNL,

$$A_{1t} = \exp(\alpha_1 + \sum_{k=1}^{K} \sum_{j=1}^{m} \beta_{k1j} x_{kjt} + \varepsilon_{1t}) \text{ e}$$

$$A_{2t} = \exp(\alpha_2 + \sum_{k=1}^{K} \sum_{j=1}^{m} \beta_{k2j} x_{kjt} + \varepsilon_{2t})$$

e portanto:

$$\text{logit}(s_{1t}) = (\alpha_1 - \alpha_2) + (\varepsilon_{1t} - \varepsilon_{2t}) + \sum_{k=1}^{K} (\beta_{k1j} - \beta_{k2j}) x_{kjt}.$$

Demonstramos assim que, para o caso de duas empresas, o modelo de atração é idêntico ao modelo logístico (secção 5.2).

6.2 Modelos de Escolha

Nesta secção será feita uma revisão do método Bayesiano proposto por Chen e Yang (2007) para modelar o comportamento individual dos consumidores utilizando dados agregados. Neste trabalho, a modelação da dinâmica de compra de um cliente é feita recorrendo a dados agregados ao nível da loja. Uma das principais dificuldades no estudo da dinâmica de compra consiste no facto de não podermos observar o comportamento passado de um cliente quando apenas existe informação agregada. Como tal, os modelos existentes ignoram a dinâmica de procura quando analisam os dados agregados, sendo construídos através de um processo de maximização da função utilidade, capturando assim a heterogeneidade dos consumidores.

Uma das vantagens principais do método apresentado neste trabalho de Chen e Yang é a ca-

pacidade de modelar o comportamento de um consumidor, tendo em conta o histórico de procura agregada. Isto será possível simulando um conjunto de escolhas consistentes com os dados observados através do método de simulação Monte Carlo. No entanto, uma vez que as escolhas individuais não são conhecidas, esta coerência entre as simulações e os dados reais não pode ser medida.

Para avaliar a validade do modelo os autores fizeram várias simulações. Primeiro fixaram os parâmetros e geraram a procura agregada através da escolhas individuais obtidas por simulação. De seguida, tendo apenas acesso às procuras agregadas, utilizam o método para verificar se obtêm uma estimativa dos parâmetros próxima do valor que usaram inicialmente (ver Chen e Yang, 2007, Tabela 1).

Geração de Escolhas a partir dos Dados Observados

Assume-se que a procura agregada, S_{jt}, do produto j no instante t é gerada por M clusters de consumidores no mercado. Dentro de um mesmo cluster assume-se que todos os clientes têm a mesma função utilidade e todos os clusters têm o mesmo número de clientes. Além disso, supõe-se que cada cliente consome apenas uma unidade de produto por cada período de tempo. Aqui, o cluster é considerado a unidade mínima de desagregação.

A função utilidade para o produto j no cluster i (no instante t) é

$$u_{ijt} = \theta_i^{\mathsf{T}} x_{ijt} + \varepsilon_{ijt}, \tag{6.1}$$

onde x_{ijt} é o vetor que inclui variáveis de marketing, θ_i é uma variável com distribuição normal multivariada ($\theta_i \sim \text{NMV}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) e ε_{ijt} é o termo erro com função distribuição de valor extremo do tipo-I (também conhecida por distribuição Gumbel). As variáveis de marketing compreendem, por exemplo, o preço e existência de campanhas promocionais e podem ainda ser incluídas as simulações do histórico de procura de forma a captar a dinâmica de procura.

Para J produtos, a função de utilidade especificada conduz a uma probabilidade de escolha logit:

$$\Pr(y_{ijt} = 1) = s_{ijt} = \frac{\exp(\theta_i x_{ijt})}{\sum_{k=0}^{J} \exp(\theta_i x_{ikt})},$$

com $y_{ijt} = 1$ se os clientes do cluster i consomem o produto j no instante t (Viton, 2010). Caso contrário, $y_{ijt} = 0$.

Os dois métodos mais comuns para estimar modelos discretos de escolha usando dados agregados são:

- Minimizar as discrepâncias entre os valores de quota observados e os previstos, por exemplo, minimizar o erro quadrático médio.
- Maximizar a função de máxima verosimilhança. De acordo com esta aproximação, a função maxima verosimilhança correpondente à equação 6.1 é:

$$L = \prod_{t=1}^{\mathsf{T}} \left\{ C^{M}_{O_{0,t},\ldots,O_{Jt}} \prod_{j=0}^{J} \left[\int s_{ijt} f(\theta_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\theta_i \right]^{O_{jt}} \right\}, \tag{6.2}$$

onde:

- $C^{M}_{O_{0,t},\ldots,O_{Jt}}$ é o coeficiente multinomial;
- $f(\theta_i|\boldsymbol{\mu},\boldsymbol{\Sigma})$ é a função densidade de θ_i;
- O_{jt} é o número de clusters, em M, que escolhe a marca j no instante t:

$$O_{jt} = \lfloor S_{jt}M + 0.5 \rfloor.$$

Esta função máxima verosimilhança é baseada no no facto de se considerar que os M clusters são permutáveis e que cada grupo tem uma probabilidade de escolha esperada dada por $\int s_{ijt} f(\theta_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) d\theta_i$.

Apesar da estimação do modelo com dados agregados ser viabilizada por estas duas abordagens, não é muito fácil incorporar o histórico de compra do consumidor visto que esta informação não está diretamente disponível. Para superar esta dificuldade é proposto um modelo Bayesiano hierárquico que permite tratar as escolhas individuais como variável, obtendo-as por "ampliação" dos dados agregados.

A ideia é usar um conjunto R de clusters de forma que a probabilidade média das escolhas neste subconjunto aproxime a dos restantes $M - R$ clusters. Esta abordagem facilita o processo de modelação com dados agregados pois o histórico simulado pode ser utilizado directamente quando é calculada s_{ij}.

O ponto de partida deste modelo é a função máxima verosimilhança dada na equação (6.2); no entanto, assume-se que existem R clusters representativos dos M cujas escolhas (denotadas por y_{rjt}, com $r \in R$ e $R \leq M$) serão obtidas por "ampliação" dos dados.

Conhecidos os dados agregados, o histórico de procura, h, em R é um conjunto de y_{rjt} tal que:

$$\sum_{r=1}^{R} y_{rjt} \leq O_{jt}.$$

Esta condição resulta do facto de R ser um subconjunto de M.

Seja H o conjunto de todos os h, ou seja, o conjunto de todos os históricos de compra possíveis. Podemos escrever a função máxima verosimilhança dos dados agregados como:

$$L = \sum_{h \in H} \left(L_{R|h} x L_{M-R|h} \right),$$

onde:

- $L_{R|h} = \prod_{r=1}^{R} \int [\prod_{t=1}^{T} \prod_{j=0}^{J} (s_{rjt})^{y_{rjt}} f(\theta_r|\boldsymbol{\mu},\boldsymbol{\Sigma})] d\theta_r$ é a função máxima verosimilhança nos R clusters;
- $L_{R-M|h} = \prod_{t=1}^{T} \left\{ C^{M-R}_{z_{0t},\ldots,C_{Jt}} \prod_{j=0}^{J} \left[\int s_{ijt} f(\theta_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) \theta_i \right]^{Z_{jt}} \right\}$, com $Z_{jt} = O_{jt} - \sum_{r=1}^{R} y_{rjt}$ e $C^{M-R}_{z_{0t},\ldots,C_{Jt}}$, o coeficiente multinomial.

Uma vez que se assume que os R clusters sao representativos dos M clusters então assume-se que $\int s_{ijt} \mathrm{f}(\theta_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) d\theta_i = \sum_{r=1}^{R} \frac{s_{rjt}}{R}$.

Com a simulação de escolhas nos R clusters, este trabalho permite introduzir um histórico de procura na função utilidade e consequentemente em s_{ijt}. Sem este aumento nos R clusters, quando

s_{ijt} é função do histórico, seria necessário integrar sobre todas as possibilidades $\int s_{ijt}(\theta_i|\bar{\theta})d\theta_i$, o que seria impraticável.

Em relação à escolha de M e R, pode tomar-se $R = M$. No entanto, nesse caso é necessário que $\sum_{r=1}^{R} y_{rjt} = O_{jt}$ e, consequentemente, o algoritmo de simulação de dados torna-se ineficiente. Um valor de R relativamente grande em relação a M reduz a taxa de aceitação na geração de escolhas simuladas, o que torna o algoritmo menos eficiente. Por outro lado, um valor de R muito pequeno, torna o algoritmo mais eficiente mas pode não ser representativo de M.

No seu trabalho, Chen e Yang, concluem que quando o valor de M é desconhecido, se os resultados de estimação forem robustos quando se varia R fixando M, então isto sugere que a escolha de M é razoável.

Vejamos agora como estimar os parâmetros do modelo ($\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$). A função $L = \sum_{h\in H}(L_{R|h} x L_{M-R|h})$ é, como referido, difícil de tratar devido à distribuição de heterogeneidade dos consumidores e ao grande número de combinações de histórico de compra admissível (h) que são consistentes com as ações agregadas. Por esta razão, é tomada uma análise bayesiana usando dados simulados. A análise bayesiana dos dados agregados é feita especificando a função distribuição conjunta de todos os parâmetros do modelo.

A função densidade de probabilidade conjunta posterior pode ser escrita como

$$f\big(\{y_t\},\{\theta_r\},\boldsymbol{\mu},\boldsymbol{\Sigma}|\{S_t\},\{x_{rt}\}\big) \propto \prod_{t=1}^{T}\Big\{f(S_t|y_t,\theta_1,...,\theta_R)\prod_{r=1}^{R}[f(y_{rt}|\theta_r,x_{rt})f(\theta_r|\boldsymbol{\mu},\boldsymbol{\Sigma})f(\boldsymbol{\mu},\boldsymbol{\Sigma})]\Big\},$$

onde

- y_t é o conjunto de escolhas $y_{rjt}, r \in \{1,...,R\}, j \in \{1,...,J\}$;
- $f(S_t|y_t,\theta_1,...,\theta_R) = L_{M-R|h,t}$;
- $f(y_{rt}|\theta_r,x_{rt}) = \prod_{j=0}^{J}(s_{rjt})^{y_{rjt}}$ é a probabilidade de escolha no instante t, do cluster r, cujas escolhas foram simuladas;
- $f(\theta_r|\boldsymbol{\mu},\boldsymbol{\Sigma})$ é a distribuição de heterogenidade;
- $f(\boldsymbol{\mu},\boldsymbol{\Sigma})$ é a distribuição a priori.

A estimação será feita através de cadeias de Markov e gerando iterativamente amostras para os parâmetros do modelo. No seu trabalho, os autores consideram ainda a opção de não escolha de nenhuma das marcas de interesse. A não escolha é denotada por $j = 0$ e tem associada uma função utilidade dada por

$$u_{i0t} = \varepsilon_{i0t}.$$

Passo 1. Geração de y_t

A chave deste algoritmo é a geração de escolhas individuais (a nível dos R clusters) gerando y_t de forma condicional a outros parâmetros do modelo. Pela equação de probabilidade conjunta

posterior, y_t é proporcional a

$$\prod_{\tau=t}^{t'} \Big[C^{M-R}_{z_{0t},\ldots,C_{Jt}} \prod_{j=0}^{J} \Big(\sum_{r=1}^{R} s_{rj\tau}/R \Big)^{Z_{jt}} \Big] \Big(\prod_{r=1}^{R} \prod_{j=0}^{J} s_{rj\tau}^{y_{rj\tau}} \Big),$$

onde

$$s_{rjt} = \frac{\exp(\theta_r^{\intercal} x_{jt})}{\sum_k \exp(\theta_r^{\intercal} x_{kt})}$$

e $t+1, \ldots, t'$ são os períodos em que a escolha dos consumidor é afetada pela escolha em t. Se não houver este efeito dinâmico $t' = t$. Uma vez que $f(y_{rt}|\theta_r, x_{rt}) = \prod_{j=0}^{J} (s_{rjt})^{y_{rjt}}$, podem gerar-se amostras de $y_{rt} (r = 1, \ldots, R)$ recorrendo a uma função distribuição discreta com $J+1$ valores possíveis.

Cada amostra y_{rt} consiste num vetor em que apenas uma das entradas é 1 (indicando a escolha do consumidor dentro das $J+1$ possibilidades). A probabilidade de cada resultado é s_{rjt}, que é a probabilidade logit apresentada. A amostra candidata é qualificada se Z_{jt} é não negativo para todo o j; caso contrário, uma nova amostra é gerada.

Considere-se que a amostra anterior é $y_t^{(p)}$ e a seguinte é $y_t^{(n)}$, a probabilidade de aceitação da amostra $y_t^{(n)}$ é dada por

$$\min \Big\{ \prod_{\tau=t}^{t'} \Big[C^{M-R}_{z_{0t},\ldots,C_{Jt}} \prod_{j=0}^{J} \Big(\sum_{r=1}^{R} s_{rj\tau}/R \Big)^{Z_{jt}} \Big] \Big(\prod_{r=1}^{R} \prod_{j=0}^{J} s_{rj\tau}^{y_{rj\tau}} \Big) \Big| y_t^{(n)} \Big/$$

$$\prod_{\tau=t}^{t'} \Big[C^{M-R}_{z_{0t},\ldots,C_{Jt}} \prod_{j=0}^{J} \Big(\sum_{r=1}^{R} s_{rj\tau}/R \Big)^{Z_{jt}} \Big] \Big(\prod_{r=1}^{R} \prod_{j=0}^{J} s_{rj\tau}^{y_{rj\tau}} \Big) \Big| y_t^{(p)}, 1 \Big\}$$

Se $y_t^{(n)}$ não for aceite, $y_t^{(p)} := y_t^{(n)}$ e é gerada uma nova amostra $y_t^{(n)}$.

Passo 2. Gerar θ_r

Depois de geradas as escolhas dos R clusters, θ_r pode ser gerado usado o algoritmo de Metropolis-Hastings (Chib e Greenberg, 1995).

A função máxima verosimilhança de θ_r é

$$l(\theta_r) \propto \prod_{t=1}^{T} \Big[\prod_{j=0}^{J} \Big(\sum_{r=1}^{R} \frac{s_{rjt}}{R} \Big)^{Z_{jt}} \prod_{j=0}^{J} s_{rjt}^{y_{rjt}} \Big], \text{ e}$$

$$s_{ijt} = \frac{\exp(\theta_r^{\intercal} x_{jt})}{\sum_k exp(\theta_r^{\intercal} x_{kt})}.$$

Por outro lado, assumindo que θ_r provém de uma variável aleatória normal multivariada com média $\boldsymbol{\mu}$ e matriz covariância $\boldsymbol{\Sigma}$, a função probabilidade posterior é

$$f(\theta_r) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-1/2(\theta_r - \boldsymbol{\mu})^{\intercal} \boldsymbol{\Sigma}^{-1} (\theta_r - \boldsymbol{\mu})) l(\theta_r).$$

Para gerar amostras de θ_r é usado o método de Metropolis-Hastings (Chib e Greenberg, 1995,

página 330). Seja $\theta_r^{(p)}$ a amostra anterior e $\theta_r^{(n)}$, a probabilidade de aceitação da amostra é

$$\min\left\{\frac{\exp[-1/2(\theta_r^{(n)}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}](\theta_r^{(n)}-\boldsymbol{\mu})I(\theta_r^{(n)})}{\exp[-1/2(\theta_r^{(p)}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}](\theta_r^{(p)}-\boldsymbol{\mu})I(\theta_r^{(p)})},1\right\}$$

Cada nova amostra é gerada a partir de $\theta_r^{(p)}$:

$$\theta_r^{(n)} := \theta_r^{(p)} + \Delta,$$

onde Δ é uma amostra de uma normal multivariada de média $\mathbf{0}$ e matriz covariância $0.015\,\mathbf{I}$, sendo $\mathbf{I}$ a matriz identidade.

Passo 3. Gerar $\boldsymbol{\Sigma}$

Gerar elementos da diagonal de $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{kk}$, $k \in \{1,\dots,K\}$ onde K é a dimensão de θ_r. A distribuição posterior de $\boldsymbol{\Sigma}_{kk}$ é tal que

$$f(\boldsymbol{\Sigma}_{kk}|\theta_{rk},\bar{\theta_k}) \propto \text{InvGamma}(a,b),$$

onde $a = s_0 + \frac{R}{2}$, $(s_0 = 3)$ e $b = \frac{2}{\sum_{r=1}^{R}(\theta_{rk}-\bar{\theta_k})^2+2q_0}$, $(q_0 = 0.2)$.

Passo 4. Gerar $\boldsymbol{\mu}$

$$f(\boldsymbol{\mu}|\theta_r,\boldsymbol{\Sigma}) = MNV(v,\Psi),$$

onde $v = \Psi(\sum_{r=1}^{R}\frac{\theta_r}{R} + \boldsymbol{\Sigma}_0)$, $\Psi = (\boldsymbol{\Sigma}_0^{-1} + R\boldsymbol{\Sigma}^{-1})^{-1}$ e $\boldsymbol{\Sigma}_0 = 100\mathbf{I}$.

Passo 5. Voltar ao Passo 1

Enquanto θ_r não for aceite ($\forall r \in \{1,\dots,R\}$) volta a 1.

Por vezes a convergência é muito lenta e um critério bastante utilizado é o número máximo de iterações.

6.2.1 Aplicação do algoritmo ao caso em estudo

No caso em estudo, estamos perante um contexto mais amplo do que uma loja. Apresentamos de forma esquemática a analogia do nosso problema ao modelo proposto por Chen e Yang:

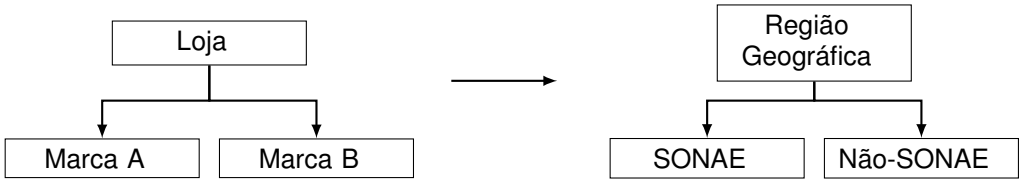


Figura 6.1: Aplicação original dos autores à esquerda versus a nossa aplicação à direita.

Os nossos dados não permitem simular a não escolha, pelo que cada cluster de clientes apenas pode escolher a SONAE ou a não-SONAE (tal como nos modelos de atração).

Na implementação consideramos as variáveis: ser SONAE (x_1), ser não-SONAE (x_2) e fazer campanha (x_3). Com este modelo estaremos a simular o peso que cada cluster homogéneo de consumidores dá ao facto dos produtos serem da SONAE ou da concorrência e ainda ao facto de haver campanha promocional.

A implementação deste algoritmo foi feita em R e a sua execução é lenta para valores elevados do número de clusters R. Foram testadas todas as combinações para os valores de clusters $M \in \{30, 40, 50\}$ e de $R \in \{10, 20, 30\}$. A estimação de parâmetros foi feita recorrendo aos meses entre janeiro de 2011 e abril de 2014 e os erros foram obtidos prevendo os últimos 7 meses. Os erros obtidos para a unidade 51, em Portugal Continental, são representados na Figura 6.2:

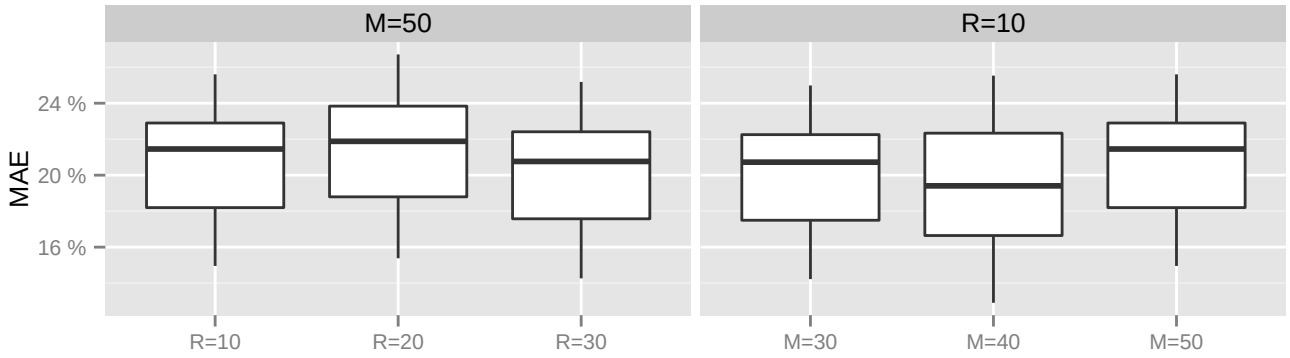


Figura 6.2: Erros absolutos usando o método da escolha.

A execução destes modelos demorou no total 14 horas tendo sido usado um processador i3 1.80GHz de 4 cores, a correr simulações em paralelo.

Cada modelo foi simulado 100 vezes, o valor inicial de μ foi $(1, 1, 1)$ e para as diagonais da matriz de covariância usamos o valor 0.1. Além disso, escolhemos $t' = 1$. Em geral a cadeia de Markov convergiu antes das 30 iterações. Com este modelo tentamos estimar os valores dos parâmetros μ e a diagonal da matriz de covariância $\mathbf{\Sigma}$. Fixando $M = 50$ observamos que a alteração do valor de R introduzia uma variação apenas a partir da segunda casa décimal nas estimativas de μ. No entanto, no que respeita à diagonal da matriz de covariância, as alterações foram maiores. Por exemplo, para $R = 10$, os valores estimados são metade dos valores estimados usando $R = 20$. Segundo os autores, se o valor de M fosse próximo do ideal, então existiria convergência dos parâmetros para um dado valor mesmo alterando o valor de R, o que não aconteceu na nossa simulação.

Os autores referem, nas suas aplicações, que iterações na ordem das dezenas ou centenas de milhar são usadas nas suas simulações. Tudo isto sugere que talvez devessemos usar valores de M e R superiores.

A análise dos resultados revela um erro absoluto elevado. No entanto, seria expectável que os erros fossem elevados uma vez que, como já foi referido, estamos a fazer uma simplificação muito grande ao considerar que existem apenas duas empresas. A título ilustrativo mostramos as estimativas $\hat{\mu}$ obtidas pelas simulações usando $M = 50$ e $R = 10$: os seus valores foram 0.030,0.037 e 0.04. Se este modelo se ajustasse bem aos nossos dados isto significaria que em média a função utilidade de um cliente atribuía um peso de 0.03 ao facto de um produto ser SONAE, 0.037 ao facto

de ser da concorrência e 0.04 à existência de campanha promocional. De notar que neste caso as campanhas teriam um peso superior na função utilidade de um cliente.

De forma a verificar se um maior número para os valores de *M* e *R* produzia melhores resultados, decidimos testar ainda a estimação de parâmetros utilizando $M = 200$ e $R = 50$. Executamos a cadeia de Markov para 600 iterações e portanto 600 amostras foram geradas para cada parâmetro. Decidimos estimar os parâmetros a partir da média das últimas 400 amostras. O desvio padrão registado para as estimativas de μ é para os três parâmetros muito próximo de 0.14 e a matriz de covariâncias obtida sugere um grande variabilidade entre os clusters. Estes parâmetros levaram a um erro absoluto muito semelhante aos anteriores.

Outras variáveis poderiam ter sido testadas, nomeadamente a proporção de lojas da SONAE e não-SONAE, no entanto, dado o tempo de execução da nossa implementação para valores elevados de *M* e *R* optamos por não fazê-lo. Uma outra razão é a simplificação feita ao usar um mercado com duas empresas, que em princípio não permitirá que o modelo tenha um bom desempenho. Além disso, não estamos a modelar um único produto da SONAE e não-SONAE mas sim uma família de produtos.

Como nota final, reforçamos que estes modelos são particularmente úteis para estimar a importância que o consumidor dá a uma dada característica da empresa e que dada a relação entre a função utilidade e a probabilidade de escolha de um produto, este modelo permite ainda estimar a quota do mesmo.

Capítulo 7

Resultados

Os resultados dos modelos produzidos ao longo da dissertação são agora expostos:

7.1 Comparação entre os Modelos

Nesta secção fazemos a escolha do melhor modelo, de entre os apresentados nos Capítulos 4 e 5, para cada uma das regiões e ainda para o caso em que um único modelo é construído para prever todas as regiões, sendo neste caso inserida uma variável explicativa extra: a região. O desempenho dos métodos é feito através do método k-fold para os métodos em data mining e através de uma janela deslizante para o modelo ARIMA. O conjunto de dados usado para esta avaliação de desempenho compreende as observações de janeiro de 2011 a abril de 2014. A razão da utilização desta janela temporal para k-fold, em vez de uma escolha aleatória de observações, é permitir uma comparação o mais justa possível dos métodos de data mining com os métodos ARIMA. Deixando para posterior avaliação do melhor modelo os últimos 7 meses estamos a garantir que estes não são utilizados pelo ARIMA na validação cruzada (ver seccão 2.7). Em relação à medida usada para escolher o modelo com melhor conjunto de erros interessa que esta tenha em conta outliers uma vez que queremos evitar um modelo com eventuais erros elevados. Por esta razão optamos por utilizar a média dos erros.

Na Figura 7.1 na próxima página, estão representados os erros médios cometidos na validação cruzada. Pela sua análise podemos observar que os modelos estimados com base em todos os dados incorporando a variável região tem melhor desempenho do que os modelos construídos usando apenas os dados região a região. Uma justificação possível para estes resultados é o facto de que, usando todos os dados, estamos a construir um modelo a partir de $5\times$ mais informação para cada uma das variáveis correspondentes às campanhas promocionais e os modelos podem por isso obter melhores estimativas para os parâmetros associados a estas. Este resultado é também um indício de que os impactos das variáveis explicativas em cada uma das regiões é semelhante. No geral, as redes neuronais tiveram um erro médio inferior aos restantes métodos.

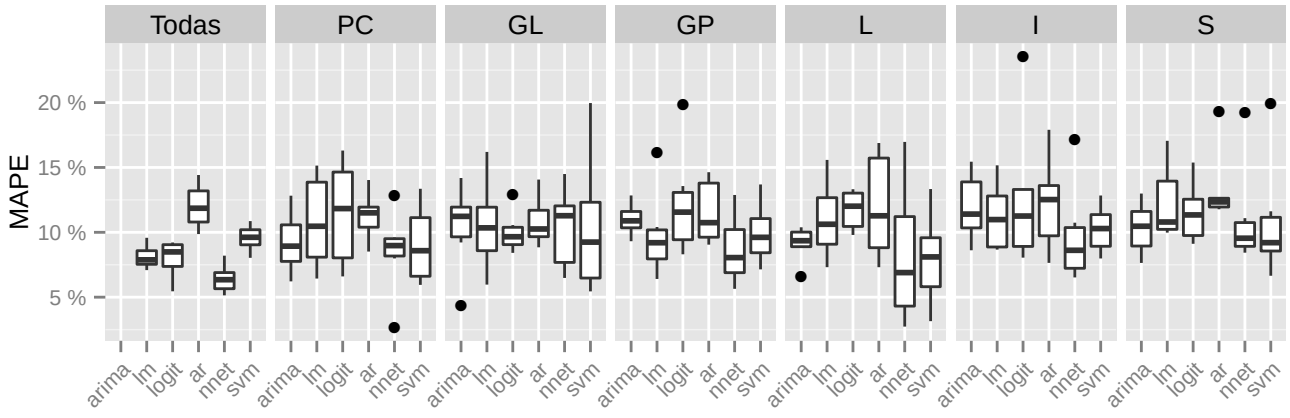


Figura 7.1: Desempenho dos modelos para a unidade 51.

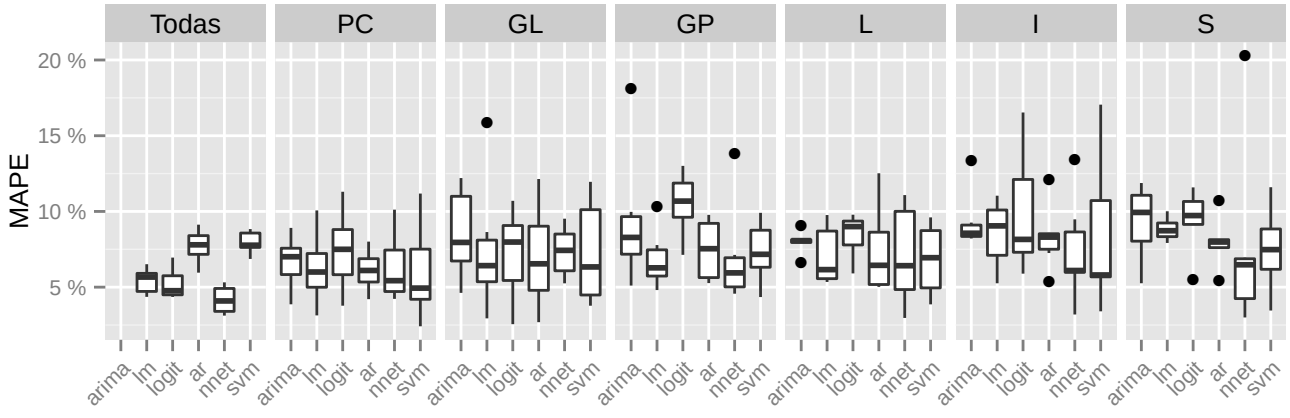


Figura 7.2: Desempenho dos modelos para a unidade 53.

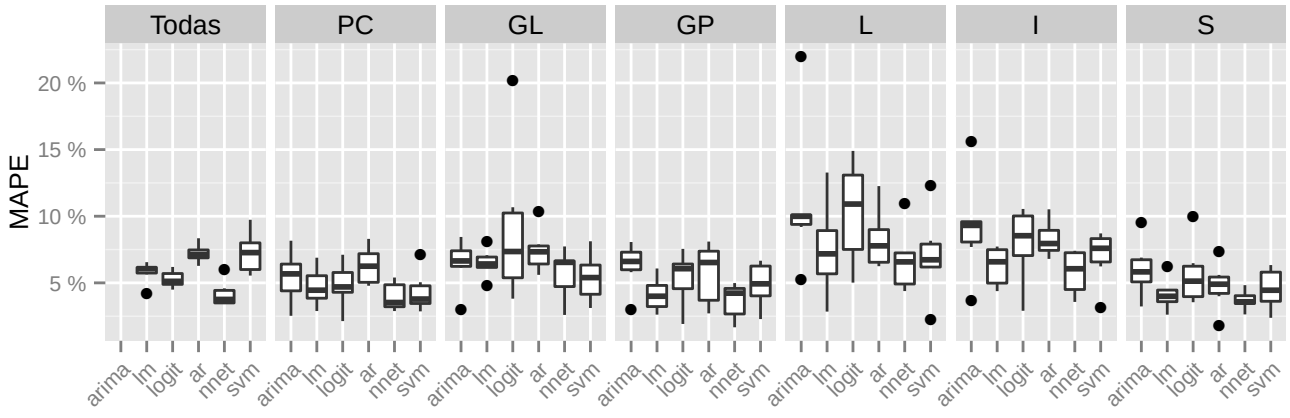


Figura 7.3: Desempenho dos modelos para a unidade 54 e 55.

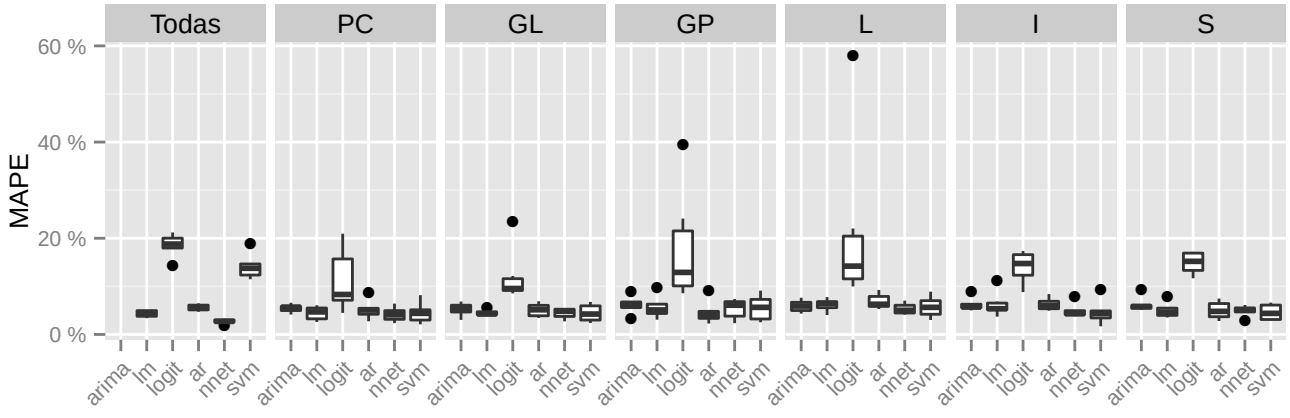


Figura 7.4: Desempenho dos modelos para a quota total (exceto unidade 52).

Analisando a Figura 7.2, referente aos erros para a unidade 53, verificamos que tal como na unidade 51, o modelo único com a variável região, que ajusta os valores para cada uma destas, induz melhores resultados. Mais uma vez, as redes neuronais apresentam erros médios menores. De notar que na Grande Lisboa e Grande Porto os erros médios de todos os modelos são bastante próximos.

Na Figura 7.3 e na Figura 7.4 estão representados os erros para a unidade 54/55 e para o total exceto 52, respetivamente. Mais uma vez, o uso de um modelo único induz um menor erro médio e as redes neuronais revelam um melhor desempenho em quase todas as regiões.

Como comentário final, podemos observar que a regressão linear revelou melhor desempenho do que o modelo logit.

7.2 Modelos Hierárquicos

No caso em estudo, as séries temporais, por unidade de negócio, possuem uma estrutura hierárquica (temos dados para cada região *e* para Portugal Continental). Podemos aproveitar esta estrutura complementando os modelos com a informação conjunta de forma a obter modelos mais robustos. Para a unidade de negócios 51 é registada a quota para cada uma das regiões e ainda para Portugal Continental:

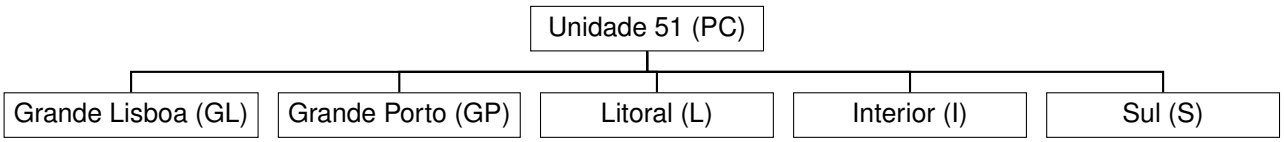


Figura 7.5: Modelo hierárquico para a unidade 51.

As previsões realizadas, por qualquer um dos modelos, devem ser ajustadas de forma a que se possam obter resultados mais robustos. Assume-se que o diagrama anterior pode ser representado como:

$$\begin{pmatrix} x_{\mathrm{PC},t} \\ x_{\mathrm{GL},t} \\ \vdots \\ x_{\mathrm{S},t} \end{pmatrix} = \begin{pmatrix} p_{GL} & p_{GP} & p_L & p_I & p_S \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{GL,t} \\ \vdots \\ x_{S,t} \end{pmatrix},$$

ou de forma mais compacta: $\mathbf{x}_t = \mathbf{S}\,\mathbf{x}_{K,t}$, com $K = \{\mathrm{GL}, \ldots, \mathrm{S}\}$. $\mathbf{x}_{K,t}$ são as observações no nível mais baixo da hierarquia e p_K representa o peso da quota na região K na quota total (nível superior da hierarquia, correspondente a Portugal Continental).

Tome-se $\hat{x}_{k,t}$ como a previsão feita pelo modelo escolhido para cada um dos níveis inferiores da hierarquia e $\hat{x}_{PC,t}$ como a previsão gerada, pelo mesmo modelo, para o nível superior da hierarquia, no instante t. Da mesma forma, $\tilde{x}_{k,t}$ e $\tilde{x}_{PC,t}$ as previsões resultantes considerando a estrutura hierárquica.

Antes de procedermos à explicação dos modelos hierárquicos, explicamos de que forma obtive-

mos as entradas da matriz **S**: se o nosso problema consistisse em, por exemplo, prever vendas, as entradas de **S** tomariam apenas valor 0 ou 1. No entanto, no caso das quotas não é verdade que a quota em Portugal Continental é a soma das quotas regionais. Foi por isso necessário determinar de que forma a quota de Portugal Continental se divide pelas regiões. Para isso utilizamos uma regressão linear da forma $\text{QUOTA}_{PC} = 0 + \beta_1 \text{QUOTA}_{GL} + \beta_2 \text{QUOTA}_{GP} + \beta_3 \text{QUOTA}_{L} + \beta_4 \text{QUOTA}_{I} + \beta_5 \text{QUOTA}_{S}$, obtendo assim estimativas $\hat{p}_{GL} = \beta_1, \ldots, \hat{p}_S = \beta_5$.

Esta abordagem, pouco convencional, não foi encontrada na literatura. Mas, uma vez que as quotas variam pouco ao longo do tempo, então os pesos também variarão pouco, pelo que pensamos justificar-se estruturar desta forma as contribuições das quotas das várias regiões para a quota de PC.

As técnicas mais comuns para construir de modelos hierárquicos são conhecidas como *bottom-up* e *top-down*:

Bottom-Up. Nesta aproximação, são feitas as previsões da base da hierarquia de forma independente, ou seja, prevêm-se as quotas para cada uma das regiões utilizando o modelo escolhido. De seguida, estas previsões são agregadas de forma a obter o nível superior da hierarquia, isto é, agregam-se as quotas regionais para obter a quota a nível de Portugal Continental. Aqui, $\tilde{x}_t = \mathbf{S}\hat{x}_t$.

Top-Down. Esta aproximação utiliza como base a previsão da quota no nível superior da série, ou seja, a quota em Portugal Continental. As quotas a nível regional são determinadas desagregando essa previsão. Para isso são usadas proporções $q_1, \ldots, q_5$ que podem ser obtidas através do histórico, por exemplo.

Existem vários métodos para desagregar um série temporal; neste trabalho usamos a média do histórico de proporções, dada por:

$$q_j = \frac{1}{T} \sum_{t=1}^{T} \frac{x_{j,t}}{x_{PC}},$$

em que T é o número de observações no histórico, $j \in \{GL, \ldots, S\}$ e y_t o total agregado (Hyndman *et al.*, 2011).

Não existe consenso sobre qual destas abordagens leva a melhores resultados. Neste estudo será ainda usada uma abordagem mais recente (Hyndman *et al.*, 2011), que em geral leva a melhores resultados que as duas anteriores:

Combinação Ótima das Previsões. Nesta terceira abordagem, todas as previsões são realizadas de forma independente, aplicando-se posteriormente um modelo de regressão linear de forma a otimizar a combinação das previsões, $\hat{\mathbf{y}}_t = S\beta_t + \varepsilon_t$. No seu trabalho, Hyndman *et al.* provou que o estimador que otimiza a combinação das previsões é dado por:

$$\tilde{\mathbf{y}}_t = \mathbf{S} \left(\mathbf{S}^{\intercal} \mathbf{S}\right)^{-1} \mathbf{S}^{\intercal} \hat{\mathbf{y}}_t$$

Este método aparece na literatura aplicado a casos como a previsão de vendas em que, como já foi dito, o total de vendas é a soma das vendas das hierarquias inferiores e, consequentemente, a matriz **S** tem entradas com valor 0 ou 1 apenas.

Para uma melhor visão do método suponhamos que a estrutura hierárquica é dada, por exemplo,

por:

$$\begin{pmatrix} x_{\mathrm{PC},t} \\ x_{\mathrm{GL},t} \\ \vdots \\ x_{\mathrm{S},t} \end{pmatrix} = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{GL,t} \\ \vdots \\ x_{S,t} \end{pmatrix},$$

então

$$\tilde{\mathbf{y}}_t = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 29/30 & -1/30 & -1/30 & -1/30 & -1/30 \\ 1/6 & -1/30 & 29/30 & -1/30 & -1/30 & -1/30 \\ 1/6 & -1/30 & -1/30 & 29/30 & -1/30 & -1/30 \\ 1/6 & -1/30 & -1/30 & -1/30 & 29/30 & -1/30 \\ 1/6 & -1/30 & -1/30 & -1/30 & -1/30 & 29/30 \end{pmatrix} \hat{\mathbf{y}}_t.$$

Neste caso, a previsão para a Grande Lisboa seria dada por $1/6$ da previsão de Portugal Continental mais $29/30$ da previsão para a Grande Lisboa menos $1/30$ da previsão obtida para o Grande Porto e por aí adiante. Os pesos negativos estão relacionados com as séries que não influenciam diretamente a série considerada. Estes coeficientes são negativos e não nulos para que seja retirado o efeito destas séries sobre a série do total e, uma vez que apenas dependem da estrutura hierárquica e não dos dados observados, são determinados um única vez.

Será feita de seguida uma análise dos resultados com e sem combinação de previsões, sobre a forma de tabelas e gráficos. Para a unidade 51 a análise será mais detalhada, sendo que para as restantes o procedimento foi análogo.

Mais um Nível. Acrescentaremos mais um nível ao modelo de forma a aproveitar a informação a nível do conjunto da unidades.

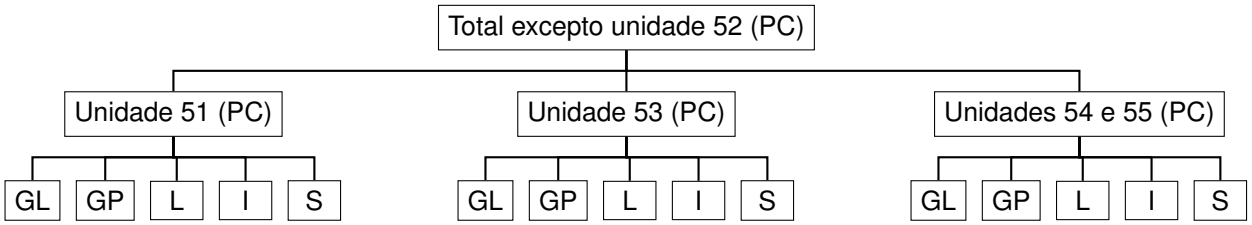


Figura 7.6: Modelo hierárquico de dois níveis: regiões e unidades de negócio.

Aplicação dos Modelos Hierárquicos aos Dados

Pela análise do desempenho dos modelos tínhamos concluído que, de uma forma geral, as redes neuronais resultam em erros médios inferiores. Para a aplicação de modelos hierárquicos utilizamos uma estrutura com três níveis. Apresentamos o erro cometido utilizando k-fold no conjunto de treino, constituído pelo conjunto dos meses compreendidos entre janeiro de 2011 e abril de 2014, na Tabela 7.1 na página seguinte. Ainda na mesma tabela, apresentamos o erro médio cometido pelo modelo com melhor desempenho no conjunto de teste, constituído pelos últimos 7 meses.

Podemos verificar que, no método bottom-up, há um aumento do erro à medida que subimos na hierarquia. Uma razão para que tal ocorra é a acumulação de erros pois as previsões dos níveis

Região	Unidade	Bottom Up	Top Down	Combinação Ótima	Sem Hierarquia
PC	Total	56.5 ± 10.7	5.1 ± 1.3	29.2 ± 7.8	5.0 ± 1.0
PC	51	14.0 ± 3.9	9.9 ± 2.6	12.6 ± 3.6	9.8 ± 2.4
PC	53	15.9 ± 1.4	6.9 ± 1.8	14.1 ± 1.1	6.6 ± 2.6
PC	54/55	49.8 ± 3.1	6.2 ± 2.2	30.9 ± 2.8	6.4 ± 0.7
GL	51	10.7 ± 2.9	12.1 ± 2.3	10.8 ± 2.9	11.3 ± 3.4
GL	53	6.9 ± 2.8	37.4 ± 2.2	7.9 ± 2.3	7.5 ± 2.1
GL	54/55	6.4 ± 2.7	16.2 ± 3.3	7.5 ± 3.0	6.5 ± 1.4
GP	51	9.6 ± 2.0	78.5 ± 12.5	9.6 ± 1.9	10.1 ± 2.1
GP	53	8.5 ± 2.6	15.5 ± 2.2	8.4 ± 2.5	8.4 ± 4.2
GP	54/55	6.4 ± 1.8	46.2 ± 6.3	11.1 ± 5.2	7.0 ± 0.9
L	51	8.8 ± 3.8	26.9 ± 8.1	12.2 ± 5.1	9.3 ± 3.2
L	53	7.0 ± 1.4	7.6 ± 1.9	7.8 ± 1.7	6.9 ± 1.8
L	54/55	7.0 ± 2.3	55.0 ± 5.8	7.7 ± 2.3	6.7 ± 1.5
I	51	10.3 ± 3.4	42.7 ± 7.4	10.5 ± 3.5	10.8 ± 2.7
I	53	7.2 ± 2.7	44.1 ± 1.4	9.8 ± 3.0	7.3 ± 2.1
I	54/55	5.8 ± 1.3	44.3 ± 2.7	22.1 ± 3.2	5.9 ± 2.4
S	51	10.1 ± 3.0	11.4 ± 4.1	11.2 ± 2.8	10.8 ± 2.5
S	53	8.0 ± 3.3	15.6 ± 1.4	8.3 ± 2.8	8.2 ± 1.9
S	54/55	4.2 ± 1.5	5.6 ± 1.3	6.4 ± 2.1	4.0 ± 1.5

Tabela 7.1: Erro médio ± desvio padrão (em %) da combinação dos modelos hierárquicos, no conjunto de validação, e erro cometido pelo melhor método no conjunto de teste.

superiores são obtidas por combinação linear das previsões feitas para cada uma das 5 regiões. No método top-down também se verifica que no nível em que é feita a previsão independente (nível superior correspondente à quota no total), o erro é baixo mas, ao descer na hierarquia, os erros vão aumentando. O método de combinação ótima revela melhores resultados do que os métodos bottom-up e top-down nos níveis em que as previsões são obtidas por combinação de outras. Podemos ver que, de uma forma geral os valores do erro médio estão bastante próximos dos valores obtidos quando prevemos cada uma das combinações unidade/região de forma independente, no entanto, são ligeiramente superiores.

Para concluir, verificamos que a previsão em cada uma das regiões de forma independente se revela mais eficaz do que a obtida recorrendo a uma estrutura hierárquica. Para terminar apresentamos, como já foi referido, na última coluna da tabela, o erro cometido pelo melhor método no conjunto dos últimos 7 meses. Em relação ao desvio padrão do erro, verificamos que no conjunto de teste este se localiza entre o valor 3.5% e 7%.

7.3 Cartogramas

Umas das mais importantes áreas em estatística, por vezes descartada em favor da mais glamorosa estatística inferencial, é a estatística descritiva. Esta disciplina tenta transmitir os dados da

forma mais intuitiva possível e condensada ao utilizador.

Uma vez que qualquer Português conhece bem o mapa de Portugal, podemos explorar essa cognição para representar a área, não em termos reais, mas em relação à variável que se quer explicar. A este tipo de mapas chamam-se **cartogramas**.

O primeiro algoritmo computacional para a elaboração de cartogramas é de Tobler (1973). Enquanto os mapas coropléticos associam quantidades a regiões através do uso de cores, os cartogramas são mapas que modificam a própria topografia da região em proporção com a quantidade a ela associada.

No seu algoritmo original, Tobler formalizou o seguinte problema de otimização: minimizar o determinante do Jacobiano de uma superfície, que é resolvido através de um método iterativo de diferenças finitas. O método contudo requer restrições convolutas para ser preciso (Dougenik *et al.*, 1985).

Um algoritmo baseado em automatos celulares foi sugerido por Appel *et al.* (1983), mas devido ao facto de requerer a conversão da topografia numa grelha pode introduzir demasiadas distorções.

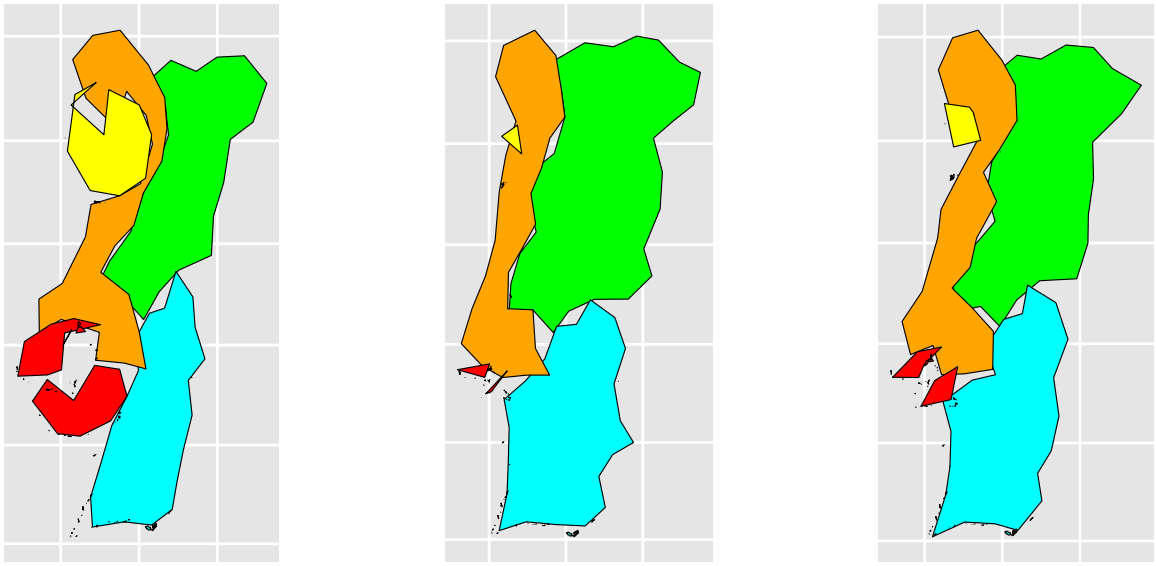
Os algoritmos mais usados contudo são baseados em dinâmicas da física: Gastner e Newman (2004) utiliza algumas ideias de Appel *et al.*, com ideias da física de fluídos em que é associado aos vários nós, a "quantidade" desejada e estes vão sendo expandidos na direcção dos nós com "quantidades" mais pequenas. O processo repete-se até todos os nós terem as mesmas "quantidades", e o mapa estar devidamente distorcido. Este método é recente e parece ser cada vez mais utilizado.

O algoritmo que ainda continua mais em uso, contudo, e que usamos, vem de Dougenik *et al.* (1985). Este deforma directamente a topografia. O algoritmo desenvolvido utiliza o conceito de forças:

$$F_{ij} = \begin{cases} (p_j - q_j)p_j d_{ij} & \text{se } d_{ij} > p_j, \\ (p_j - q_j)((4p_j - 3d_{ij})/p_j)(d_{ij}^2/p_j^2) & \text{caso contrário,} \end{cases} \tag{7.1}$$

em que F_{ij} é a força excercída pelo polígono j no ponto i, p_j é a área actual do polígono normalizada entre todos os polígonos, q_j é a área da variável desejada, também normalizada, e d_{ij} é a distância entre os vários pontos i e os centros do polígono j. O "caso contrário" do sistema é um ajustamento que evita problemas nos casos em que uma coordenada se encontra muito próxima do centro do polígono. Este método é aplicado várias vezes aos vários vértices do mapa, usando um múltiplo da força F_{ij} até o resultado ser visualmente satisfazível.

Estes mapas são formas eficientes de comunicar resultados. O mapa original está na Figura 2.2 na página 4. Utilizamos então um cartograma para mostrar as estimativas do impacto das campanhas promocionais ao longo do país, como estimado pela regressão linear (secção 5.1). Para a transformação da Figura 7.7 na página seguinte foi utilizado o software de sistema de informação geográfica QGIS (2009), com base num mapa administrativo LAU 1 (distritos) do Instituto Geográfico Português (IGEO, 2015), adaptado para as 5 regiões "administrativas" da SONAE e submetido às dinâmicas em (7.1).



(a) PROMO1 (b) PROMO2 (c) PROMO4

Figura 7.7: Cartogramas do efeito de várias promoções na quota da unidade 51.

Capítulo 8

Conclusão

A previsão de séries temporais de quota de mercado ainda é um assunto infante. O problema revelou-se ainda mais severo face à recolha e limitação dos dados. Após uma exploração da literatura e dos dados nos primeiros capítulos, dentro dos dados que nos foram disponibilizados, as campanhas promocionais são as variáveis mais úteis para explicar a variação na quota da empresa.

Os modelos de data mining mais convencionais, e ainda os modelos ARIMA, são comparados no Capítulo 7, o modelo que melhor parece descrever os dados são as redes neuronais. Isto deve-se ao facto das redes neuronais facilmente embeberem efeitos de segunda ordem das variáveis e modelarem qualquer função contínua por partes (Hornik, 1991). Várias outras abordagens menos ortodoxas são avaliadas no Capítulo 6, nomeadamente os vários modelos anteriores são combinados num **modelo hierárquico** de forma a melhorar os resultados. Além disso, é utilizado um **modelo de escolha** e analisamos o **modelo de atração**.

O modelo de atração é comum na literatura de estudo de quotas (Cooper e Nakanishi, 1988); para os nossos dados, mostramos que o modelo se desenvolve num modelo logístico, que tinha sido já explorado no Capítulo 5.

Os modelos hierárquicos são métodos conhecidos que permitem melhorar a qualidade dum modelo quando se tem informação a vários níveis de agregação (por exemplo, complementado as previsões das vendas nas várias regiões com um modelo do total do país), mas a aplicação a quotas não é de todo trivial. Estudamos no Capítulo 6 uma forma de o fazer.

Os modelos de escolha utilizam um método de Monte Carlo para estimar parâmetros na função utilidade dos consumidores e foram alvo dum prémio Nobel em 2000. São modelos recentes que costumam ser aplicados para vendas (Chen e Yang, 2007); esta foi uma tentativa de os utilizar em quotas. A nossa implementação foi desenvolvida em R.

Visto que a medida em estudo é fração de vendas brutas, e é possível obter o seu histórico, seria possível abordar este problema desenvolvendo um modelo de previsão para as vendas da worten e um modelo de previsão para as vendas totais do mercado. De facto, numa fase inicial esta metodologia foi também testada mas os erros obtidos eram superiores.

Contribuições. Para além duma exploração horizontal da aplicação de várias técnicas à previsão das quotas de mercado, realçamos os seguintes pontos originais:

- A utilização de modelos hierárquicos para quotas, cuja aplicação costuma resumir-se a vendas: foi feita esta aplicação porque verificamos que os pesos da quota por região se mantiveram bastante estacionários no período em estudo;
- A aplicação de modelos de escolha para a quota de mercado, num contexto mais genérico que uma loja.

Trabalho Futuro. Com informação adicional sobre os concorrentes seria possível extrair mais informação, e mais precisa, do modelo da atração ou do modelo de escolha. Por outro lado, com o auxílio de séries geoespaciais, seria interessante estudar o impacto da distância das lojas em relação aos focos urbanos através de modelos gravíticos como o modelo de Huff (Huff, 1964).

Poderiam ainda ser desenvolvidos métodos mais robustos para a estimação das vendas brutas da worten e das vendas brutas totais, obtendo posteriormente a quota.

Relativamente à análise do desempenho dos modelos poderia ser feita uma análise estatística dos erros cometidos, com o objetivo de avaliar de forma mais exata a diferença entre o seu desempenho nos dados.

Bibliografia

Akaike H (1998). "Information Theory and an Extension of the Maximum Likelihood Principle." Em "Selected Papers of Hirotugu Akaike," volume 71, pp. 199–213. Springer. doi: 10.1007/978-1-4612-1694-0{_}15. (pg. 36)

Aksoy S, Haralick RM (2001). "Feature normalization and likelihood-based similarity measures for image retrieval." *Pattern Recognition Letters*, **22**(5), 563–582. doi: 10.1016/S0167-8655(00)00112-4. (pg. 7)

Almeida LB (1997). "Multilayer perceptrons." *Handbook of Neural Computation,Oxford University Press,UK*, pp. 1–30. (pg. 43)

Appel A, Evangelisti C, Stein A (1983). "Animating Quantitative Maps with Cellular Automata." *IBM Technical Discovery Bulletin*. (pg. 67)

Bell DE, Keeney RL, Little JD (1975). "A Market Share Theorem." *Journal of Marketing Research*, **12**(2), 136–141. (pg. 19)

Ben-Akiva M, Bierlaire M (1999). "Discrete Choice Methods and their Applications to Short Term Travel Decisions." Em "Handbook of Transportation Science. Kluwer," pp. 5–33. Springer. doi: 10.1007/978-1-4615-5203-1{_}2. (pg. 20)

Brockwell PJ, Davis RA (1987). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer New York. doi: 10.1007/978-1-4899-0004-3. (pg. 21)

Brockwell PJ, Richard A Davis (2002). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer New York. doi: 10.1007/b97391. (pg. 23), (pg. 24), (pg. 32)

Chen Y, Yang S (2007). "Estimating Disaggregate Models Using Aggregate Data Through Augmentation of Individual Choice." *Journal of Marketing Research*, **44**(4), 613–621. doi: 10.1509/jmkr.44.4.613. (pg. 2), (pg. 20), (pg. 52), (pg. 53), (pg. 55), (pg. 57), (pg. 69)

Chib S, Greenberg E (1995). "Understanding the Metropolis-Hastings Algorithm." *The American Statistician*, **49**(4), 327–335. doi: 10.1080/00031305.1995.10476177. (pg. 56)

Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990). "STL: A seasonal-trend decomposition procedure based on loess." *Journal of Official Statistics*, **6**(1), 3–73. doi: 10.1016/j.jnca.2015.06.008. (pg. 25), (pg. 26)

Cleveland WS (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association*, **74**(368), 829–836. doi: 10.2307/2683591. (pg. 25)

Cooper LG, Nakanishi M (1988). *Market-Share Analysis*. Springer. doi: 10.13140/2.1.1004.6402. (pg. 2), (pg. 19), (pg. 49), (pg. 50), (pg. 51), (pg. 69)

Cowpertwait PS, Metcalfe AV (2009). "State Space Models." Em Springer (ed.), "Introductory Time Series with R," pp. 229–246. Springer New York. doi: 10.1007/978-0-387-88698-5{_}12. (pg. 22), (pg. 28), (pg. 29)

Cryer JD, Chan KS (2008). *Time Series Analysis*. Springer Texts in Statistics. Springer New York. doi: 10.1007/978-0-387-75959-3. (pg. 22), (pg. 32)

Dougenik JA, Chrisman NR, Niemeyer DR (1985). "An Algorithm to Construct Continuous Area Cartograms." *The Professional Geographer*, **37**(1), 75–81. doi: 10.1111/j.0033-0124.1985.00075.x. (pg. 67)

Faraway JJ (2006). *Extending the linear model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC. (pg. 37), (pg. 44)

Fritsch S, Guenther F, following earlier work by Marc Suling (2012). *neuralnet: Training of neural networks*. R package version 1.32. (pg. 44)

Gastner MT, Newman MEJ (2004). "From The Cover: Diffusion-based method for producing density-equalizing maps." *Proceedings of the National Academy of Sciences of the United States of America*, **101**(20), 7499–7504. doi: 10.1073/pnas.0400280101. (pg. 67)

Granger C (1979). "Seasonality: causation, interpretation, and implications." Em A Zellner (ed.), "Seasonal Analysis of Economic Time Series," pp. 33–56. NBER. (pg. 24)

Greene WH (2011). *Econometric Analysis*. Pearson Education Limited (Verlag). (pg. 15)

Gujarati DN (2004). *Basic Econometrics*. McGraw Hill. (pg. 30), (pg. 31), (pg. 36)

Haitovsky Y (1969). "Multicollinearity in regression analysis: Comment." *The Review of economics and statistics*, pp. 92–107. doi: 10.2307/1926450. (pg. 11)

Hand D, Mannila H, Smyth P (2000). *Principles of data mining*. Cambridge University Press. (pg. 40)

Hornik K (1991). "Approximation capabilities of multilayer feedforward networks." *Neural Networks*, **4**(2), 251–257. doi: 10.1016/0893-6080(91)90009-T. (pg. 44), (pg. 69)

Hotelling H (1929). "Stability in Competition." *The Economic Journal*, **39**(153), 41. doi: 10.2307/2224214. (pg. 17)

Huff DL (1964). "Defining and Estimating a Trading Area." *Journal of Marketing*, **28**(3), 34. doi: 10.2307/1249154. (pg. 17), (pg. 70)

Huff DL (2003). "Parameter Estimation in the Huff Model." *ArcUser*, (October-December 2003), 3. (pg. 18)

Hyndman RJ (2014). "Forecasting: Principles & Practice." URL: http://robjhyndman.com/uwafiles/fpp-notes.pdf. (pg. 14)

Hyndman RJ, Ahmed Ra, Athanasopoulos G, Shang HL (2011). "Optimal combination forecasts for hierarchical time series." *Computational Statistics & Data Analysis*, **55**(9), 2579–2589. doi: 10.1016/j.csda.2011.03.006. (pg. 64)

IGEO (2015). "Mapa LAU de Portugal, Instituto Geográfico Português." URL: http://www.gadm.org, visitado a 15 de janeiro de 2015. (pg. 67)

INE (2011). "Poder de compra per capita por Localização geográfica (NUTS - 2002); Bienal - INE." (pg. 5), (pg. 7)

INE (2013). "População média anual residente (Nº) por Local de residência (NUTS - 2013). Sexo e idade; Anual - INE, Estimativas Anuais da População Residente." (pg. 5), (pg. 7)

Kohavi R (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." Em "IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2," pp. 1137–1143. Morgan Kaufmann Publishers Inc. (pg. 15)

Kotler P (1984). *Marketing Management: Analysis, Planning and Control*. Prentice-Hall. (pg. 18)

Kwiatkowski D, Phillips PC, Schmidt P, Shin Y (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root." *Journal of Econometrics*, **54**(1-3), 159–178. doi: 10.1016/0304-4076(92)90104-Y. (pg. 31)

Leeflang PSH, Reuyl JC (1984). "On the Predictive Power of Market Share Attraction Models." *Journal of Marketing Research*, **21**(2), 211. doi: 10.2307/3151703. (pg. 17)

Lira SA, Neto AC (2006). "Coeficientes de correlação para variáveis ordinais e dicotómicas derivados do coeficiente linear de pearson." *Ciência & Engenharia*, **15**, 45–53. (pg. 11)

Meyer D (2014). "Support Vector Machines: The Interface to libsvm in package e1071." (pg. 47)

Naert P, Weverbergh M (1981). "On the Prediction Power of Market Share Attraction Models." *Journal of Marketing Research*, **18**(2), 146. doi: 10.2307/3150949. (pg. 51)

Nalbantov GI, Franses PH, Groenen PJF, Bioch JC (2010). "Estimating the Market Share Attraction Model using Support Vector Regressions." *Econometric Reviews*, **29**(5-6), 688–716. doi: 10.1080/07474938.2010.481989. (pg. 51)

Philips PCB, Perron P (1988). "Testing for a unit root in time series regression." *Biometrika*, **75**(2), 335–346. doi: 10.1093/biomet/75.2.335. (pg. 31)

QGIS (2009). "Geographic Information System." URL: http://qgis.osgeo.org. (pg. 67)

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revelle W (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.5.6. (pg. 12)

Rizopoulos D (2013). *ltm: Latent Trait Models under IRT*. R package version 1.0-0. (pg. 11)

Sarma DD (2009). *Geostatistics with Applications in Earth Sciences*. Springer. doi: 10.1007/978-1-4020-9380-7. (pg. 18)

Seltman HJ (2015). "Experimental Design and Analysis." URL: http://www.stat.cmu.edu/{~}hseltman/309/Book/Book.pdf. (pg. 36)

Shumway RH, Stoffer DS (2011). *Time Series Analysis and Its Applications With R Examples*. Springer. doi: 10.1007/978-1-4419-7865-3. (pg. 29), (pg. 31)

Smola AJ, Sch B, Schölkopf B (2004). "A Tutorial on Support Vector Regression." *Statistics and Computing*, **14**(3), 199–222. doi: 10.1023/B:STCO.0000035301.49549.88. (pg. 46)

Tobler WR (1973). "A Continuous Transformation Useful for Districting." *Annals of the New York Academy of Sciences*, **219**(1), 215–220. doi: 10.1111/j.1749-6632.1973.tb41401.x. (pg. 67)

Trapletti A, Hornik K (2013). *tseries: Time series analysis and computational finance*. R package version 0.10-32. (pg. 31)

Viton PA (2010). "Derivation of the Logit Choice Probabilities." URL: http://facweb.knowlton.ohio-state.edu/pviton/courses2/crp5700/logit.pdf. (pg. 53)