MARGARIDA PEREIRA FERREIRA

PREVISÃO HORÁRIA DO NÚMERO DE ADMISSÕES NUM SERVIÇO DE URGÊNCIA

FORECASTING HOURLY ADMISSIONS IN AN EMERGENCY DEPARTMENT

MARGARIDA PEREIRA FERREIRA

PREVISÃO HORÁRIA DO NÚMERO DE ADMISSÕES NUM SERVIÇO DE URGÊNCIA

FORECASTING HOURLY ADMISSIONS IN AN EMERGENCY DEPARTMENT

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica do Doutor Luís Silva, Professor Auxiliar Convidado do Departamento de Matemática da Universidade de Aveiro.

o júri / the jury

presidente / president

Prof. Doutor Eugénio Alexandre Miguel Rocha
Professor Auxiliar da Universidade de Aveiro

vogais / examiners committee

Prof. Doutora Isabel Maria Marques da Silva Magalhães
Professora Auxiliar do Departamento de Engenharia Civil da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor Luís Miguel Almeida da Silva
Professor Auxiliar Convidado da Universidade de Aveiro (orientador da UA)

agradecimentos /
acknowledgements

Quero deixar um grande e caloroso agradecimento aos meus pais e aos meus irmãos que sempre me apoiaram, mesmo longe, e sem eles nada disto seria possível. À minha madrinha, Vera, que nunca deixou de se preocupar e que representa um grande exemplo para mim.

Ao Fábio, o meu verdadeiro suporte ao longo dos últimos anos, pela paciência, compreensão e empenho em motivar-me.

À Inês, à Rita e à Letícia pelas conversas de desespero e descompressão e ao Cláudio pela preocupação.

Como não podia deixar de ser, quero deixar o meu agradecimento especial ao Professor Luís Silva, meu orientador, pela paciência que teve comigo e por toda a ajuda que me prestou. Por último mas não menos importante, agradeço a todos os colegas da Prologica, nomeadamente ao Diogo Reis pela oportunidade e ao Bernardo, ao Diogo Costa e ao Paulo pela ajuda que sempre disponibilizaram.

Resumo

Na triagem dos Serviços de Urgência, o Protocolo de Triagem de Manchester define tempos de espera para o atendimento para cada uma das prioridades que, no entanto, são muitas vezes excedidos. Isso pode revelar, para além de outras causas, uma gestão de recursos menos boa e, consequentemente, a necessidade de esta ser apoiada por previsões do número de admissões.

Neste trabalho propõe-se um modelo de previsão horária do número de admissões para uma janela temporal de 10 dias. Por ser uma previsão horária, ou seja, bastante fina do ponto de vista temporal, esta ferramenta revela-se de extrema utilidade no apoio à tomada de decisão num serviço de urgência. Para a construção deste modelo são usados dados reais de uma Unidade de Saúde portuguesa, de janeiro de 2014 a outubro de 2018.

Os modelos mais utilizados na literatura para lidar com este tipo de problemas são os modelos clássicos de regressão ou os modelos clássicos para séries temporais (nomeadamente ARIMA), uma vez que o número de admissões horárias é um conjunto de observações registadas ao longo do tempo.

Para uma modelação mais versátil e adequada a séries temporais de contagem, recorreu-se a modelos lineares generalizados para séries temporais de contagem. Com esta abordagem houve a possibilidade de efetuar mais escolhas acerca das componentes a incluir no modelo e que efetivamente influenciam o número de admissões. É o caso de componentes relativas à série, como observações passadas ou médias passadas, e de componentes externas, chamadas co-variáveis, que complementam a informação da série temporal.

Com um RMSE de 5.5 admissões horárias, os resultados mostram o potencial da abordagem proposta no apoio à tomada de decisão, quer a nível de recursos humanos quer materiais, de tal forma que a empresa de acolhimento incluiu, na sua plataforma Meliora, um *dashboard* com os resultados do modelo e que já está a se utilizado pela Unidade de Saúde.

Abstract

At emergency department's triage, the Manchester Triage Protocol defines waiting times to receive medical care for each priority which, however, are often exceeded. This may be a consequence of a not such good resources management, among others, revealing a necessity of automated admission forecast systems.

This work proposes an hourly forecast model of the number of admissions for a 10-day time window. Because it's a hourly forecast, it means that it's very detailed from a temporal point of view, this tool is extremely useful to support decision making in an emergency department. For the construction of this model, real data of one Portuguese Health Unit is used, from January 2014 to October 2018.

The most used models in the literature to deal with this type of problem are the classic regression models or the classic models for time series (namely ARIMA), since the number of hourly admissions is a set of observations recorded over time.

For a more versatile modeling and more appropriate to count time series, generalized linear models for count time series were used. With this approach, there was the possibility of making more choices about components to include in the model and which effectively influence the number of admissions. It's the case of components related to the series, such as past observations or past avarages, and external components, called covariables, that complement the information of the time series.

With an RMSE of 5.5 hourly admissions, the results show the potential of the proposed approach to support decision making, for human and material resources management, in such a way that the host company included, in its Meliora platform, a dashboard with the results of the model that is already being used by the Health Unit.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Enquadramento e motivação

Dadas as opções de dissertação, projeto ou estágio, a escolha de estágio deveu-se ao interesse no mercado de trabalho, tendo em vista um primeiro contacto que proporciona alguma experiência empresarial. Estando inserido em ambiente empresarial, o estágio permite o desenvolvimento de *soft skills* como a polivalência, a autonomia e o trabalho de grupo.

O estágio desenrolou-se na Prologica, uma empresa portuguesa de renome, pioneira no desenvolvimento de tecnologias baseadas em conhecimento extraído dos dados (mais informações na secção 1.1.1).

O tema proposto, na área da saúde, está inserido num projeto da empresa. As entidades de saúde recolhem diariamente uma quantidade de dados enorme que têm a necessidade de ver utilizada de forma a melhorar os seus serviços.

O trabalho desenvolvido no âmbito do estágio teve como base dados reais de uma Unidade Local de Saúde (ULS) portuguesa, que é uma unidade de saúde mais abrangente do que um só hospital (mais informações na secção 1.1.2). Para efeitos de proteção dos dados, ao longo do documento não serão referidas nem evidenciadas quaisquer informações sobre a entidade de saúde, os seus profissionais ou utentes.

1.1.1 Prologica

A Prologica é uma empresa fundada em 1984, com sede em Lisboa e escritório de Investigação e Desenvolvimento em São João da Madeira. Tendo como objetivo o uso adequado da tecnologia, explorando os dados e tirando o melhor partido deles, implementa soluções nas áreas da saúde, educação e cidadania. Estas soluções visam o suporte à decisão, para uma melhor prevenção e monitorização. Com uma forte presença em mercados desenvolvidos e emergentes da Europa, África e América, a Prologica tornou-se uma referência de qualidade, inovação e excelência no setor público e privado, contando já com várias distinções. É constituída por uma equipa acolhedora, jovem e proativa que trabalha com tecnologias recentes e que procura acompanhar os avanços das mesmas.

O escritório de São João da Madeira, onde decorreu o estágio, está dividido em duas secções: secção dos dados e secção de desenvolvimento. A secção dos dados, na qual me inseri, é composta por *data engineers* e *data scientists*. Grande parte do trabalho desenvolvido aqui é no âmbito da saúde e está inserido no Meliora, uma plataforma de conhecimento e tecnologia de suporte à decisão da Prologica. Contém dados, algoritmos e ferramentas para proporcionar às organizações e seus colaboradores, as respostas que procuram para melhorar continuadamente as suas operações (Prologica, 2018).

1.1.2 Entidades de Saúde

O Estado Português assegura o direito à saúde a todos os cidadãos de Portugal através do Serviço Nacional de Saúde (SNS). Esta estrutura, que engloba entidades desde o Ministro da Saúde às Unidades

Básicas de Saúde, está organizada de forma hierárquica, como mostra a Figura 1.1 (Serviço Nacional de Saúde, 2017).

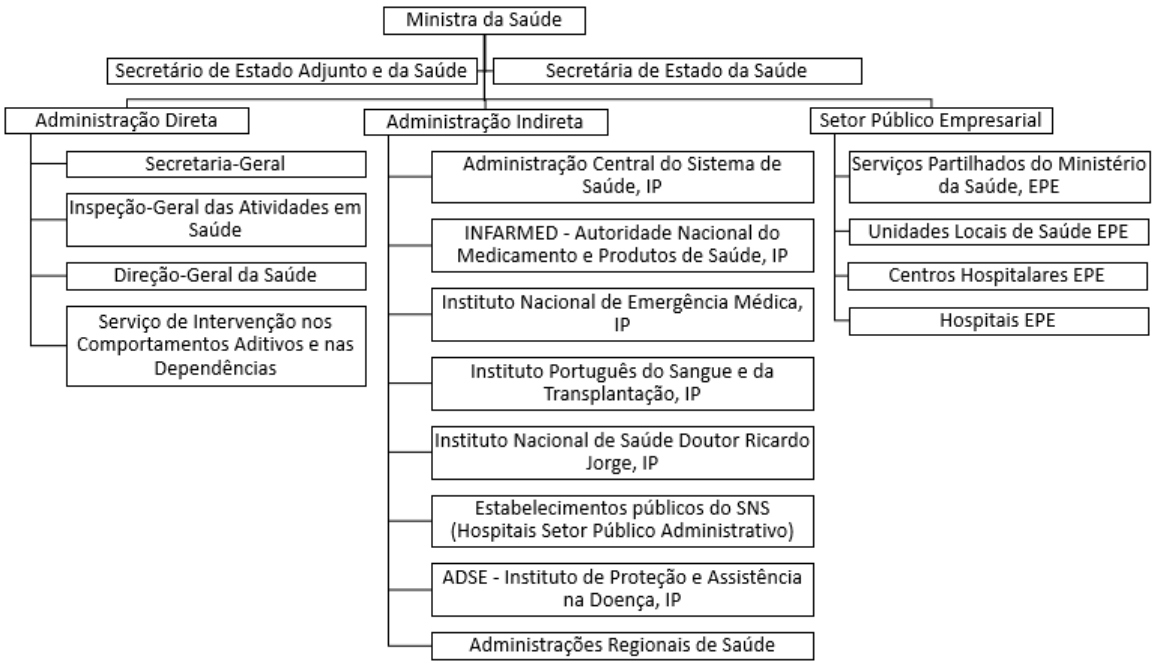


Figura 1.1: Esquema hierárquico do Serviço Nacional de Saúde de Portugal

No setor empresarial estão incluídas diversas Entidades Públicas Empresariais (EPE) entre as quais constam as que se ocupam da prestação de cuidados de saúde, como é o caso de Unidades Locais de Saúde, Centros Hospitalares e Hospitais.

Como referido, alguns hospitais não são abrangidos por entidades maiores mas existem outros que fazem parte de Centros Hospitalares ou de Unidades Locais de Saúde (ULS). Os Centros Hospitalares são compostos por Hospitais. Já as ULS são compostas maioritariamente por Hospitais e Centros de Saúde (agrupados em Agrupamentos de Centros de Saúde (ACES)), tendo também integradas outras unidades de saúde de modo a abranger os diferentes tipos de cuidados de saúde - ver Figura 1.2.

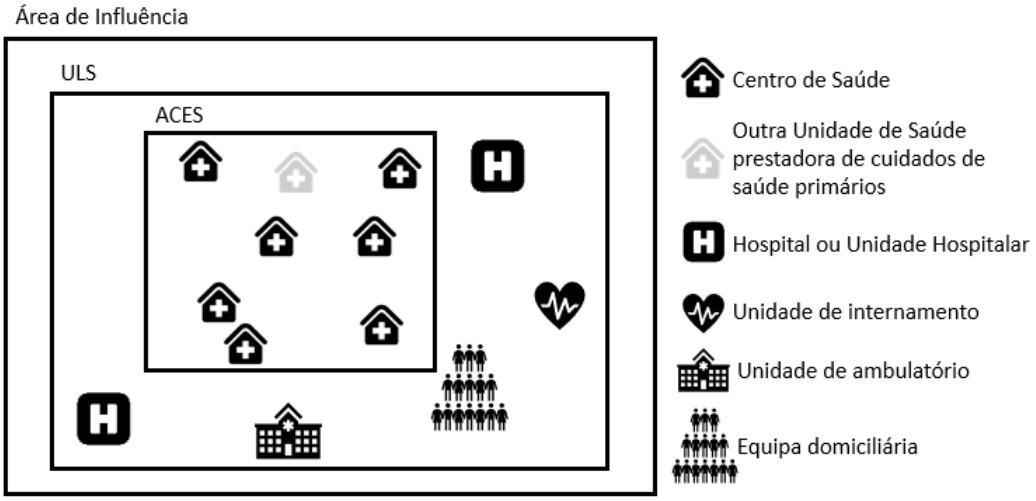


Figura 1.2: Ilustração da composição de uma ULS

As ULS são unidades de saúde com uma área de influência específica, muitas vezes um distrito, que

começaram a ser idealizadas em 1999, sendo distribuídas pelo país como mostra a Figura 1.3 (Entidade Reguladora da Saúde, 2011; Serviço Nacional de Saúde, 2017).

Figura 1.3: Distribuição das ULS, nome e respetivo ano de criação, em Portugal

Estas unidades prestam cuidados de saúde primários, diferenciados e continuados (Serviço Nacional de Saúde, 2017).

Os cuidados de saúde primários são aqueles que são essenciais e, por isso, representam o primeiro nível de contacto dos indivíduos, da família e da comunidade com o SNS. Assim, é conveniente que estes cuidados sejam levados mais proximamente possível aos lugares onde as pessoas vivem e trabalham, usualmente em Centros de Saúde (Entidade Reguladora da Saúde, 2011).

Os cuidados de saúde diferenciados ou hospitalares podem ser definidos como o conjunto de atividades de prevenção, promoção, restabelecimento ou manutenção da saúde, bem como de diagnóstico, tratamento e reabilitação, em ambiente hospitalar (Entidade Reguladora da Saúde, 2011).

Os cuidados de saúde continuados designam o conjunto de intervenções sequenciais de saúde que visam promover a autonomia melhorando a funcionalidade da pessoa em situação de dependência, através da sua reabilitação, readaptação e reinserção familiar e social (Entidade Reguladora da Saúde, 2011). Estes podem ser prestados em unidades de internamento, unidades de ambulatório e equipas domiciliárias ou hospitalares (Serviço Nacional de Saúde, 2017).

Os serviços de prestação de cuidados de saúde diferenciados são efetuados sob várias linhas de atividade, entre elas:

- Consulta Externa;
- Urgência e Emergência Médica;
- Bloco Operatório;
- Cirurgia de Ambulatório;
- Internamento;

- Hospital de Dia.

Em particular, o Serviço de Urgência funciona por etapas, como ilustra a Figura 1.4.

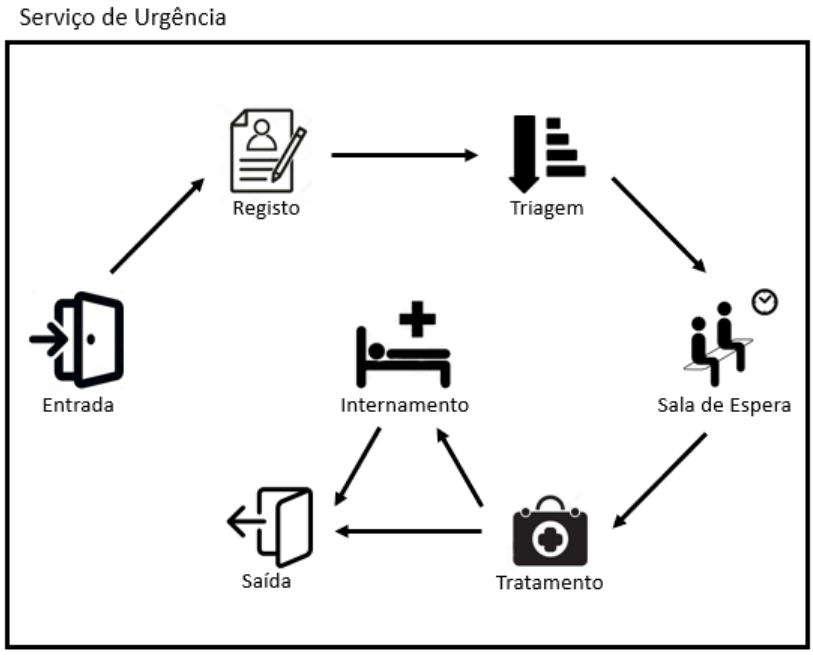


Figura 1.4: Ilustração do fluxo mais frequente no Serviço de Urgência de um Hospital (baseada na figura 1 de Zhao e Lie, 2010)

Geralmente, e numa fase inicial, existe um registo dos dados do utente, dando-se aquilo a que se chama a admissão.

Posteriormente, o utente passa por um processo de triagem no qual lhe é atribuída uma cor consoante a prioridade que detém.

Este processo baseia-se no Protocolo de Triagem de Manchester, introduzido em Portugal pelo Grupo Português de Triagem em 1999 e implementado atualmente a nível nacional (Grupo Português de Triagem, 2015).

Este Protocolo implica que a triagem seja realizada na presença do doente que recorre aos Serviços de Urgência, por um enfermeiro com formação em Triagem de Manchester, permitindo assim ao profissional recolher uma história detalhada sobre o motivo que traz o doente ao serviço de urgência e a recolha/medição de parâmetros fisiológicos apresentados. Desta forma, determina uma prioridade clínica que atribui a cor vermelha a casos de emergência, laranja a casos muito urgentes, amarelo a urgentes, verde a pouco urgentes e azul a não urgentes e, consequentemente, associa um tempo máximo previsto para a primeira observação médica (ver Figura 1.5).

Depois da triagem e do tempo de espera, o utente recebe os cuidados médicos e, habitualmente, procede à saída ou poderá ser internado caso se verifique necessário. O utente pode, em alguns casos, ser transferido para outra unidade de saúde mais adequada ou ser encaminhado para uma consulta.

Não obstante do fluxo normal de um utente no Serviço de Urgência, este não invalida o facto de haver utentes a abandonar o serviço sem os cuidados de saúde, podendo ser por motivos de tempo de espera excessivo e/ou outros.

De certo modo, a previsão do número de admissões relaciona-se com este problema, apoiando a decisão de alocação de recursos, esperando-se a minimização dos tempos de espera excessivos.

1.2 Plano de estágio

O estágio enquadra-se na aplicação de modelação de dados para previsão na realidade hospitalar.

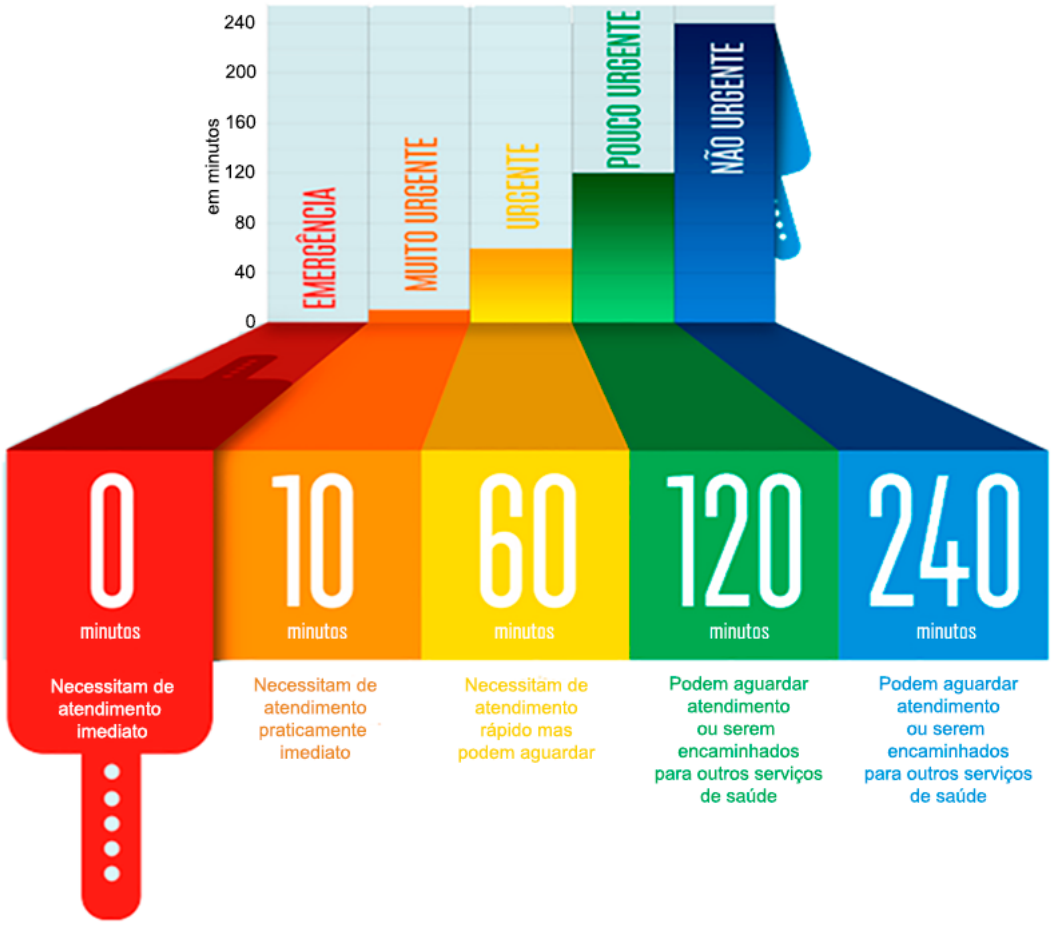


Figura 1.5: Tempo previsto de atendimento para as cinco cores da Triagem de Manchester (Grupo Português de Triagem, 2015)

O objetivo proposto pela Prologica prende-se com o estudo da aplicabilidade de metodologias de modelação para a previsão do número de admissões no Serviço de Urgência por dia e hora, para um período de 10 dias, e com o desenvolvimento de protótipos para endereçar o problema.

O desenvolvimento do trabalho encontra-se dividido por várias etapas encadeadas logicamente.

Primeiramente, procedeu-se ao estudo sobre o contexto da estatística em saúde, com principal foco em modelação e algoritmos de distribuição de contagens.

Como complemento ao descrito anteriormente, analisou-se o estado da arte sobre a aplicação de modelos preditivos e aprendizagem computacional no âmbito das Urgências.

Aquando da receção dos dados de uma entidade de saúde, procedeu-se à análise dos mesmos e relacionamento com dados externos. Para isso, houve um pré-processamento dos dados e sua reestruturação.

Posteriormente, prepararam-se os dados para o desenvolvimento das metodologias selecionadas e efetuaram-se os testes às mesmas e a validação da solução.

Por último, elaborou-se uma proposta de *front-end* para o utilizador final, ou seja, a forma como são apresentados os resultados aos utilizadores e como estes podem interagir com os parâmetros.

1.3 Estrutura da Dissertação

O documento segue uma estrutura baseada nas etapas propostas no Plano de Estágio e no próprio trabalho desenvolvido.

No segundo capítulo, iniciado pelo problema que o trabalho visa colmatar, apresenta-se com uma aná-

lise de artigos escritos com trabalhos dentro do mesmo objetivo deste e uma introdução das metodologias mais frequentemente usadas e da necessária para este trabalho.

O caso de estudo é descrito no terceiro capítulo, que inclui descrição, processamento e análise dos dados e também as implementações para a previsão do número de admissões num Serviço de Urgências.

É apresentado, no quarto capítulo, o resultado final do trabalho, que consiste na construção de um *dashboard* no Meliora e todo o procedimento conceptual que o antecedeu.

Por último, no quinto capítulo, finaliza-se com as conclusões, com inclusão de discussão do trabalho, suas limitações, trabalho futuro e considerações finais.

Todo o relatório é complementado com as referências acedidas ao longo do trabalho, que se situam no final do relatório.

Capítulo 2

Revisão da Literatura

No que diz respeito à primeira fase dos trabalhos, é importante explorar e avaliar trabalho no mesmo âmbito e as metodologias neles usadas.

Já inúmeros estudos se debruçaram sobre a previsão das admissões nas urgências, dada a problemática à volta desta questão. É, de facto, importante conseguir-se usar técnicas de previsão para melhorar os sistemas que usamos no dia-a-dia, nomeadamente os sistemas de saúde, que são tão imprescindíveis. Existe uma procura incessante da melhor forma de prever as admissões nas urgências, no que diz respeito a modelos, variáveis externas à realidade hospitalar e dimensões de amostras, entre outras.

Neste capítulo introduz-se o tema da alocação dos recursos no Serviço de Urgência, apresenta-se um estudo bibliográfico de artigos com propósitos parecidos ao deste estudo e exploram-se várias metodologias frequentemente usadas.

2.1 Serviço de Urgência: Recursos

Cada vez mais os profissionais responsáveis pela gestão de recursos dos Serviços de Urgência (SU) estão obrigados a ter cuidados redobrados. Em Portugal, a ideia generalizada de que uma melhor acessibilidade aos cuidados de saúde primários, nomeadamente ao médico de família, levaria a uma diminuição das admissões nas urgências hospitalares, está a ser analisada. De facto, os SU estão constantemente a ultrapassar as suas capacidades, denunciando algumas falhas, ou na própria gestão dos SU, ou nas unidades básicas de saúde (Sá, 2002).

Em particular, as ULS devem, entre outras ações, ajustar os recursos disponíveis às necessidades de saúde, como é previsto nos Termos de Referência para contratualização de cuidados de saúde do SNS para 2019 (Administração Central do Sistema de Saúde, 2018).

Ora, este ajuste de recursos nem sempre é simples pois, se por um lado a experiência consegue ajudar a prever algumas situações, por outro existem situações dificilmente previstas. Assim, a evolução tecnológica veio trazer mais e diferentes formas de conseguir monitorizar o passado e perspetivar o futuro através do estudo e modelação dos dados.

2.2 Estado da Arte

Muitos artigos foram analisados na procura de escolhas mais acertadas e diversificadas, consoante os estudos já efetuados. Nesse sentido, exploram-se os artigos do ponto de vista dos objetivos e dados usados (quantidade de dados e tipos de variáveis), das metodologias usadas e das conclusões obtidas.

2.2.1 Objetivos e Dados

O objetivo dos estudos à volta da previsão do número de admissões num Serviço de Urgência difere um pouco. Ao contrário de muitos dos estudos, cujo objetivo é a previsão diária das admissões (Tabela 2.1), este estudo objetiva a previsão das admissões por hora, a cada dia. As previsões mais finas, hora a

hora, são mais vantajosas do ponto de vista da gestão de recursos pois a distribuição das admissões varia ao longo do dia e os recursos são alocados em função dessa variação.

Por outro lado, a própria estrutura de dados também é diferente.

Existem estudos que se baseiam em 6 anos de dados (Jones, 2007) e outros que apenas têm como base 1 ano de dados (McCarthy et al., 2008). Esta dimensão, relativa ao número de observações, tem impacto no treino do modelo, podendo originar *overfitting*[1] ou *underfitting*[2].

Algumas abordagens, ao contrário da maioria, optam por modelar dados provenientes de hospitais distintos de modo a tornar os modelos mais gerais e aplicáveis a qualquer hospital (Boyle, Jessup et al., 2011; Boyle, Wallis et al., 2008). Um modelo geral seria, sem dúvida, ideal para o SNS, de modo a estar apto para o uso de qualquer hospital deste sistema. No entanto, quanto mais específico puder ser o modelo em relação à realidade do hospital, melhor desempenho terá.

Uma das questões mais importantes é relativa às variáveis. Por ser uma previsão que não se pode basear em características da entidade de saúde nem dos utentes, a base é sempre a contagem de admissões num dado SU. Coloca-se a questão: "O que poderá influenciar e explicar este número de admissões?".

Muitos são aqueles que incluem no seu estudo variáveis de calendário - dias da semana, meses, estações do ano, feriados - (Batal et al., 2001; McCarthy et al., 2008; Jones et al., 2008). É fácil perceber o impacto que estas variáveis podem ter no número de admissões; basta verificar que a quantidade de gripes no inverno é superior ou que depois de um feriado festivo, como o Carnaval, a afluência ao SU é maior.

Uma outra questão importante, que bastantes estudos exploram, são os fatores meteorológicos (Batal et al., 2001; Díaz et al., 2001; McCarthy et al., 2008; Jones et al., 2008; Calegari et al., 2016) e ambientais (Díaz et al., 2001). Informação como a temperatura, a precipitação, a humidade ou a poluição podem ajudar a explicar a variação do número de admissões.

Outros estudos com objetivos mais específicos, como a previsão de camas ocupadas, entram em conta com fatores populacionais ou hospitalares. Todo o estudo da literatura descrito acima está sumariado na Tabela 2.1.

2.2.2 Metodologias

Sendo que a previsão a ser feita é quantitativa e pode ser explicada por algumas variáveis externas, um tipo de metodologias usado é modelos lineares (Batal et al., 2001; Boyle, Wallis et al., 2008; McCarthy et al., 2008). Este tipo de modelos é obtido através da interação de outras variáveis com a variável resposta[3].

Para estudos semelhantes a este, como as admissões são registadas ao longo de instantes de tempo, a distribuição de contagem destas admissões pode ser interpretada como uma série temporal (consultar mais sobre o conceito na Secção 2.3.2). Desta forma, são utilizadas outras metodologias adequadas a séries temporais, como modelos Auto-regressivos Integrados e de Médias Móveis (ARIMA[4]), ou particularizações dos mesmos, que permitem criar um modelo sobre a própria série temporal usando observações e médias passadas, e amaciamento exponencial (Díaz et al., 2001; Champion et al., 2007; Jones et al., 2008).

As metodologias anteriormente mencionadas são as clássicas da Estatística. Embora não seja usual a utilização de métodos de aprendizagem computacional, alguns estudos exploram redes neuronais (Jones et al., 2008).

2.2.3 Conclusões

Apesar de muitos estudos considerarem que os fatores meteorológicos influenciam o número de admissões, a maior parte deles constata que as variáveis meteorológicas pioram a performance dos

[1] *Overfitting* é a palavra inglesa para designar o sobre-ajuste do modelo aos dados, ou seja, um ajuste excessivo no treino que se reflete num baixo desempenho com novos dados.

[2] *Underfitting*, ao contrário de *overfitting*, designa o sub-ajuste do modelo aos dados que origina um erro elevado no treino e no teste.

[3] Designação para a variável que se pretende prever.

[4] Do inglês, *Autoregressive Integrated Moving Average*.

modelos. Ao contrário destas, as variáveis relacionadas com o calendário são fortemente elogiadas pelo seu forte poder preditivo no número de admissões (Batal et al., 2001; McCarthy et al., 2008; Calegari et al., 2016).

Em relação a metodologias, não existe supremacia evidente ao nível do desempenho. Cada abordagem tem pontos fortes mas também limitações. Existe sim superioridade numérica de modelos temporais em relação aos modelos lineares clássicos.

Se por um lado alguns investigadores obtiveram bons resultados com modelos lineares (Boyle, Wallis et al., 2008; Jones et al., 2008; McCarthy et al., 2008; Marcilio, Hajat e Gouveia, 2013), outros conseguiram-no através de metodologias de séries temporais, por vezes superiorizando-se o amaciamento exponencial (Champion et al., 2007; Boyle, Jessup et al., 2011) e outras vezes os modelos ARIMA (Díaz-Hierro et al., 2012; Calegari et al., 2016).

2.3 Modelos de Previsão

É importante ter em atenção alguns conceitos e teoria por detrás desta temática. De facto, as admissões podem ou não ser interpretadas como sendo uma série temporal e, consoante isso, ser estudadas e modeladas através de diferentes metodologias. Neste capítulo, são apresentados brevemente os métodos mais usados na literatura, referidos no capítulo anterior, e, posteriormente, é introduzido o modelo utilizado - Modelos Lineares Generalizados para Séries Temporais de Contagem.

2.3.1 Modelos de Regressão Clássicos

As regressões são modelos estatísticos para modelar a relação entre variáveis (Montgomery, Peck e Vining, 2006).

Para o efeito, interessa referir particularmente dois tipos de regressão: Regressão Linear Múltipla e Regressão de Poisson. Ambos os modelos pertencem a um grande grupo de modelos designado por Modelos Lineares Generalizados (GLM[5]).

Os GLM têm a particularidade de ter como pressuposto que a distribuição da variável resposta pertence à família exponencial, ou seja, a sua função densidade de probabilidade pode escrever-se na forma

$$f(y,\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

onde θ é parâmetro e $a,\ b,\ c$ e d são funções conhecidas, ou, de forma equivalente,

$$f(y,\theta) = s(y)t(\theta)e^{a(y)b(\theta)},$$

para $t(\theta) = e^{c(\theta)}$ e $s(y) = e^{d(y)}$.

Em particular, com $s(y) = (y!)^{-1}$, $t(\theta) = e^{-\theta}$, $a(y) = y$ e $b(\theta) = \log(\theta)$ obtém-se

$$f(y,\theta) = (y!)^{-1}e^{-\theta}e^{y\log(\theta)} \Leftrightarrow f(y,\theta) = \frac{e^{-\theta}\theta^y}{y!} \tag{2.1}$$

- função massa de probabilidade da distribuição de Poisson, distribuição da variável resposta de uma regressão de Poisson - enquanto que, com $a(y) = y$, $b(\theta) = \frac{\theta}{\sigma^2}$, $c(\theta) = -\frac{\theta^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$ e $d(y) = -\frac{y^2}{2\sigma^2}$, se obtém

$$f(y,\theta) = \exp\left[y\frac{\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right] \Leftrightarrow f(y,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\theta)^2}{2\sigma^2}} \tag{2.2}$$

- função densidade de probabilidade da distribuição Normal, distribuição da variável resposta da regressão linear (Figura 2.1). Estas distribuições distinguem-se logo à partida pela natureza da variável que descrevem - a distribuição Normal descreve uma variável contínua (com valores infinitos e incontáveis) e a distribuição de Poisson descreve variáveis discretas (com valores contáveis, inteiros não negativos) - o que pode ser um fator de escolha da regressão.

[5]Do inglês, *Generalized Linear Models*.

Artigo	Objetivo	Metodologias	Variáveis	Conclusões
(Batal et al., 2001)	Admissões diárias	Regressão linear	Calendário, Meteorológicas	Variáveis de calendário são significativas e variáveis meteorológicas são pouco relevantes.
(Díaz et al., 2001)	Admissões diárias	Correlação cruzada, ARIMA	Meteorológicas, Ambientais	As admissões estão relacionadas com a temperatura, a poluição e concentração de ozono.
(Champion et al., 2007)	Admissões mensais	Amaciamento exponencial, ARIMA		Amaciamento exponencial prevê algumas quedas que o ARIMA não prevê.
(Jones, 2007)	Admissões horárias	Auto-regressão vetorial		A flexibilidade do modelo permite que seja capaz de ter uma boa performance.
(Boyle, Wallis et al., 2008)	Admissões mensais	Regressões		O melhor método é aquele que tem menor erro absoluto médio no teste.
(McCarthy et al., 2008)	Admissões horárias	Regressão de Poisson	Calendário, Meteorológicas	Variáveis meteorológicas não têm efeito significativo.
(Jones et al., 2008)	Admissões diárias	Regressões, ARIMA, Amaciamento exponencial, Redes neuronais artificiais	Calendário, Meteorológicas	A regressão de séries temporais acrescenta pouco em relação à regressão linear com variáveis de calendário.
(Boyle, Jessup et al., 2011)	Admissões horárias	Regressão múltipla, ARIMA, Amaciamento exponencial	Calendário	O ARIMA é preferível quando se considera informação mais recente.
(Díaz-Hierro et al., 2012)	Admissões horárias	ARIMA, Outros de séries temporais	Meteorológicas, Ambientais, Populacionais	ARIMA alerta para mudanças inesperadas. Modelos com variáveis exógenas podem ser melhores.
(Marcilio, Hajat e Gouveia, 2013)	Admissões diárias	Modelos de regressão lineares, Estimação de equações, ARIMA	Calendário, Meteorológicas	Informação da temperatura não influencia a performance do modelo. Modelos de regressão lineares e de estimação são melhores do que os ARIMA.
(Calegari et al., 2016)	Admissões diárias	Modelos baseados em séries temporais	Meteorológicas	A melhor performance foi do ARIMA e os fatores meteorológicos não melhoram o modelo.

Tabela 2.1: Sumário dos artigos cujos temas são semelhantes ao presente estudo

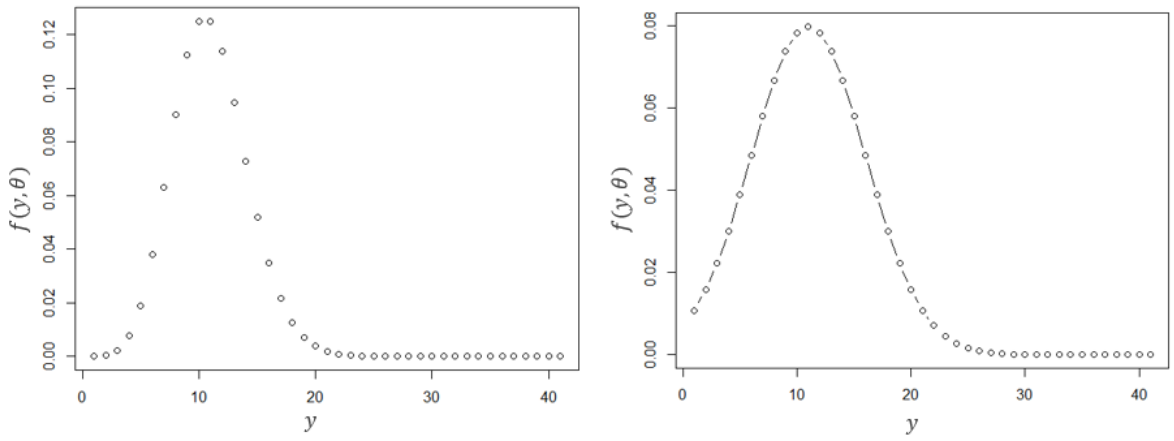


Figura 2.1: Exemplo de representação gráfica da função massa de probabilidade da distribuição de Poisson, com $\theta = 10$ (gráfico da esquerda) e da função densidade de probabilidade da distribuição Normal, com $\theta = 10$ e $\sigma = 5$ (gráfico da direita).

Regressão Linear Múltipla

O objetivo é descrever variável resposta y através de um conjunto de variáveis explicativas/preditivas, tal que

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \ ,$$

onde β_0 é o coeficiente independente, x_j denota cada uma das variáveis explicativas, β_j os respetivos parâmetros desconhecidos, com $j = 1, 2, ..., k$ em que k representa o número de variáveis explicativas e, por fim, ε representa a componente de erro não observável (muitas vezes designada de componente de ruído, representa todos os fatores que influenciam y mas não podem ser descritos pelas variáveis explicativas). Para um conjunto de observações ($\boldsymbol{x_i}$,y_i), $i = 1, ..., n$, o problema pode ser formulado matricialmente como

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \Leftrightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

O vetor de parâmetros $\boldsymbol{\beta}$ é estimado de modo que este modelo teórico se ajuste aos dados reais, obtendo-se assim o modelo de estimação $\hat{y}$ da variável resposta, para cada observação i,

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} \ , i = 1, ..., n \ .$$

O método clássico para estimação de $\boldsymbol{\beta}$ é o Método dos Mínimos Quadrados que resulta do Método de Máxima Verosimilhança assumindo que $\varepsilon \sim N(0, \sigma^2)$.

Desta forma, a função de verosimilhança, baseada na função f de 2.2, é dada por

$$L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}.$$

Naturalmente, maximizar $L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2)$ é equivalente a maximizar

$$\ln L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln(2\pi) - n \ln(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

e, para um valor fixo de σ, a maximização prende-se apenas em minimizar o termo

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \ .$$

Ora,

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

e obtém-se assim o Método dos Mínimos Quadrados. Seja

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 .$$

Então os estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ devem satisfazer

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) = 0$$

e

$$\frac{\partial S}{\partial \beta_j} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right)x_{ij} = 0,\ j = 1, 2, \ldots, k ,$$

as chamadas equações normais.

Na forma matricial, a forma fechada[6] da solução deste método é dada por

$$\begin{aligned}\frac{\partial S}{\partial \boldsymbol{\beta}} =& -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0\\ &\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}\\ &(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\ &\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} .\end{aligned}$$

Assim, sabendo os estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, obtém-se o estimador para a variável resposta, $\hat{y}$,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k .$$

Regressão de Poisson

No caso da regressão de Poisson, assume-se que a variável resposta y segue então uma distribuição de Poisson, que representa normalmente uma contagem de um qualquer evento (Montgomery, Peck e Vining, 2006).

O modelo de Poisson pode ser escrito como

$$y_i = E(y_i) + \varepsilon_i,\ \ i = 1, 2, \ldots, n ,$$

em que $E(y_i)$ denota o valor esperado dos valores observados e, assumindo que $E(y_i) = \mu_i$, assume-se que existe uma função de ligação g tal que

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k = \begin{bmatrix}1 & x_1 & \ldots & x_k\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_k\end{bmatrix} = \mathrm{x}_i'\boldsymbol{\beta} .$$

A função de ligação g associa os valores esperados da variável resposta às variáveis explicativas no modelo. Uma função de ligação transforma as probabilidades dos níveis de uma variável de resposta, neste caso discreta, em uma escala contínua que é ilimitada (Montgomery, Peck e Vining, 2006). Depois

[6]Uma forma fechada é aquela que pode ser expressa analiticamente em termos de um número delimitado de certas funções bem conhecidas.

de concluída a transformação, a relação entre as variáveis explicativas e a resposta pode ser modelada com regressão linear. Usualmente, a função de ligação na regressão de Poisson é a identidade, em que se pretende estimadores de $\boldsymbol{\beta}$ tais que $g(\mu_i) = \mu_i = \mathrm{x}_i'\boldsymbol{\beta}$, ou a logarítmica, em que se estima $\boldsymbol{\beta}$ tal que $g(\mu_i) = \ln(\mu_i) = \mathrm{x}_i'\boldsymbol{\beta}$, que tem uma correspondência direta com $\mu_i = g^{-1}(\mathrm{x}_i'\boldsymbol{\beta}) = e^{\mathrm{x}_i'\boldsymbol{\beta}}$ (Montgomery, Peck e Vining, 2006).

Aplicando o método de máxima verosimilhança, a função de verosimilhança, baseada na função f de 2.1, fica

$$\begin{aligned} L(\boldsymbol{y}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \frac{\prod_{i=1}^{n} \mu_i^{y_i} \exp\big(-\sum_{i=1}^{n} \mu_i\big)}{\prod_{i=1}^{n} y_i!} \, . \end{aligned}$$

Naturalmente, mais uma vez, maximizar $L(\mathbf{y}, \boldsymbol{\beta})$ é equivalente a maximizar a função de log- verosimilhança

$$\ln L(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \ln(y_i!) \, .$$

A maximização desta função requer métodos numéricos a fim de obter as estimativas dos parâmetros, uma vez que não é possível obter uma solução de forma fechada. Um dos métodos que pode ser usado é o método dos Mínimos Quadrados Ponderados Iterativamente (Montgomery, Peck e Vining, 2006, Apêndice C.14), que usa o método de Newton-Raphson.

Obtidos os parâmetros estimados $\hat{\boldsymbol{\beta}}$, o modelo da regressão de Poisson é

$$\hat{y}_i = g^{-1}(\mathrm{x}_i'\boldsymbol{\beta}) \, .$$

2.3.2 Séries temporais

Tendo em conta que serão apresentados a seguir modelos de séries temporais, há necessidade de abordar os conceitos associados às séries temporais. Enquanto que, nos modelos anteriores, a variável resposta é vista como uma variável quantitativa (não estando associada a uma dimensão de tempo), nos modelos seguintes, a dimensão do tempo é fundamental para as noções de auto-regressividade e sazonalidade, posteriormente abordadas.

Uma série temporal[7] é um conjunto de observações medidas geralmente ao longo de vários instantes de tempo. Matematicamente, pode ser definida como um conjunto de vetores $y(t), t = 0, 1, 2, ..., T$, onde t representa o tempo decorrido.

A representação gráfica de uma série temporal, ponto de partida para o seu estudo, faz-se geralmente em coordenadas cartesianas, em que no eixo das abcissas figuram os tempos e no das ordenadas os valores da série. À linha poligonal que une os pontos obtidos, ordenadamente (Figura 2.2), dá-se o nome de cronograma (Murteira, Müler e Turkman, 1993).

Existem várias motivações para se estudarem séries temporais, nomeadamente a descrição da série e consequente explicação de determinados acontecimentos ou a previsão e controlo do futuro.

O principal objetivo da modelação de séries temporais é estudar as observações passadas, rigorosamente, para desenvolver um modelo apropriado que descreva a estrutura inerente da série. Pode, então, entender-se como um processo de previsão do futuro, compreendendo o passado (Adhikari e Agrawal, 2013).

Os métodos de previsão podem ser divididos em dois grandes grupos: os causais ou múltiplos e os não causais ou não múltiplos (Murteira, Müler e Turkman, 1993). Os métodos causais procuram relacionar a

[7] O termo "Série Temporal"é mais ou menos consensual, apesar de, em (Murteira, Müler e Turkman, 1993), os autores considerarem o termo "Sucessão Cronológica"devido à diferença matemática entre série e sucessão.

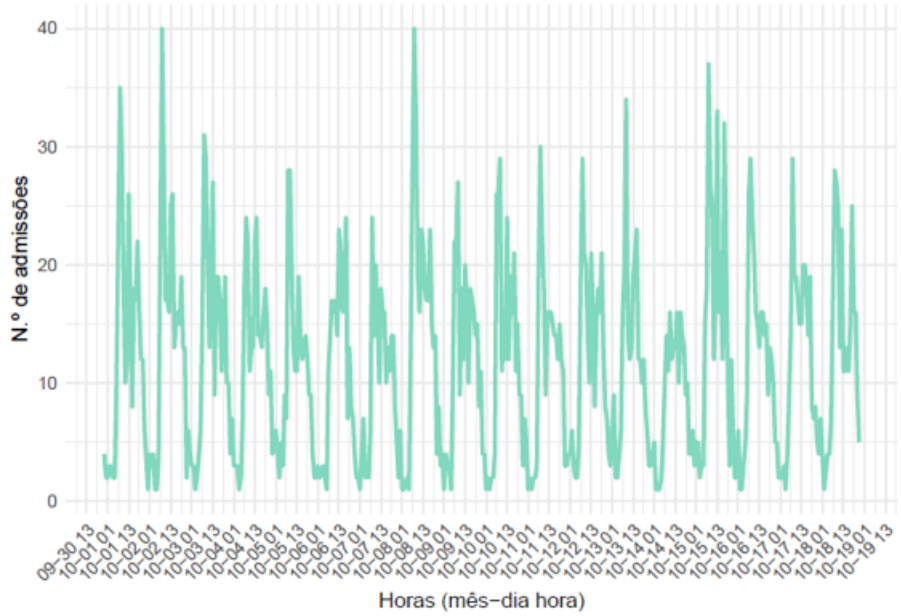


Figura 2.2: Exemplo de uma série temporal que representa o número de admissões a cada 12 horas.

variável a prever com outras variáveis através de modelos cujos parâmetros são estimados com base nas observações feitas no passado. Os métodos não causais baseiam-se exclusivamente na própria sucessão a prever e em modelos construídos com esse pressuposto.

Uma série pode-se decompor em tendência, sazonalidade e outros movimentos oscilatórios - aos quais se juntam os resíduos ou componente aleatória (Murteira, Müler e Turkman, 1993).

A tendência pode ser entendida como a variação "em média", enquanto a sazonalidade é a variação periódica em relação à tendência.

Os movimentos oscilatórios restantes estão associados a expansões e depressões recorrentes mas não periódicas. A componente aleatória, ou ruído, é composta por tudo o que não é possível modelar.

A Figura 2.3 mostra a decomposição de uma série nestas componentes.

Uma série estacionária de segunda ordem Y_t é aquela que satisfaz as condições (Harvey, 1993)

$$\begin{aligned} E(Y_t) &= \mu \\ E[(Y_t - \mu)^2] &= \sigma^2 \\ E[(Y_t - \mu)(Y_{t+h} - \mu)] &= \gamma(h), \quad h = 1, 2, ... \end{aligned}$$

ou seja, em que a média e a variância são constantes e a auto-covariância só depende de h, desfasamento dos pares de valores do processo temporal. A Figura 2.4 apresenta exemplos de estacionaridade e não estacionaridade. De agora em diante, sempre que for referida a estacionaridade, pretender-se-á fazer referência a estacionaridade de segunda ordem, aquela que permite tirar algumas conclusões sobre inferência.

Um ruído branco é um caso particular de série estacionária, conjunto de variáveis aleatórias não correlacionadas com média zero e variância finita.

2.3.3 Modelos clássicos para Séries Temporais

Os modelos clássicos mais complexos para séries temporais são uma conjugação de diversos conceitos mais simples. Os conceitos essenciais são a auto-regressão, as médias móveis, a integração (diferenciação) e a sazonalidade.

É importante definir o operador de desfasamento L que, aplicado a qualquer processo temporal y_t, é

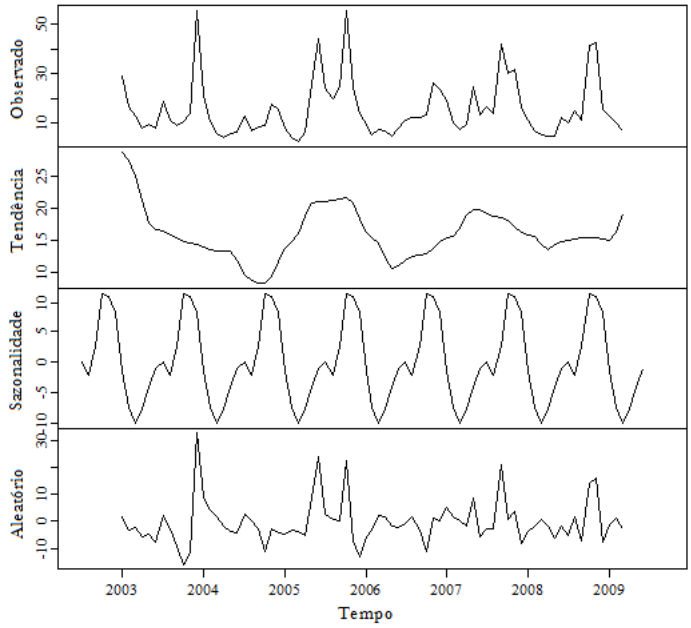


Figura 2.3: Exemplo de decomposição de uma série temporal. De cima para baixo: série original, tendência, sazonalidade e componente aleatória. (fonte: https://www.researchgate.net/figure/Figura-4-Decomposicao-da-serie-temporal-em-componentes-de-sazonalidade-de-tendencia-e_fig1_274194810)

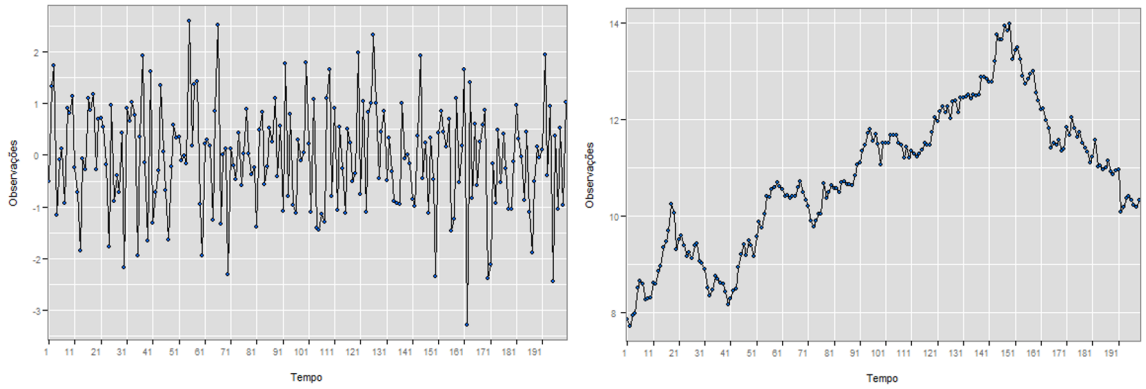


Figura 2.4: Exemplo de série estacionária e não estacionária, respetivamente, da esquerda para a direita. (fonte: http://www.portalaction.com.br/series-temporais/11-estacionariedade)

definido da seguinte forma:

$$Ly_t = y_{t-1}, \quad Ly_{t-1} = y_{t-2}, \quad ...$$
$$L^h y_t = y_{t-h}, \ h = 1, 2, 3, ... \ .$$

Modelo Auto-Regressivo

Os modelos auto-regressivos (AR) são aqueles que podem ser escritos como

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t, \quad t = 1, \ldots, T \ ,$$

onde $\phi_1, ..., \phi_p$ são os parâmetros a estimar e ε_t é um ruído branco, ou

$$\phi_p(L) y_t = \varepsilon_t,$$
$$\phi_p(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p \ .$$

O modelo pode ser denotado por $y_t \sim \text{AR}(p)$, dizendo-se um modelo auto-regressivo de ordem p (Harvey, 1993).

Este modelo é conhecido pela fácil interpretação, uma vez que apenas estabelece, para o valor no instante t, uma dependência de valores que a série tomou em instantes anteriores.

Neste tipo de modelos, assume-se que a série temporal y_t é estacionária e as raízes do polinómio AR devem estar fora do círculo unitário.

A estimação dos parâmetros pode ser realizada a partir do método de máxima verosimilhança, utilizando otimização numérica, para a função de log-verosimilhança

$$\log L(\boldsymbol{\phi}, \sigma^2) = -\frac{1}{2}T \log 2\pi - \frac{1}{2}T \log\sigma^2 - \frac{1}{2} \log|\boldsymbol{V_p}| - \frac{1}{2}\sigma^{-2}\big[\boldsymbol{y_p}'\boldsymbol{V_p}^{-1}\boldsymbol{y_p} + \sum_{t=p+1}^{T} (y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2\big] ,$$

onde $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)'$, $\boldsymbol{y_p} = (y_1, ..., y_p)'$ e $\boldsymbol{\sigma^2 V_p}$ é a matriz de covariâncias de $\boldsymbol{y_p}$.

Os estimadores podem, equivalentemente, ser obtidos por mínimos quadrados, minimizando a função

$$S(\boldsymbol{\phi}) = \sum_{t=p+1}^{T} (y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2 ,$$

ou pelo método que usa equações de Yule-Walker (Harvey, 1993).

Modelo de Médias Móveis

Os modelos de médias móveis (MA)[8] são aqueles que podem ser escritos como

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}, \;\; t = 1, \ldots, T ,$$

onde $\theta_1, ..., \theta_q$ são os parâmetros a estimar e ε_t é um ruído branco, ou

$$y_t = \theta_q(L)\varepsilon_t,$$

$$\theta_q(L) = 1 - \theta_1 L - \theta_2 L^2 - ... - \theta_q L^q .$$

Este modelo resulta da ideia de exprimir y_t em termos de um processo mais simples ε_t, em que os efeitos produzidos por inovações só perduram por um curto período de tempo, ao contrário do AR (Murteira, Müler e Turkman, 1993). O modelo pode ser denotado por $y_t \sim \text{MA}(q)$, dizendo-se um modelo de médias móveis de ordem q.

Um processo de médias móveis finito é sempre estacionário.

Caso $\varepsilon_t \sim N(0, \sigma^2)$, os parâmetros podem ser estimados pelo método da máxima verosimilhança, recursivamente, maximizando a função

$$\log L(\boldsymbol{\theta}, \boldsymbol{\sigma^2}) = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \theta\varepsilon_{t-1})^2$$

e, posteriormente, considerando

$$\varepsilon_t = y_t - \hat{\theta_1}\varepsilon_{t-1} - \ldots - \hat{\theta_q}\varepsilon_{t-q}, \;\; t = 1, \ldots, T$$

com $\varepsilon_{1-q} = \varepsilon_{2-q} = \ldots = \varepsilon_0 = 0$.

Podem também obter-se os estimadores dos parâmetros através de mínimos quadrados implementando o método de Gauss-Newton para obter a solução (Harvey, 1993).

[8] Do inglês, Moving-Avarage

Modelo Auto-Regressivo e de Médias Móveis

Quando uma série temporal pode ser gerada por um processo com uma estrutura que resulte da combinação de modelos AR e MA, é vantajoso um processo misto, modelo cujo nome é dado por ARMA(p,q), dado pela expressão geral

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}, \quad t = 1, \ldots, T$$

ou

$$\phi_p(L) y_t = \theta_q(L) \varepsilon_t \ .$$

Caso sejam experimentados cada um dos modelos AR e MA separadamente, estes podem exigir um número excessivos de parâmetros. Assim, o modelo ARMA surge como uma alternativa para a diminuição do número de parâmetros dos modelos (Murteira, Müler e Turkman, 1993).

Pode proceder-se à estimação dos parâmetros de forma semelhante ao descrito no modelos de médias móveis (Harvey, 1993).

Modelos Auto-Regressivos Integrados e de Médias Móveis

O conceito por detrás deste tipo de modelo é muito importante do ponto de vista prático dos modelos de séries temporais, pelo facto do descrito abaixo.

Nos modelos ARMA, a série temporal tem de ser estacionária. Caso não seja, que acontece muito frequentemente na prática, os modelos não podem ser usados. Para solucionar essa questão, pode ser usado o cálculo de diferenças de modo a obter uma série estacionária (Harvey, 1993).

Denotando o operador diferença por

$$\begin{aligned}
\Delta y_t &= y_t - y_{t-1} \\
\Delta^2 y_t &= \Delta(\Delta y_t) \\
&= (1-L)^2 y_t \\
&= (1 - 2L + L^2) y_t \\
&= y_t - 2y_{t-1} + y_{t-2} \\
&\vdots \\
\Delta^d y_t &= (1-L)^d y_t \ ,
\end{aligned}$$

onde d é designado por ordem das diferenças ou da diferenciação, um ARIMA(p,d,q) pode ser representado por

$$\phi_p(L) \Delta^d y_t = \theta_0 + \theta_q(L) \varepsilon_t \ .$$

Este modelo é muitas vezes usado para modelações económicas e financeiras.

À semelhança do modelo anterior, os parâmetros são calculados recursivamente através da consideração de um operador não estacionário de grau $p + d$ (Murteira, Müler e Turkman, 1993, p.146).

Sazonalidade

Sendo a sazonalidade uma variação periódica em relação à tendência, é uma informação importante para um modelo que se pretende ajustar aos dados.

Desta forma, define-se a sazonalidade S como o período em que existem tendências ou padrões.Pode-se ter modelos sazonais mais simples, como AR sazonais AR(P)$_S$ tais que

$$y_t = \Phi_1 y_{t-S} + \ldots + \Phi_P t_{t-PS} + \varepsilon_t$$

ou

$$\Phi_P(L^S)y_t = \varepsilon_t \ ,$$
$$\Phi_P(L^S) = 1 - \Phi_1 L^S - \ldots - \Phi_P L^{PS}.$$

ou modelos MA(Q)$_S$

$$y_t = \varepsilon_t - \Theta_1\varepsilon_{t-S} - \ldots - \Theta_Q\varepsilon_{t-QS}$$

ou

$$y_t = \Theta_Q(L^S)\varepsilon_t \ ,$$
$$\Theta_Q(L^S) = 1 - \Theta_1 L^S - \ldots - \Theta_Q L^{PS}$$

e modelos mais complexos que passam a integrar parâmetros não sazonais (p, d, q) e sazonais $(P, D, Q)_S$, designando-se SARIMA(p,d,q)x(P,D,Q)$_S$[9] e formulado da seguinte forma:

$$\phi_p(L)\Phi_P(L^S)\Delta^d\Delta^D y_t = \theta_q(L)\Theta_Q(L^S)\varepsilon_t \ .$$

Neste caso, os parâmetros são estimados recursivamente, como no modelo ARIMA (Harvey, 1993, p.141).

2.3.4 Modelos Lineares Generalizados para Séries Temporais de Contagem

Este tipo de modelos, usados neste estudo, são particularmente úteis quando se lida com uma série temporal de contagem sazonal (Liboschik, Fokianos e Fried, 2017).

Este modelo, que pode ser visto como uma junção de regressões com os modelos clássicos de séries temporais, tem a vantagem de se poder formular usando componentes passadas da própria série temporal e co-variáveis - variáveis explicativas externas que complementam a informação dada pela série.

Seja uma série temporal de contagem $\{Y_t : t \in \mathbb{N}\}$ e um vetor de co-variáveis r-dimensional que varia com o tempo $\{\mathbf{X}_t : t \in \mathbb{N}\}$, em que $\mathbf{X}_t = (X_{t,1}, ..., X_{t,r})^\top$. Denotando por $\mathcal{F}_t$ a história do processo conjunto $\{Y_t, \lambda_t, \mathbf{X}_{t+1} : t \in \mathbb{N}\}$, ou seja, a informação da série em t e das co-variáveis em $t+1$, considere-se que média condicional $E(Y_t|\mathcal{F}_{t-1})$ é modelada pelo processo $\{\lambda_t : t \in \mathbb{N}\}$, de modo que $E(Y_t|\mathcal{F}_{t-1}) = \lambda_t$. Pode-se considerar o modelo geral

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{h=1}^{q} \alpha_h g(\lambda_{t-j_h}) + \boldsymbol{\eta}^\top \mathbf{X}_t, \tag{2.3}$$

onde $g : \mathbb{R}^+ \to \mathbb{R}$ é uma função de ligação, $\tilde{g} : \mathbb{N}_0 \to \mathbb{R}$ é uma função de transformação e $\boldsymbol{\eta} = (\eta_1, ..., \eta_r)^\top$ é o vetor de parâmetros que corresponde ao efeitos das co-variáveis. Note-se que são definidos os conjuntos $P = \{i_1, i_1, ..., i_p : 0 < i_1 < i_2 < ... < i_p < \infty, p \in \mathbb{N}_0\}$ e $Q = \{j_1, j_2, ..., j_q : 0 < j_1 < j_2 < ... < j_q < \infty, q \in \mathbb{N}_0\}$ para os índices que definem o desfasamento considerado em observações e médias condicionadas, respetivamente. Esta definição dos desfasamentos é também uma vantagem em relação aos modelos clássicos para séries temporais, uma vez que, neste podem ser considerados vários períodos diferentes de sazonalidade no mesmo modelo pois é apenas exigido que eles tenham uma ordem em vez de serem um incremento fixo.

É usual, tratando-se de uma série temporal de contagem, que se tome $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$, o que implica que

$$P(Y_t = y|\mathcal{F}_{t-1}) = \frac{\lambda_t e^{-\lambda_t}}{y!}, \ \ y = 0, 1, ... \ . \tag{2.4}$$

Denotando por $\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_p, \alpha_1, ..., \alpha_q, \eta_1, ..., \eta_r)^\top$ o vetor de todos os parâmetros do modelo, o espaço dos parâmetros é dado por

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : \beta_0 > 0, \beta_1, ..., \beta_p, \alpha_1, ..., \alpha_q, \eta_1, ..., \eta_r \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{h=1}^{q} \alpha_h < 1 \right\}$$

[9] Modelo Sazonal Auto-Regressivo Integrado e de Médias Móveis, do inglês, Seasonal Autorregressive Integrated Moving-Avarage.

ou por

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : |\beta_1|, ..., |\beta_p|, |\alpha_1|, ..., |\alpha_q| < 1, \left| \sum_{k=1}^{p} \beta_k + \sum_{h=1}^{q} \alpha_h \right| < 1 \right\},$$

caso a função de ligação seja logarítmica.

O estimador de quasi máxima verosimilhança (QMLE) $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$, abordagem escolhida pela simplicidade e utilidade em derivar estimadores consistentes, é a solução do problema de otimização com restrições não lineares

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}), \tag{2.5}$$

onde

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_t(y_t; \boldsymbol{\theta}) = \sum_{t=1}^{n} \left(y_t \ln \left(\lambda_t(\boldsymbol{\theta}) \right) - \lambda_t(\boldsymbol{\theta}) \right) \tag{2.6}$$

é a função condicional de quasi log-verosimilhança, onde $p_t(y; \boldsymbol{\theta})$ é a função densidade de probabilidade da distribuição de Poisson, definida na equação 2.4.

Para resolver numericamente este problema de maximização é utilizado um algoritmo que basicamente impõe as restrições adicionando um valor de barreira à função objetivo e, de seguida, aplica um algoritmo para otimização sem restrições a essa nova função objetivo, repetindo estes dois passos, se necessário (Liboschik, Fokianos e Fried, 2017).

2.3.5 Critérios de seleção de modelos

Para selecionar os modelos existem várias medidas de seleção. Entre elas estão o RMSE[10] e o AIC[11], usados neste estudo.

O critério de informação de Akaike (AIC) é um estimador da qualidade relativa de modelos estatísticos. A formulação deste critério é dada por

$$\text{AIC} = -2\tilde{\ell}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) + 2k$$

onde $\tilde{\ell}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \sum_{t=1}^{n} \log(p_t(y_t))$ é a função de log-verosimilhança "verdadeira"e k é o número de parâmetros estimados, ou seja, o número de componentes de $\hat{\boldsymbol{\theta}}$ (Liboschik, Fokianos e Fried, 2017).

A outra medida usada é a raiz do erro quadrático médio (RMSE). Como o próprio nome indica, esta medida consiste no erro que existe entre as previsões e os valores verdadeiros correspondentes. Assim, esta medida é calculada através da fórmula

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}},$$

em que $\hat{y}_t$ são os valores previstos para os valores verdadeiros de teste y_t e n é o número de valores previsto.

Em ambas as medidas se pretende selecionar o menor valor, apesar de apresentarem ordens de grandeza muito distintas.

[10] Do inglês, Root Mean Square Error, é a raíz dos erros quandráticos médios.

[11] Do inglês, Akaike Information Criterion, é o critério de informação de Akaike.

Capítulo 3

Exploração dos Dados e Modelação

Depois de estudada a modelação mais adequada ao objetivo, segue-se a parte prática do trabalho, na qual se efetuou todo o processamento dos dados, a análise dos mesmos e a modelação e seleção do modelo mais vantajoso.

3.1 Exploração dos Dados

Os dados utilizados neste trabalho são reais e provenientes de uma Unidade Local de Saúde (ULS) portuguesa.

O conjunto de dados original é composto por **2253876 observações** e **27 variáveis**.

Estas observações correspondem ao período entre as 0 horas de 1 de janeiro de 2014 e as 11 horas de 19 de outubro de 2018. Originalmente, cada observação não corresponde a uma admissão, dado que existem admissões caracterizadas em várias observações, como se poderá ver adiante.

As variáveis originais estão descritas na Tabela 3.1.

Note-se que nos dados utilizados existe uma cor adicional às da Triagem de Manchester, o branco. A cor branca é atribuída para outros casos, como por chamada do médico ou outros.

Para que os dados estejam adequados para a fase de modelação, é necessário realizar algumas etapas de pré-processamento e reestruturação, que a seguir se descrevem.

3.1.1 Pré-processamento

Anonimização

A primeira alteração efetuada aos dados consistiu na sua anonimização pois, tal como a ULS, os dados que provêm dela também carecem de anonimato, para efeitos de proteção de dados.

As variáveis que necessitam de ser totalmente anonimizadas são aquelas que comprometem, na íntegra, a proteção dos dados pessoais, profissionais e da ULS. Essas variáveis são:

- ADMISSAO_ID;
- ENFERMEIRO;
- IDENTIFICACAO;
- LOCALIDADE;
- LOCAL_URGENCIA;
- MEDICO;
- NUM_TRANSF;

Variável	Descrição
ADMISSAO_ID	identificação única de uma admissão
CAUSA_ADMISSAO	causa da admissão do utente
COD_VALOR_ANALISADO	código do valor analisado na triagem (por exemplo, medindo-se a temperatura corporal do utente, essa medida tem o código 1 associado)
COD_VIA_VERDE	código da via verde (por exemplo, tratando-se de uma via verde do AVC, associa-se o código 1)
CONDICIONANTES_NO_REGISTO	condicionantes no registo na triagem
COR_PULSEIRA	cor da pulseira (azul, verde, amarelo, laranja, vermelho, branco)
DESCRICAO_VALOR_ANALISADO	descrição do valor analisado na triagem (Temperatura, Fluxo expiratório máximo, Frequência respiratória, Oximetria Periférica, Escala de Coma de Glasgow, Dor (na régua da dor), Frequência cardíaca (no pulso), Glicémia, Pressão arterial diastólica, Pressão arterial sistólica)
DES_SALA	descrição da sala em que a admissão ocorreu
DTA_ADMISSAO	data e hora a que a admissão ocorreu (formato aaaa-mm-dd hh:mm:ss)
DTA_TRIAGEM	data e hora da triagem
ENFERMEIRO	identificação única do enfermeiro triador
ESPECIALIDADE	especialidade de atendimento do utente
IDADE	idade do utente
IDENTIFICACAO	identificação única do utente
LOCALIDADE	localidade de habitação do utente
LOCAL_URGENCIA	local em que a urgência se situa (no caso, existem várias urgência na ULS)
MEDICO	identificação única do médico médico responsável pelo atendimento na especialidade
NUM_TRANSF	número identificador de transferência caso o utente tenha sido transferido entre cuidados diferenciados da ULS
OBSERVACOES	observações acerca da admissão
PRIORIDADE	prioridade atribuída ao utente após triagem de Manchester (não urgente, pouco urgente, urgente, muito urgente, emergente, outros casos)
PROVENIENCIA	proveniência do utente na admissão (exterior, inem, centro de saúde, outro hospital, etc)
QUADRO_CLINICO	quadro clínico do utente (sintomas) na triagem
REGIAO	região de habitação do utente
TIPO_UNIDADE_SAUDE	tipo de unidade de saúde de onde provem (caso o campo PROVENIENCIA seja preenchido)
UNIDADE_SAUDE_PROVENIENCIA	unidade de saúde de onde provem (caso tenha vindo de outro hospital, de centro de saúde, etc)
VALOR_REGISTADO	valor analisado na triagem (por exemplo, quando medindo-se a temperatura corporal do utente, o valor registado neste campo é o valor numérico resultante 37.5)
VIA_VERDE	tipo de via verde (do AVC, Coronária)

Tabela 3.1: Dados: nome das variáveis e sua descrição.

- REGIAO;
- UNIDADE_SAUDE_PROVENIENCIA.

Para a anonimização destas variáveis, usou-se, para substituir cada um dos valores que elas tomavam,

- uma letra maiúscula, escolhida de modo que houvesse uma associação à variável em questão - ver Tabela 3.2;
- um número inteiro que é atribuído sequencialmente a cada valor quando ordenados consoante o sistema de numeração hexatrigesimal (composto pelos algarismos arábicos e o alfabeto latino), em que os algarismos são antes das letras e as letras minúsculas são antes das letras maiúsculas.

Este procedimento está ilustrado na Figura 3.1, com exemplos não reais.

Existem também variáveis que apenas necessitam de ser parcialmente anonimizadas pois os campos são compostos por expressões importantes para o estudo e expressões que comprometem o anonimato, tais como ESPECIALIDADE e PROVENIENCIA. Para essas, foi analisado em cada caso qual a expressão importante para permanecer, procurada linha a linha essa expressão e substituído cada valor em que a expressão foi encontrada apenas pela expressão pretendida, eliminando assim palavras comprometedoras (Figura 3.2).

Variável	**Letra**
ADMISSAO_ID	A
ENFERMEIRO	E
IDENTIFICACAO	I
LOCALIDADE	L
LOCAL_URGENCIA	U
MEDICO	M
NUM_TRANSF	T
REGIAO	R
UNIDADE_SAUDE_PROVENIENCIA	P

Tabela 3.2: Letras escolhidas para a anonimização de cada variável.

Coluna MEDICO original

MEDICO
1A35DF9
H321X5F
QO25101
4DR78PT
A3F527U
RV89LIY
A3f95G1

Componentes de anonimização

Letra	**Número**
M	1
M	5
M	6
M	2
M	4
M	7
M	3

Coluna MEDICO anonimizada

MEDICO
M1
M5
M6
M2
M4
M7
M3

Figura 3.1: Exemplo de anonimização para a variável MEDICO - exemplos não reais.

Reestruturação

Para este estudo, objetiva-se que cada uma das observação (linhas) da tabela sejam referentes a admissões distintas para que aquando da contagem do número de admissões por hora, não seja contada a mesma admissão várias vezes.

Coluna ESPECIALIDADE original

ESPECIALIDADE
CLINICA GERAL DE LOCAL_A
MEDICINA INTERNA
MEDICINA INTERNA DE LOCAL_F
CLINICA GERAL DE LOCAL_B
CLINICA GERAL
MEDICINA INTERNA DE LOCAL_D
MEDICINA INTERNA DE LOCAL_C

Coluna ESPECIALIDADE anonimizada

ESPECIALIDADE
CLINICA GERAL
MEDICINA INTERNA
MEDICINA INTERNA
CLINICA GERAL
CLINICA GERAL
MEDICINA INTERNA
MEDICINA INTERNA

Figura 3.2: Exemplo de anonimização para a variável ESPECIALIDADE - campos não reais.

ADMISSAO_ID	IDADE	MEDICO	...	COD_VALOR_ANALISADO	DESCRICAO_VALOR_ANALISADO	VALOR_REGISTADO
A123	45	M6		1	TEMPERATURA	37
A123	45	M6		2	OXIOMETRIA PERIFERICA	99
A123	45	M6		3	ESCALA DE COMA GLASGOW	15
A123	45	M6		4	DOR (REGUA DA DOR)	1
A123	45	M6		5	FREQUENCIA CARDIACA (PULSO)	116

Figura 3.3: Exemplo da estruturação inicial do conjunto de dados.

A reestruturação dos dados é fundamental, principalmente porque os valores analisados na triagem vão fazendo repetir a restante informação do conjunto de dados, como está ilustrado sucintamente na Figura 3.3.

Os dez valores analisados na triagem, presentes no conjunto de dados

- Temperatura;
- Fluxo expiratório máximo;
- Frequência respiratória;
- Oxiometria Periférica;
- Escala de Coma de Glasgow;
- Dor (na régua da dor);
- Frequência cardíaca (no pulso);
- Glicémia;
- Pressão arterial diastólica;
- Pressão arterial sistólica;

são, posteriormente, considerados como variáveis (colunas). Assim, efetua-se a sua extração para um conjunto de dados auxiliar - *dadosValores* - ficando uma observação para cada uma das 783895 admissões (Figura 3.4).

Posteriormente, elimina-se do conjunto base as três colunas correspondentes a estes valores e a CONDICIONANTES_NO_REGISTO, que não é relevante para o estudo e as linhas que ficaram duplicadas com a eliminação dessas colunas. O conjunto de dados base fica com **1103263 observações** e **23 variáveis**.

ADMISSAO_ID	TEMPERATURA	EXPIRATORIO	RESPIRATORIO	OXIOMETRIA	COMA	DOR	PULSO	GLICEMIA	DIASTOLICA	SISTOLICA
A123	37			99	15	1	116			
A124	37					2	145			
A125		10	34				123	7		
A126	37,5			89					23	45

Figura 3.4: Extração dos valores registados na triagem para colunas do conjunto de dados auxiliar - *dadosValores*.

O objetivo é conseguir juntar a informação deste conjunto auxiliar com a do conjunto base mas, enquanto o conjunto *dadosValores* tem o número de observações correspondente com o número de admissões, o conjunto base ainda não, ou seja, ainda existem admissões repetidas.

Decidiu-se, entretanto, eliminar as observações do dia 19 de outubro de 2018, uma vez que apenas existiam até às 11 horas da manhã, podendo, numa representação diária do número de admissões, ser mal interpretada em relação aos restantes dias. O conjunto base fica assim com **1103056 observações**, correspondentes a **783754 admissões**, que será o número de observações do conjunto auxiliar.

Averiguou-se o número de triagens por admissão e concluiu-se que era esse facto que fazia repetir as admissões no conjunto, ou seja, havia admissões com várias datas distintas de triagem (como mostra o exemplo da figura 3.5.

ADMISSAO_ID	IDADE	MEDICO	...	PRIORIDADE	DTA_TRIAGEM
A123	45	M6		Não urgente	2018-05-23 10:18:05
A123	45	M6		Urgente	2018-05-23 10:18:21
A123	45	M6		Pouco urgente	2018-05-23 10:19:34
A123	45	M6		Urgente	2018-05-23 10:19:46
A123	45	M6		Pouco urgente	2018-05-23 10:19:58

Figura 3.5: Exemplo de admissão com instantes de triagem diferentes.

Perante este problema e em conversa com os profissionais da empresa e da ULS, decidiu-se considerar a data de triagem mais recente para cada admissão, uma vez ter sido a efetivada e desencadeadora dos tratamentos, resultando um total de **783754 observações** no conjunto base. Na sequência deste feito, criou-se uma nova variável, TRIAGENS, de modo a guardar a quantidade de triagens por admissão e eliminou-se a variável DTA_TRIAGEM por se verificar inútil doravante. Deste modo, as admissões já não se encontram repetidas e podem juntar-se ao conjunto base com as 783754 observações do conjunto dos valores registados na triagem. O conjunto base fica assim com **783754 observações** e **33 variáveis**.

Ao longo do estudo, verificou-se crucial ter em conta as admissões apenas de um local das urgências para que não fossem previstas todas as admissões da ULS. Uma vez que a ULS tem Serviços de Urgência em cidades distintas, para o suporte à alocação de recursos que é certamente independente em cada Serviço de Urgência, apenas fará sentido considerar um deles. Assim, foram eliminadas as admissões em que o local da urgência correspondesse a U1 e U2, ficando com **481204 observações**.

Concluiu-se também que mais algumas variáveis seriam irrelevantes para o estudo, como DES_SALA, NUM_TRANSF, OBSERVACOES, TIPO_UNIDADE_SAUDE e UNIDADE_SAUDE_ PROVENIENCIA. Posto isto, estas variáveis foram eliminadas dando um total de **28 variáveis**.

Posteriormente, verifica-se relevante para a análise separar a informação da data de admissão em hora (HORA), dia (DIA), mês (MES), ano (ANO), dia da semana (DIA_S), só data (DATA), só mês e ano (MES_ANO) e só data e hora (DATA_HORA), sem minutos e segundos. Assim, o conjunto fica com **36 variáveis**. A variável DIA_S (dia da semana) toma os valores 1 se for domingo, 2 se for segunda-feira, 3 quando é terça-feira, 4 quarta-feira, 5 quinta-feira, 6 sexta-feira e 7 quando é sábado.

Considerou-se também relevante ter em conta o tipo de dia em que as admissões ocorriam, no que diz respeito à localização dos feriados, e criou-se uma nova variável cujos valores são 1 se a admissão

ocorre no dia anterior a um feriado, 2 se a admissão ocorre num feriado, 3 se ocorre num dia seguinte a um feriado e 4 se for outra situação.

Os dias considerados como feriados são:

- dia de ano novo (1 de janeiro);
- dia de Carnaval (móvel);
- Sexta-Feira Santa (móvel);
- dia de Páscoa (móvel);
- Dia da Liberdade (25 de abril);
- Dia do Trabalhador (1 de maio);
- Dia de Portugal (10 de junho);
- Corpo de Deus (móvel);
- Assunção de Nossa Senhora (15 de agosto);
- feriado Municipal;
- dia da Implantação da República (5 de outubro);
- Dia de Todos os Santos (1 de novembro);
- dia da Restauração da Independência (1 de dezembro);
- Dia da Imaculada Conceição (8 de dezembro);
- dia de Natal (25 de dezembro).

Entendeu-se que a estação do ano pudesse também contribuir para o padrão temporal das admissões, criando-se uma nova variável. A regra usada para tal é:

- dias entre 21 de dezembro, inclusive, e 21 de março correspondem a Inverno;
- dias entre 21 de março, inclusive, e 21 de junho correspondem a Primavera;
- dias entre 21 de junho, inclusive, e 21 de setembro correspondem a Verão;
- dias entre 21 de setembro, inclusive, e 21 de dezembro correspondem a Outono.

Com a variável relativa aos feriados e a variável das estações do ano, o conjunto de dados fica com **38 variáveis**.

Por fim, decidiu-se acrescentar variáveis meteorológicas, hora a hora, como a temperatura, humidade e velocidade do vento, ficando o conjunto com **41 variáveis**.

O histórico das variáveis meteorológicas de 2014 a 2018, por dia e hora, foi extraído do site `https://openweathermap.org/history` e teve um custo associado.

Resulta o conjunto de dados com 481204 observações (admissões) e 41 variáveis que são:

Resumidamente, a Figura 3.6 apresenta um fluxograma com o procedimento de reestruturação.

3.1.2 Análise descritiva

Para a análise descritiva dos dados, não foi efetuado tratamento de valores omissos pois poderia influenciar a análise.

Numa primeira instância, analisaram-se algumas variáveis que podem ser importantes do ponto de vista de melhoria do serviço de urgência (SU) mas que não podem ser incluídas no modelo, como a idade do utente, a causa da admissão e a especialidade.

Pela Figura 3.7, pode-se concluir que a maior parte das admissões são de utentes com idade de vida ativa (dos 20 aos 70 anos). Esta informação pode desencadear ações de mobilização de cuidados ao

ADMISSAO_ID	ESPECIALIDADE	OXIOMETRIA
ANO	ESTACAO_ANO	PRIORIDADE
CAUSA_ADMISSAO	EXPIRATORIO	PROVENIENCIA
COD_VIA_VERDE	FERIADO	PULSO
COMA	GLICEMIA	QUADRO_CLINICO
COR_PULSEIRA	HORA	REGIAO
DIA	HUMIDADE	RESPIRATORIO
DIA_S	IDADE	SISTOLICA
DATA	IDENTIFICACAO	TEMPERATURA_AMBIENTE
DATA_HORA	LOCALIDADE	TEMPERATURA_TRIAGEM
DIASTOLICA	LOCAL_URGENCIA	TRIAGENS
DOR	MEDICO	VENTO
DTA_ADMISSAO	MES	VIA_VERDE
ENFERMEIRO	MES_ANO	

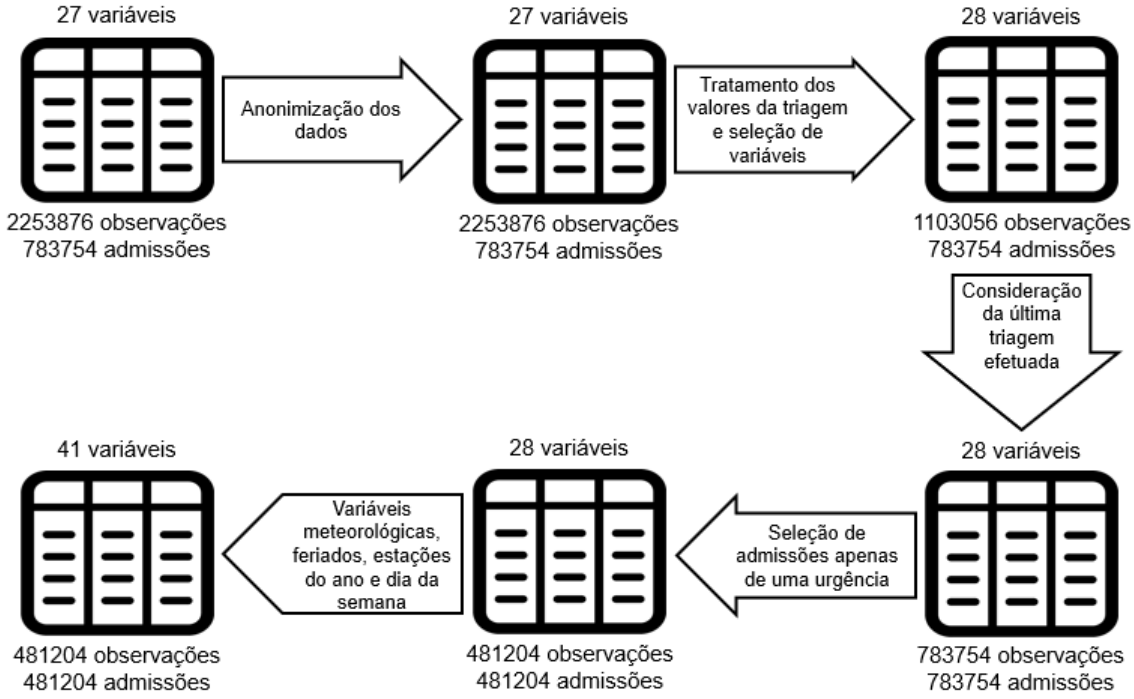


Figura 3.6: Fluxograma com o resumo da reestruturação para análise.

domicílio, dado que a população, no geral, é envelhecida e pode haver dificuldade de acesso aos cuidados de saúde por parte das pessoas de mais idade.

Por outro lado, a Figura 3.8 mostra que mais de 75% das admissões no SU são causadas por doença. Este facto leva a pensar que, para diminuir estas admissões, de modo a que os SU não ultrapassem a sua capacidade e tenham um melhor funcionamento, poderá ter de haver um acompanhamento mais estrito de utentes com doença nos centros de saúde.

Consegue-se também fazer uma melhor alocação de recursos (humanos e outros) através da informação dada pela Figura 3.9. Pode verificar-se que as especialidades médicas mais afluentes são Urgência-Triagem (25%), Pediatria (15%), Cirurgia Geral (12%), Ortopedia (11%) e Medicina Interna (10%).

Uma outra informação que é bastante importante no SU é a prioridade das admissões ou, equivalentemente, a cor da pulseira atribuída. Pode verificar-se, na Figura 3.10, que existem muito mais admissões com pulseira de cor amarela (32%) do que das restantes, o que pode ser justificado por ser a cor "central", e que existem bastantes admissões com cor verde (17%), ou seja, pouco urgentes. Estas últimas podem

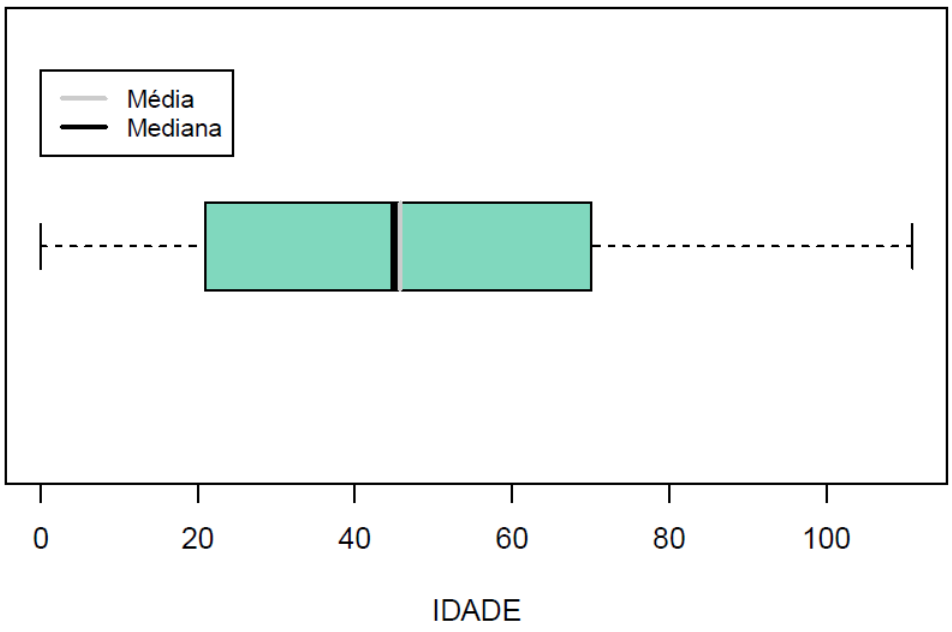


Figura 3.7: *Boxplot* da idade dos utentes.

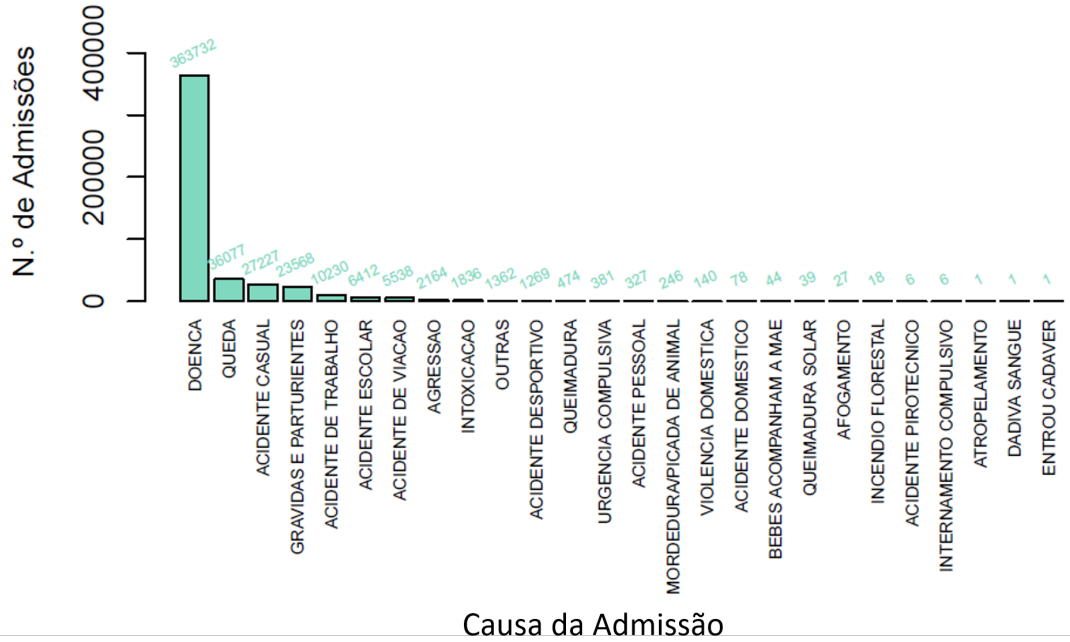


Figura 3.8: Gráfico de barras da distribuição do número de admissões pela sua causa.

denunciar falhas nas unidades básicas de saúde ou até mesmo falta de informação e/ou falha na seleção da unidade de saúde por parte dos utentes.

Numa outra fase, analisaram-se variáveis que, empiricamente, pudessem influenciar o número de admissões no SU e, por isso, pudessem fazer parte do modelo.

Relativamente aos dias da semana, pela Figura 3.11, verifica-se que há um decréscimo do número de admissões ao longo dos dias, de aproximadamente 81000 admissões na segunda-feira para cerca de 58000 no domingo. Isto incita a que esta seja uma das co-variáveis no modelo aplicado pois consoante o dia da semana, que se sabe à partida, a previsão há-de ser afetada.

O número de admissões também sofre alguma variação em termos das estações do ano (Figura 3.12),

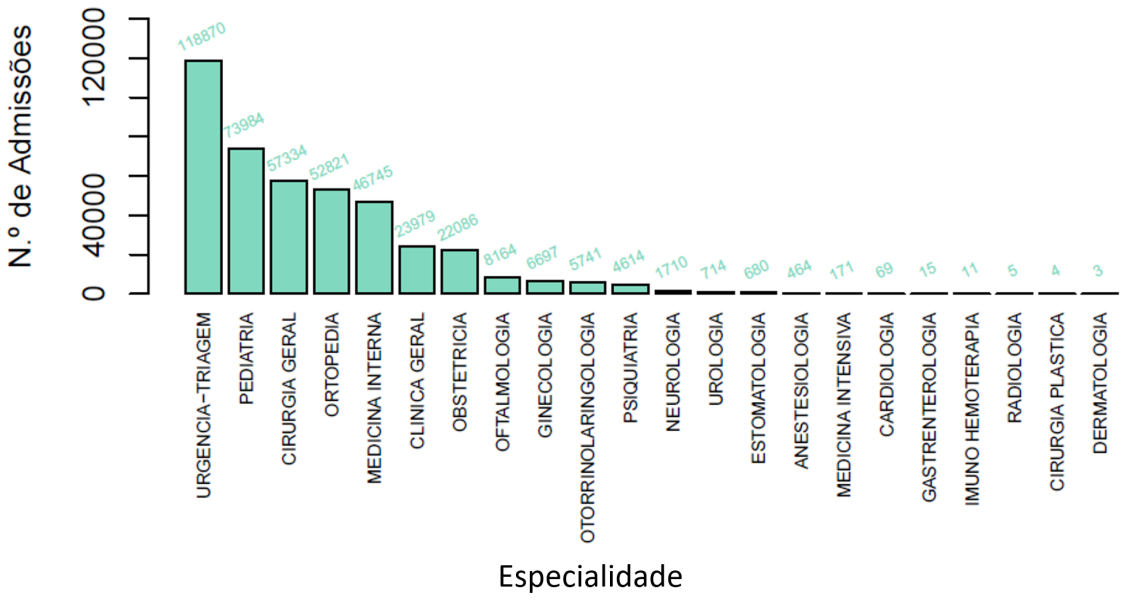


Figura 3.9: Gráfico de barras da distribuição do número de admissões pelas especialidades.

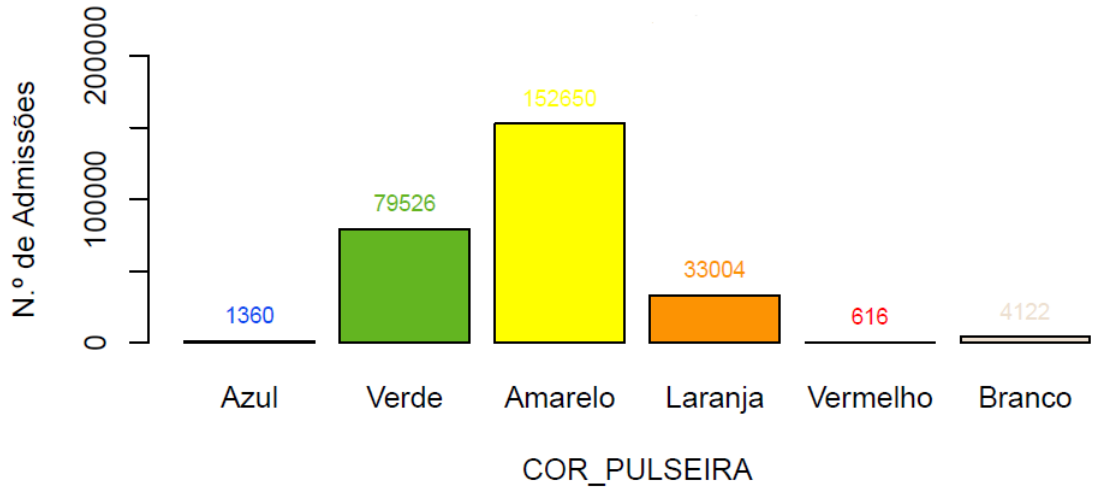


Figura 3.10: Gráfico de barras da distribuição do número de admissões pelas cores da pulseira.

dado que o outono apresenta claramente menos admissões que as restantes estações, cerca de 4% a menos. Então esta também é uma variável que provavelmente é significativa no modelo enquanto variável explicativa.

O tipo de dia em relação a feriados, aparentemente, não tem grande impacto no número de admissões (Figura 3.13) mas esta variável poderá fazer a diferença em instantes específicos da série.

O gráfico da Figura 3.14 teve como objetivo confirmar a conjetura de que o número de admissões dependia da hora que decorre. É claro que na madrugada (das 0h às 7h) o número de admissões, na ordem dos milhares, é muito mais baixo do que nas restantes horas, na ordem das dezenas de milhares de admissões.

Numa fase final de análise, analisou-se o comportamento da série temporal ao longo da várias dimensões temporais, através de cronogramas.

Na Figura 3.15, que representa a série temporal do número de admissões por hora de janeiro de 2014 a outubro de 2018, pode observar-se determinadas zonas em que, periodicamente, a série aumenta ou

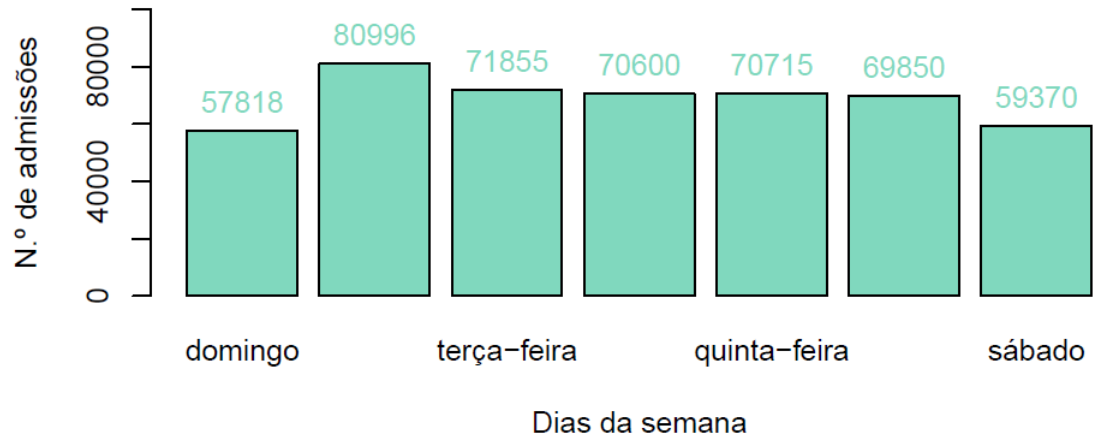


Figura 3.11: Gráfico de barras da distribuição do número de admissões pelos dias da semana.

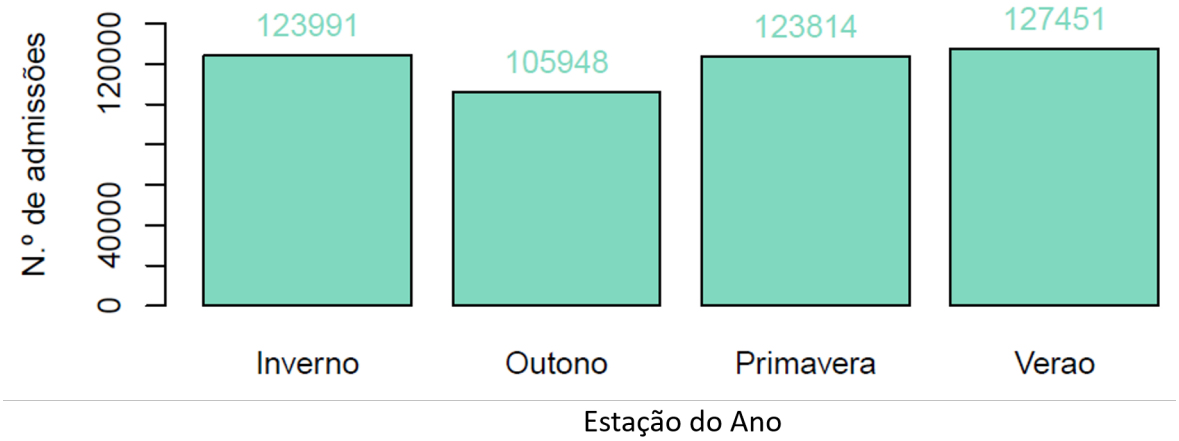


Figura 3.12: Gráfico de barras da distribuição do número de admissões pelas estações do ano.

diminui os seus valores.

Para ter mais pormenor acerca desta questão de sazonalidade, analisou-se a série numa janela temporal mais pequena (de 19 dias), na Figura 3.16. Pode observar-se que existe uma sazonalidade ao nível das horas, mais precisamente, que de 24 em 24 horas é evidente uma tendência.

Uma análise que se revelou bastante importante foi a do cronograma do número de admissões discriminando a cor das pulseiras (Figura 3.17). Foi através desta representação que se verificou que apenas as admissões posteriores a março de 2016 tinham informação acerca da triagem na base de dados da empresa. Relativamente à dispersão, constatou-se que existem mínimos e máximos locais em cada uma das cores nos mesmos instantes de tempo, o que realça o carácter sazonal da série, independentemente da cor da pulseira.

Ainda numa análise da sazonalidade, o cronograma das admissões mensais (Figura 3.18), em que se destacaram alguns mínimos e máximos, revelou uma tendência periódica, interpretada ao nível dos meses - no caso do agosto repetido em dois máximos - ou de estações do ano - no caso dos mínimos em meses de estações do ano normalmente mais frias.

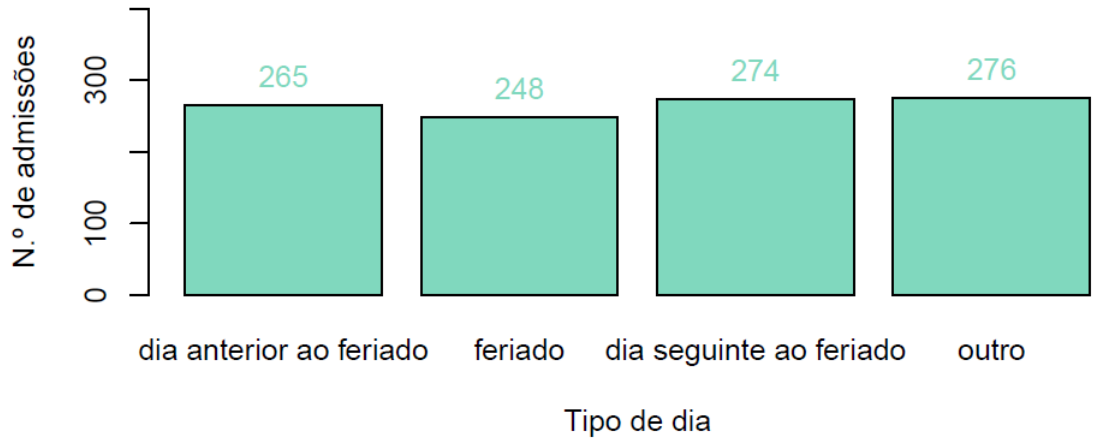


Figura 3.13: Gráfico de barras da distribuição do **número médio das admissões** pelos tipos de dia em relação aos feriados.

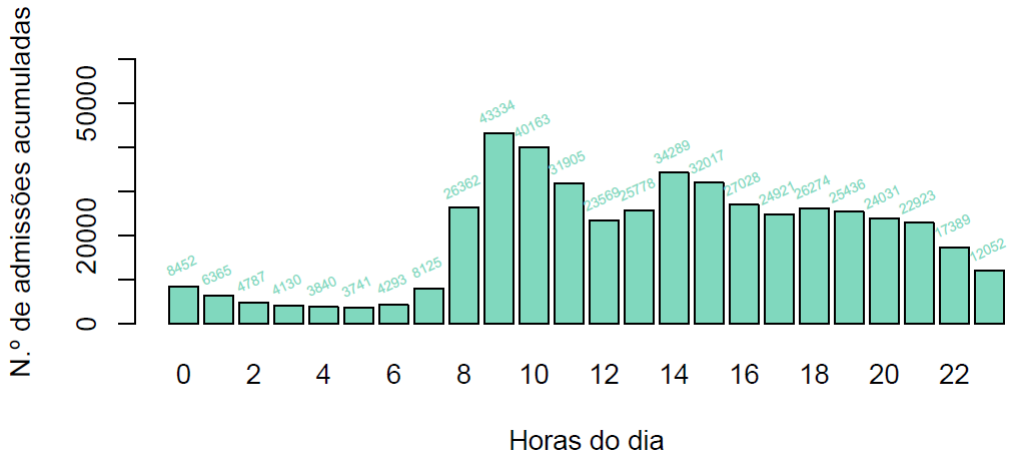


Figura 3.14: Gráfico de barras da distribuição do número de admissões acumuladas pelas horas.

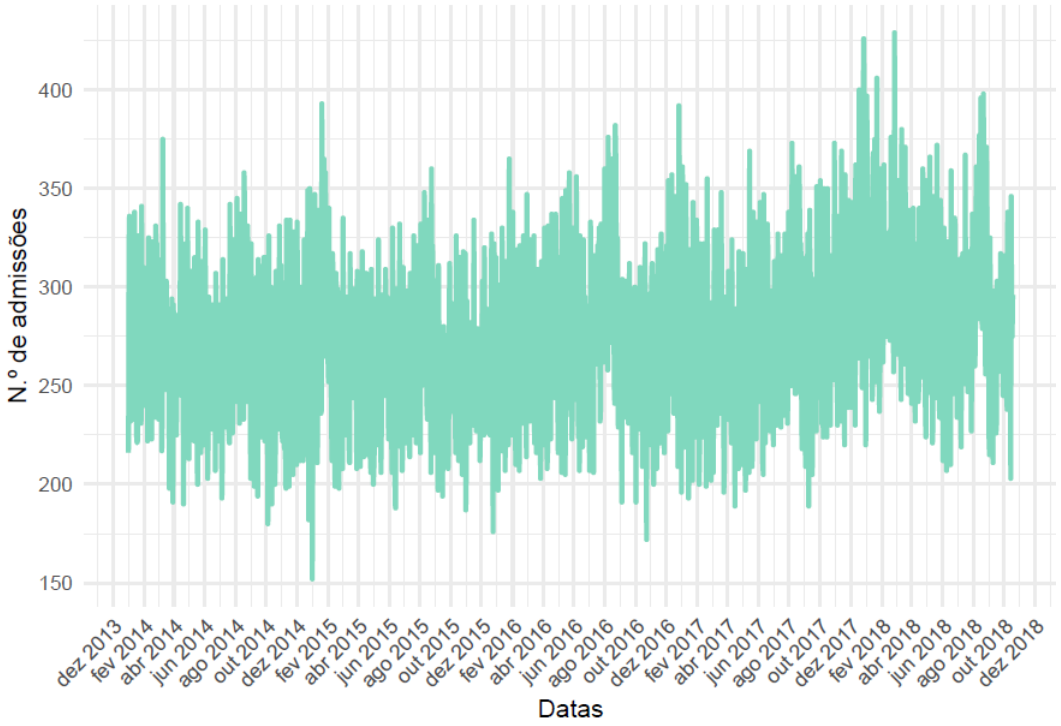


Figura 3.15: Cronograma da série temporal do número de admissões por hora, de janeiro de 2014 a outubro de 2018.

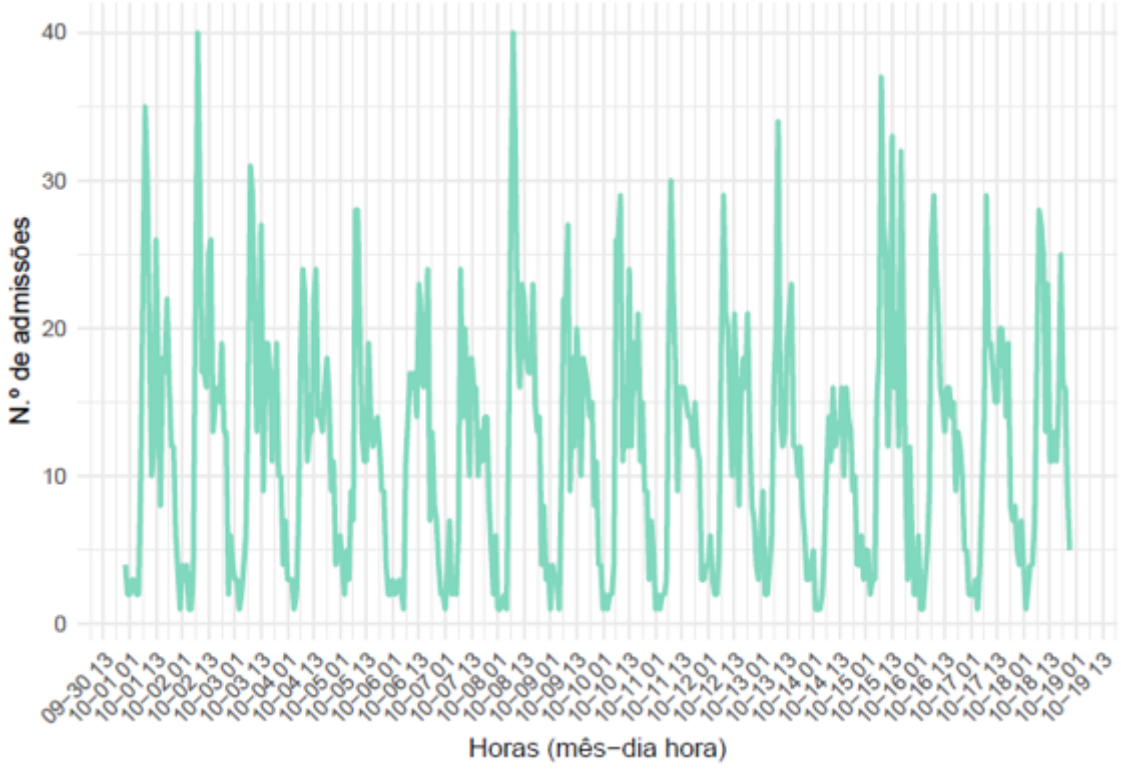


Figura 3.16: Cronograma da série temporal do número de admissões por hora, de 1 de outubro de 2018 a 19 de outubro de 2018.

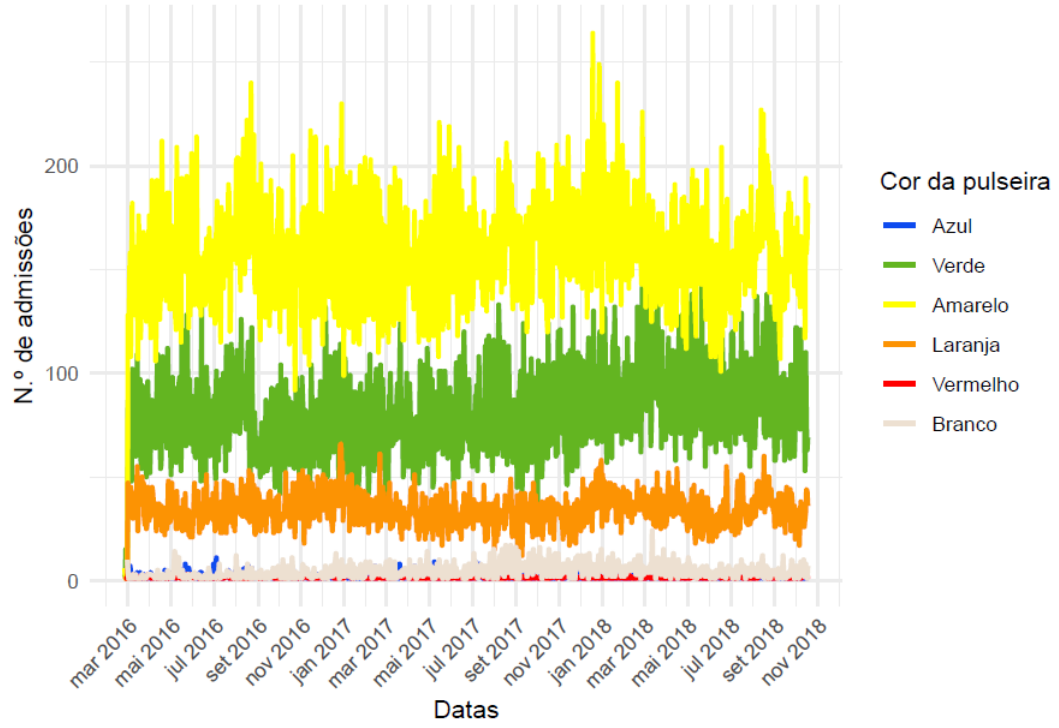


Figura 3.17: Cronograma da série temporal do número de admissões por hora, discriminado por cor da pulseira, de março de 2016 a outubro de 2018.

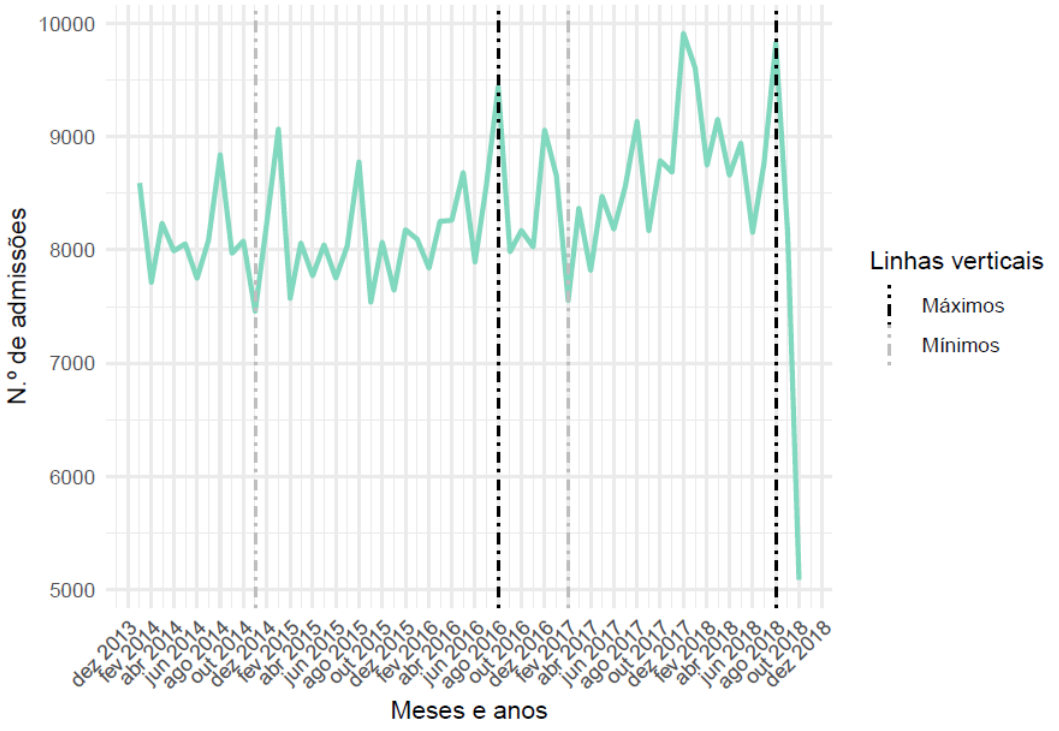


Figura 3.18: Cronograma da série temporal do número de admissões mensais, de janeiro de 2014 a outubro de 2018.

3.2 Modelação

Tal como no processamento dos dados, o *software* usado para implementar o modelo e obter as previsões foi o *R*, um software gratuito para Estatística Computacional (The R Foundation, 2018), apoiado

pelo ambiente de desenvolvimento (IDE) *RStudio* (RStudio, 2018).

Para além do processamento anteriormente descrito, foi necessário reestruturar novamente os dados para uma estrutura adequada à modelação.

Como foi referido anteriormente, a questão das variáveis é bastante importante, uma vez que a previsão não se pode basear em informações de utentes ou do SU, que não são conhecidas apriori. O que se pretende prever é o número de admissões por hora ao SU e as variáveis que podem influenciar são, naturalmente, externas ao SU.

As informações que à partida teriam influência na variável resposta são o dia da semana, a estação do ano, a localização em relação a feriados e variáveis meteorológicas (temperatura, humidade e velocidade do vento). Assim, foram selecionadas as colunas das variáveis acima mencionadas e as admissões foram agrupadas por data e hora, ou seja, houve uma inserção de uma coluna nova com a contagem de admissões na respetiva data e hora (Figura 3.19).

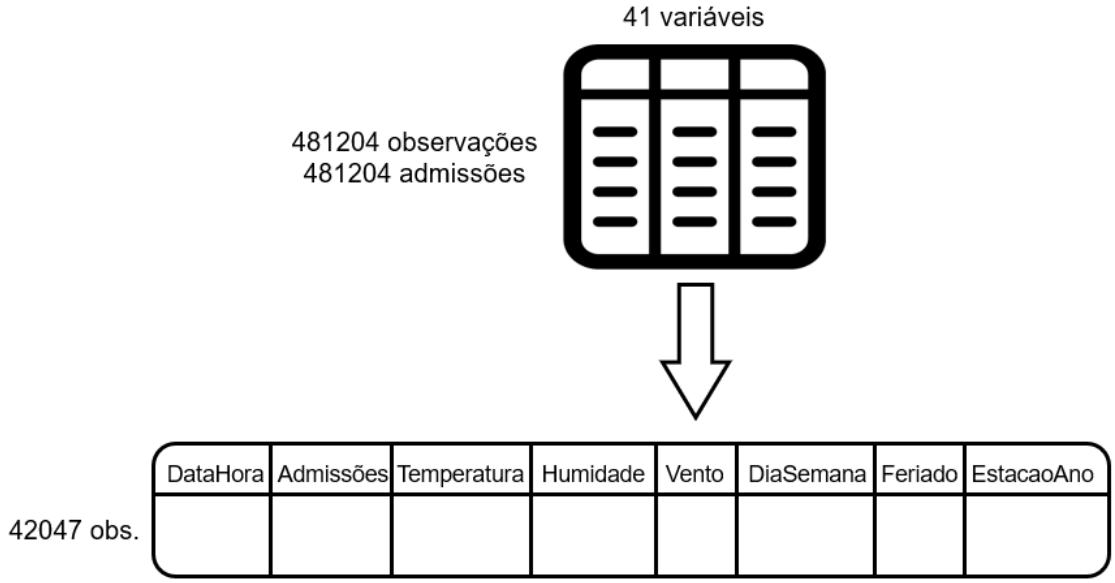


Figura 3.19: Fluxograma da reestruturação para a previsão.

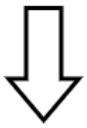
É importante referir que para o uso das variáveis categóricas[1] na modelação, foi necessário a transformação dessas variáveis em numéricas, originando variáveis dummy. No sentido literal da palavra, variéveis *dummy* são variáveis falsas ou fictícias que representam níveis de fatores, ou seja, representam algo que não possui valores numéricos (Heij et al., 2004). A ideia é criar uma variável para cada uma das classes da variável categórica que toma valor zero caso a classe não se verifique na observação ou valor um caso se verifique (Figura 3.20).

Para poder implementar o modelo e testar o ajuste do modelo, tem de se dividir o conjunto de dados em subconjuntos de treino, para ajustar o modelo aos dados, e de teste, para averiguar o comportamento do modelo perante novos dados. Usualmente, estes subconjuntos são constituídos aleatoriamente por observações do conjunto original mas, como se trata de uma série temporal, em que a ordem cronológica é extremamente importante, é necessário que as observações de treino sejam anteriores às de teste. Considerando que uma grande quantidade de dados dará ao modelo maior capacidade de captar padrões sazonais, selecionam-se quatro ano de treino e dez dias de teste, que se prende com o objetivo final da previsão. A previsão é a pretensão de antecipar a evolução no futuro das séries.

Para obter as previsões foi implementado um modelo linear generalizado para séries temporais de contagem, da biblioteca de funções *tscount* (Liboschik, Fokianos e Fried, 2017), que, recorde-se, se

[1] Variáveis com uma escala nominal ou ordinal em que cada categoria representa uma classe.

ADMISSAO_ID	EstacaoAno
A123	Verão
A124	Verão
A125	Outono
A126	Outono



ADMISSAO_ID	Verão	Outono	Inverno	Primavera
A123	1	0	0	0
A124	1	0	0	0
A125	0	1	0	0
A126	0	1	0	0

Figura 3.20: Exemplo de criação de variáveis *dummy* sobre as classes de estações do ano.

formula pelo seguinte

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{h=1}^{q} \alpha_h g(\lambda_{t-j_h}) + \boldsymbol{\eta}^\top \mathbf{X}_t,$$

definida na secção anterior.

Como já se verificou, este modelo pode ter componentes auto-regressivas da série (em Y_{t-i}), componentes de médias móveis da série (em λ_{t-j}) e co-variáveis (em X_t). Com vista a selecionar um modelo melhor (com AIC e RMSE menores), construíram-se vários modelos, na sua maioria representados na Figura 3.21. Estes modelos diferem em vários aspetos, nomeadamente nas variáveis que são incluídas, a quantidade de dados de treino utilizada ou a utilização de novas observações. A Figura 3.21 está assim ordenada, de cima para baixo.

Existem algumas outras escolhas que não estão representadas, nomeadamente ao nível de médias móveis e funções de ligação e transformação. Verificou-se que incluir médias móveis não acrescentava poder preditivo aos modelos e aumentava a dificuldade de interpretação do modelo, e ainda que usar funções de ligação e transformação diferentes da identidade também não favorecia os modelos ao nível do desempenho. Daí que se fixaram as funções de ligação e transformação como funções identidade e decidiu-se não incluir médias móveis.

Foi escolhido, à partida, o tempo para o qual são efetuadas as previsões - 10 dias (240 horas) - aquele usado como teste do modelo.

As variáveis e as componentes de auto-regressividade que faziam sentido foram inseridas no modelo e depois, por tentativa, foram retiradas aquelas cuja a ausência fazia diminuir as medidas de seleção, tendo em conta o respetivo coeficiente. A escolha recaiu num balanço entre menores valores de AIC simultaneamente com menores valores de RMSE e também na pretensão empresarial. O modelo escolhido foi o modelo 7 com RMSE de 5.50 e AIC de 189056.78 (Figura 3.21); apesar dos modelos 4, 5, 9 e 10 terem valores de RMSE menores, têm valores de AIC maiores. Apenas do modelo 8 resultaram ambas as medida de desempenho melhores mas optou-se pelo modelo 7 por incluir a variável da temperatura, o que do ponto de vista do estágio é benéfico pois traz mais complexidade na implementação, como é explicado no próximo capítulo.

MODELO	VARIÁVEIS EXTERNAS																		AUTO-REGRESSIVIDADE						DADOS			DESSEMPENHO			
	T	H	V	F1	F2	F3	F4	EI	EV	EO	EP	D1	D2	D3	D4	D5	D6	D7	1	2	24	48	168	8760	TR	TE	NO	AIC	RMSE	TTR	TTE
Modelo 1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	4A	10D		190073.56	5.67	16m	13s
Modelo 2	•			•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	4A	10D		189483.89	5.54	2m	15s
Modelo 3	•			•	•	•		•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	4A	10D		189862.71	5.59	4m	14s
Modelo 4	•				•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	4A	10D		189457.41	5.48	3m	13s
Modelo 5	•				•	•		•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	4A	10D		189366.00	5.49	4m	14s
Modelo 6	•				•	•		•		•	•	•	•	•	•	•	•		•		•	•	•	•	4A	10D		189519.57	5.83	2m	13s
Modelo 7	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	4A	10D		189056.78	5.56	2m	13s
Modelo 8					•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	4A	10D		188757.37	5.35	2m	13s
Modelo 9					•	•		•	•	•		•	•	•	•	•	•		•		•	•	•		4A	10D		189255.23	5.50	2m	13s
Modelo 10	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•		4A	10D		189719.26	5.53	2m	13s
Modelo 11	•				•	•	•	•	•	•		•	•	•	•	•	•		•		•	•	•		4A	10D		189275.52	5.62	2m	13s
Modelo 12	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	3A	10D		142415.80	5.55	1m	13s
Modelo 13	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	2A	10D		95534.60	5.61	1m	13s
Modelo 14	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	1A	10D		48107.21	5.85	41s	13s
Modelo 7	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	4A	20D		189056.78	5.74	2m	26s
Modelo 7	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	4A	10D	10D	189056.78	4.56	2m	0.2s
Modelo 7	•				•	•		•	•	•		•	•	•	•	•	•		•		•	•	•	•	4A	10D	20D	189056.78	5.24	2m	0.4s

Legenda (da esquerda para a direita, de cima para baixo)

T – temperatura; H – humidade; V – velocidade do vento; F1 – dia anterior a feriado; F2 – feriado; F3 – dia posterior a feriado; F4 – outro dia; EI – Inverno; EV – Verão; EO – Outono; EP – Primavera; D1 – Domingo; D2 – Segunda-feira; D3 – Terça-feira; D4 – Quarta-feira; D5 – Quinta-feira; D6 – Sexta-feira; D7 – Sábado; 1 - número de admissões de 1 hora antes; 2 – número de admissões de 2 horas antes; 24 – número de admissões de 24 horas (1 dia) antes; 48 – número de admissões de 48 horas (2 dias) antes; 168 – número de admissões de 168 horas (1 semana) antes; 8760 – número de admissões de 8760 horas (1 ano) antes; TR – quantidade de dados para treino; TE – quantidade de dados para teste; NO – quantidade de novas observações; TTR – tempo de duração do treino do modelo; TTE – tempo de duração do teste do modelo; A – anos; D – dias; m – minutos; s – segundos.

Figura 3.21: Tabela resumo dos testes de modelos.

Também ao nível dos dados de treino, experimentaram-se modelos com as mesmas componentes deste modelo 7 diferindo apenas a quantidade de dados de treino, originando os modelos 12, 13 e 14. No entanto, escolheu-se manter o modelo 7 uma vez que, repare-se, ao diminuir a quantidade de dados do treino do modelo, o valor de AIC diminui substancialmente, ao contrário do RMSE que aumenta - isto deve-se ao facto do AIC ser uma medida do ajuste do modelo aos dados e o RMSE ser uma medida de precisão da previsão e, com menos dados no treino, o modelo ajusta-se mais aos dados e, por isso, no teste tem um erro maior. Poder-se-ia considerar o modelo 12 pois ambas as medidas melhoram mas, como o valor de RMSE apenas melhora em uma centésima, poderá ter sido por ser determinados dados e, com dados novos, essa melhoria pode não se verificar.

O modelo resultante tem a forma

$$\begin{aligned}\lambda_t =& 0.096 + 0.298Y_{t-24} + 0.199Y_{t-48} + 0.373Y_{t-168} + 0.107Y_{t-8760} + 0.002temperatura \\ &+0.293feriado + 0.698posferiado + 0.335inverno + 0.014verao + 0.081outono \\ &+0.059domingo + 1.433segunda + 0.037terca + 0.241quarta + 0.082quinta + 0.242sexta\end{aligned}$$

Do modelo selecionado, acima referido, resulta o cronograma na Figura 3.22 onde é possível comparar o ajuste do modelo com os valores da série originais, incluindo os intervalos de confiança para as previsões. Verifica-se que as previsões seguem as tendências da série original mas que o modelo tende a efetuar previsões por defeito, tanto nos máximos como nos mínimos, o que pode ser um ponto fraco.

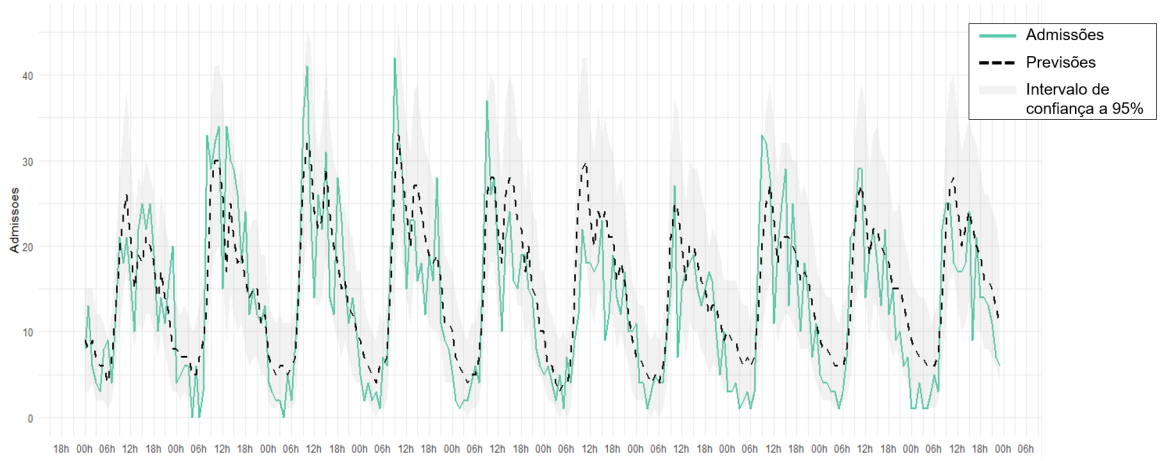


Figura 3.22: Cronograma da série temporal do número de admissões por hora, do ajuste do modelo e seus intervalos de confiança, desde 1 de janeiro de 2018 a 10 de janeiro de 2018.

Em relação às novas observações, elas tratam-se de um argumento da implementação em *R* que é muito importante aquando da utilização do modelo a "longo prazo". As novas observações podem ser inseridas para colmatar o problema associado a esta abordagem temporal em que são efetuadas previsões imediatamente a seguir à série modelada, ou seja, neste caso 240 previsões (10 dias) a partir da última hora que foi considerada no treino do modelo (Figura 3.23 em cima). Em dados atualizados, passando ao modelo novas observações que decorreram após a série usada para o treino, o modelo já treinado com a série original, tendo as novas observações, não é necessário treinar o modelo novamente e as novas observações são consideradas para a previsão, que começa a partir da última nova observação (Figura 3.23 em baixo).

Como trabalho adicional ao estágio, foram criados modelos com as mesmas variáveis do modelo 7 para cada uma das seis cores de pulseira, separando o número de admissões por cada uma das cores, com treino sobre 2 anos de dados. Depois, efetuou-se a soma das previsões de cada um dos modelos, da qual resulta uma previsão equivalente ao modelo anteriormente falado com o total de admissões, e verificou-se que a soma das admissões não é melhor que o modelo geral (Tabela 3.3).

Uma particularidade analisada nesta experiência foi ao nível dos resultados do modelo com menor RMSE, da cor vermelha. Verificou-se que o modelo previa sempre a zero, conseguindo com isso um

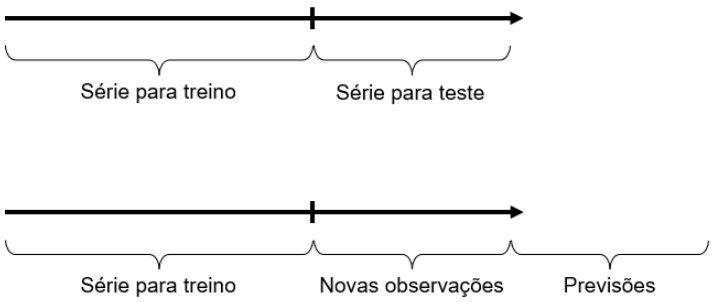


Figura 3.23: Ilustração da utilidade das novas observações na previsão com o modelo treinado.

Modelo	**AIC**	**RMSE**
Amarelo	78130.26	3.65
Azul	7680.56	0.27
Branco	12767.26	0.52
Laranja	48806.91	1.55
Verde	63241.80	2.87
Vermelho	3979.40	0.13
Soma das cores		5.78
Modelo total (2 anos de treino)	88292.93	5.84
Modelo total (4 anos de treino)	189056.78	5.56

Tabela 3.3: Resultados dos modelos para o número de admissões de cada cor de pulseira e comparação com modelos sem discriminação de cor.

menor RMSE, dado que as admissões de cor vermelha rondavam o zero ou um.

Capítulo 4

Implementação no Meliora

A modelação foi efetuada por forma a obter as previsões mas estas acabam por ser apenas números que necessitam de uma apresentação aprimorada e contextualizada para que façam sentido. Assim, os resultados do modelo foram incluídos num *dashboard* no Meliora que está já a ser utilizado pela Unidade de Saúde. Um *dashboard*, traduzido para português como "painel de bordo", é uma tela em que constam indicadores de desempenho (KPI's) relevantes para um determinado objetivo de forma organizada e apelativa. O Meliora é a plataforma da Prologica em que são disponibilizados todos os dados, algoritmos e ferramentas para equipar pessoas e organizações com as respostas necessárias para melhorar suas operações (Prologica, 2018). Ao longo deste capítulo apresenta-se todo o procedimento de preparação das ferramentas necessárias e da concessão do *layout* final.

4.1 Procedimento

Para efetuar a implementação dos resultados do modelo no Meliora, o código de R teve de ser organizado em diversos ficheiros *.r*: *01_basedados.r*, *02_api.r*, *03_treino.r* e *04_previsao.r*. Paralelamente, criaram-se as tabelas, através de linguagem SQL[1], necessárias em base de dados para que o processo pudesse assentar numa estrutura de dados, de modo a que fossem lidos os dados necessários e guardados os dados gerados: *Admissoes*, *PrevisaoAdmissoes*, *HistoricoAdmissoes*, *HistoricoPrevisaoAdmissoes*.

É importante referir que, não sabendo a temperatura apriori das horas seguintes para as quais são efetuadas as previsões, estabelece-se uma ligação à API Wunderground (TWC Product and Technology LLC, 2019) para obter as temperaturas atuais e previsões das temperaturas. API, sigla inglesa para Application Programming Interface, significa em em português Interface de Programação de Aplicações. Resumidamente, uma API é um conector que faz a interligação entre diferentes sistemas com linguagens de programação distintas, de forma rápida e com toda a segurança necessária (PTisp, 2019). Dessa ligação entre o R e a API, efetuada em tempo real, são obtidos dados em formato JSON[2]. Este tipo de formato é bastante útil para extrair informação pois esta está armazenada de forma organizada e fácil de aceder.

Em relação ao código, o ficheiro *01_basedados.r* foi criado de modo a estabelecer as ligações com as diversas tabelas de dados, seja de leitura ou escrita, enquanto o ficheiro *02_api.r* contém funções de ligação à api da temperutura e funções de preparação dos dados. Estes dois ficheiros não são executados mas sim importados pelos outros abaixo mencionados.

Quanto aos ficheiros preparados para serem executados, o *03_treino.r* atualiza o histórico de admissões, reestruturando informação nova, e treina o modelo com os dados atualizados, guardando o modelo treinado no ficheiro *modelo.rda*. Enquanto o *04_previsao.r* acede à API e efetua o registo da temperatura atual para suporte do treino, acede a previsões de temperatura para a previsão e, acedendo ao modelo previamente treinado, efetua as previsões e guarda-as.

Serve a Figura 4.1 para ilustrar as interações descritas acima.

[1] Sigla de Structured Query Language, em português Linguagem de Consulta Estruturada, é uma linguagem de pesquisa declarativa padrão para bases de dados relacionais.

[2] Abreviatura de JavaScript Object Notation.

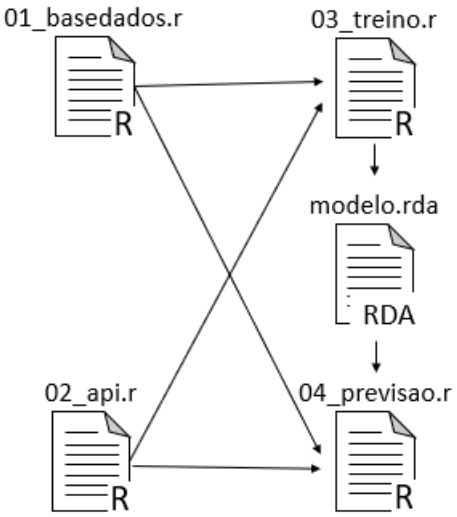


Figura 4.1: Diagrama de objetos do processo de previsão no Meliora.

Variável	**Descrição**
Data	Data do dia a considerar
Hora	Hora a considerar
Admissões	Número de admissões acumuladas na data e hora indicadas
Temperatura	Temperatura ambiente do local da Unidade de Saúde na data e hora indicadas

Tabela 4.1: Estrutura da tabela *Admissoes*.

Quanto às tabelas, no caso da tabela *Admissoes* (cujas variáveis se encontram descritas na Tabela 4.1), existe um processo de ETL[3] que atualiza a tabela com os dados da Unidade Local de Saúde, com a exceção da temperatura que é inserida com a execução do ficheiro *04_previsao.r* e respetiva ligação à API. Desta forma, esta tabela tem como finalidade ter os dados mais recentes (por processar para histórico) de número de admissões extraídos da base de dados da Unidade de Saúde.

A tabela *HistoricoAdmissoes*, como o próprio nome indica, contém toda a informação histórica (passada) já estruturada de forma a ser usada para o treino do modelo. Esta é uma tabela algo extensa pois tem a particularidade de guardar as variáveis *dummy* referentes às co-variáveis categóricas, de modo a integrarem o modelo (Tabela 4.2).

As tabelas *PrevisaoAdmissoes* e *HistoricoPrevisaoAdmissoes* têm a mesma estrutura (Tabela 4.3) pois esta segunda é o acumular da informação histórica da primeira. No fundo, nesta tabela de histórico é guardada a última previsão efetuada para uma certa hora de um certo dia. Em relação à tabela *PrevisaoAdmissoes* esta contém a informação que provém da previsão do número de admissões para cada hora de 10 dias em diante e de outras informações úteis para a construção do *dashboard* do Meliora. É, portanto, desta tabela que o Meliora vai recolher a informação que constrói o *dashboard*.

Foram também desenhados diagramas de sequência (Figura 4.2) que mostram a interação entre as tabelas e que ilustram a forma como a implementação do modelo no Meliora se torna automática em vários aspetos, tanto ao nível de atualização dos dados e do *dashboard*, como do treino do modelo e a atualização de parâmetros.

No que diz respeito ao treino do modelo, a execução do ficheiro *03_treino.r* retira os dados da tabela *Admissoes*, apagando-os, e reestrutura-os com a estrutura necessária a inseri-los na tabela *HistoricoAdmissoes*, de modo a atualizá-la. Posteriormente, lêem-se os dados atualizados dessa mesma tabela e procede-se ao treino do modelo (com as variáveis selecionadas e, como tal, apenas há atualização dos parâmetros),

[3]Do inglês, Extract Transform Load, são processos cuja função é a extração de dados de diversos sistemas, transformação desses dados e carregamento dos mesmo para as tabelas de trabalho.

Variável	Descrição
DataHora	Data do dia e hora considerada
Admissoes	Número de admissões acumuladas na data e hora indicadas
Feriado	Indicação se a data corresponde a um dia anterior a um feriado, a um feriado, a um dia posterior a um feriado ou a nenhum dos anteriores
EstacaoAno	Indicação da estação do ano a que pertence a data correspondente
DiaSemana	Indicação do dia da semana a que pertence a data correspondente
Feriado1	Variável *dummy* para a data ser anterior a um feriado
Feriado2	Variável *dummy* para a data ser um feriado
Feriado3	Variável *dummy* para a data ser posterior a um feriado
Feriado4	Variável *dummy* para a data ser nenhuma das anteriores
Inverno	Variável *dummy* para a data ser no inverno
Primavera	Variável *dummy* para a data ser na primavera
Verao	Variável *dummy* para a data ser no verão
Outono	Variável *dummy* para a data ser no outono
DiaSemana1	Variável *dummy* para a data ser um domingo
DiaSemana2	Variável *dummy* para a data ser uma segunda-feira
DiaSemana3	Variável *dummy* para a data ser uma terça-feira
DiaSemana4	Variável *dummy* para a data ser uma quarta-feira
DiaSemana5	Variável *dummy* para a data ser uma quinta-feira
DiaSemana6	Variável *dummy* para a data ser uma sexta-feira
DiaSemana7	Variável *dummy* para a data ser uma sábado
Temperatura	Temperatura ambiente do local da Unidade de Saúde na data e hora indicadas

Tabela 4.2: Estrutura da tabela *HistoricoAdmissoes*.

Variável	Descrição
Dia	Data do dia considerada
Hora	Hora considerada
DiaSemana	Dia da semana da data considerada
TipodeDia	Descrição do dia em dia da semana ou fim de semana/ feriado
Previsao	Número de admissões previstas na data e hora indicadas
LIC	Extremo inferior do intervalo de confiança da previsão a 95%
UIC	Extremo superior do intervalo de confiança da previsão a 95%

Tabela 4.3: Estrutura das tabelas *PrevisaoAdmissoes* e *HistoricoPrevisaoAdmissoes*.

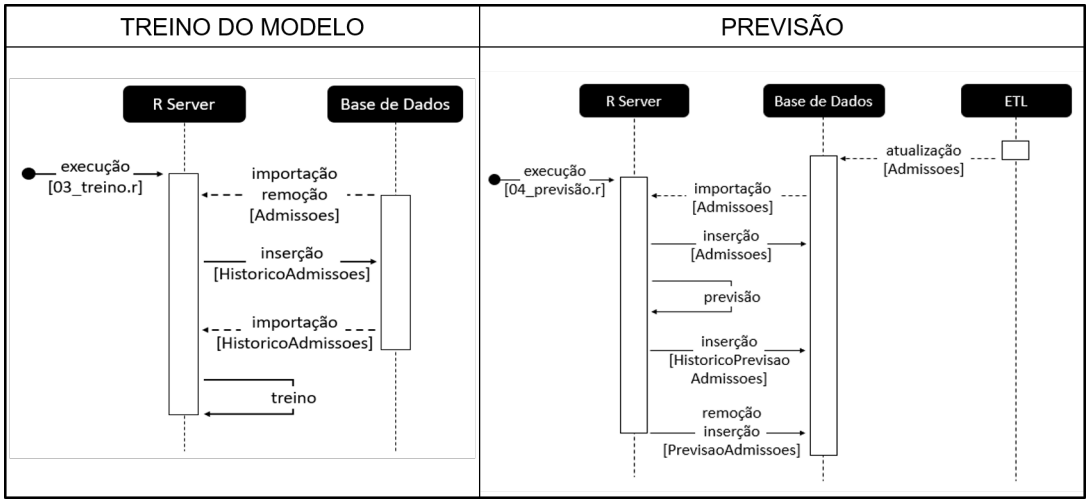


Figura 4.2: Diagramas de sequência da interação entre código e tabelas de dados.

guardando-o no ficheiro *modelo.rda*. Para as previsões, tendo a informação das novas admissões atualizada, a execução de *04_previsao.r* faz leitura dos dados de *Admissoes* para ter acesso às novas admissões desde o treino, inserção da temperatura atual na mesma tabela para que fique registada, previsão através do modelo treinado posteriormente, cópia das previsões anteriores para *HistoricoPrevisaoAdmissoes* e inserção das novas previsões em *PrevisaoAdmissoes*.

Note-se que os ficheiro *03_treino.r* e *04_previsão.r* são executados e é nessa execução que se baseia o diagrama. A sua execução, apesar de cruzar informações, é independente, no sentido em que o ficheiro *03_treino.r* pode ser executado mês a mês e o *04_previsão.r* hora a hora.

4.2 Meliora

Os resultados do modelo, depois de guardados na tabela *PrevisaoAdmissoes*, são usados no *dashboard* (Figura 4.3).

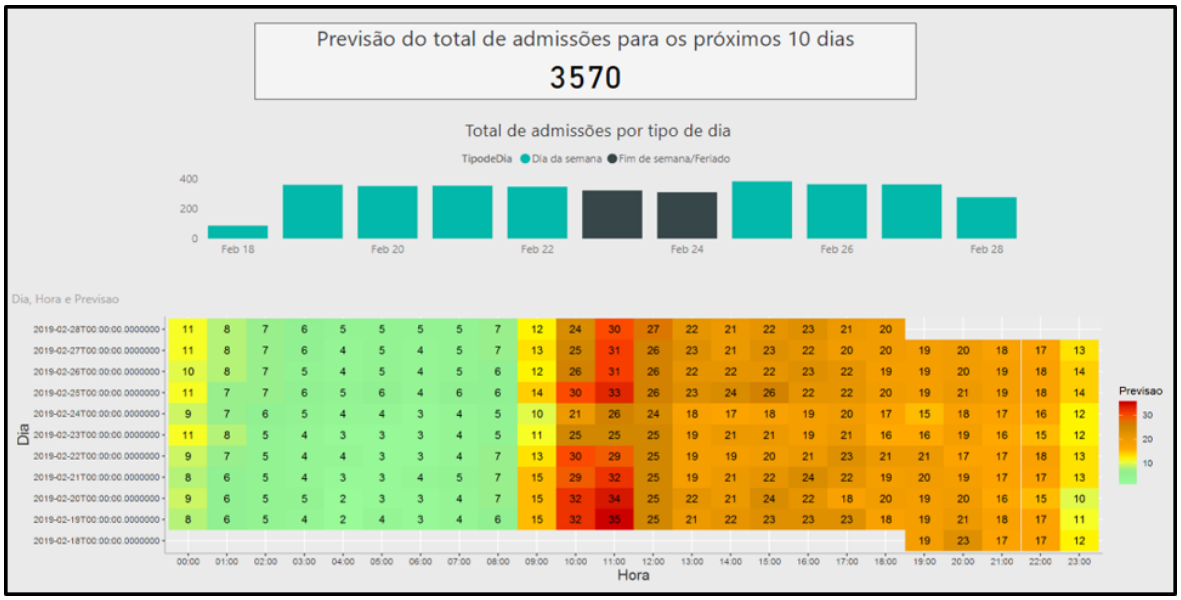


Figura 4.3: Dashboard do Meliora com resultados do modelo de previsão do número de admissões.

Esta plataforma Meliora tem a possibilidade de utilização da ferramenta da Microsoft *PowerBI* (Microsoft, 2019) para a construção dos *dashboards*. Assim, o *dashboard* é facilmente criado fazendo uma ligação a uma tabela, neste caso a *PrevisaoAdmissoes*, e escolhendo os visuais.

O *dashboard* está dividido em três partes, com informações de diferentes granularidades. Numa primeira parte está inserido um chamado cartão que contém um único valor, que corresponde ao número total de admissões previstas nos 10 dias.

O segundo visual, um gráfico de barras, tem representado o número de admissões previstas por dia, com a informação adicional do tipo de dia de que se trata (dia útil ou fim de semana/feriado).

O último visual, aquele que mais pormenor fornece e que faz jus ao propósito do trabalho, é um *heatmap* com o número de admissões previstas a cada hora e a cada dia em 10 dias. Este tem uma escala de cores que, visualmente, permite uma análise mais imediata, o que é bastante útil.

Uma particularidade que este visual tem é que, ao contrário dos outros visuais que são do *PowerBI*, este *heatmap* é integrado no mesmo *software* através de código *R*. Esta é também uma vantagem do *PowerBI*, para além de ser de fácil utilização, tem a possibilidade de integração de visuais de *R*.

Capítulo 5

Conclusão

O objetivo do estágio resume-se a estudar a aplicabilidade de metodologias de modelação na realidade hospitalar, nomeadamente na previsão do número de admissões num Serviço de Urgência, e a desenvolver protótipos.

Inicialmente houve um estudo exaustivo das metodologias mais usadas neste contexto especificamente, como se apresentou no capítulo 2. Para além das metodologias mais clássicas, apresentadas nesse capítulo, foram ainda consideradas outras alternativas: as regressões lineares generalizadas para séries temporais de contagem e um método de aprendizagem computacional - redes neuronais recorrentes, as LSTM (Long Short-Term Memory).

Na necessidade de algum pragmatismo, escolheu-se aquele mais facilmente interpretável e com um bom ajuste, a regressão linear generalizada para séries temporais de contagem.

Considera-se que o modelo escolhido obteve bons resultados tendo em conta o seu ajuste aos dados originais, obtendo um valor de RMSE de 5.56 admissões. Apesar do objetivo do trabalho ser muito específico pois é uma previsão horária, o modelo tem uma versatilidade muito maior em relação aos modelo clássicos, dado que permite usar características da série e também co-variáveis.

Um dos pontos fracos do modelo é que há uma tendência de previsão por defeito, como já havia sido referido, o que, do ponto de visto da alocação de recursos, poderá ser menos bom e levar a tempos de espera maiores.

É certo que não é fácil que um modelo se ajuste de tal forma que esteja a par e passo com os valores reais; mas a utilização de modelos que, na prática, podem ser mais fortes a detetar padrões, independentemente da interpretação, poderá ser vantajosa neste caso.

Também o protótipo visual dos resultados do modelo foi bem sucedido já que foi possível implementá-lo diretamente no Meliora 4.3 de tal modo que começasse a ser imediatamente apresentado e utilizado pela Unidade de Saúde.

Houve algumas dificuldades que se destacaram no decorrer do trabalho. Uma das maiores dificuldades foi a falta de conhecimento em séries temporais. Esta foi ultrapassada com a ajuda paciente de Professores e com a leitura paciente de livros sobre a matéria. Uma outra limitação foi o facto dos dados não estarem estruturados de forma a que fossem analisados, o que acabou por ser um desafio bastante interessante.

O trabalho futuro prende-se em aprofundar a previsão por cor da pulseira e efetuar previsões por área da urgência (Geral, Obstetrícia, Pediatria). Por serem previsões mais finas, permitirão previsões mais precisas.

Considero que foi uma mais valia poder trabalhar neste projeto, neste tema de conceito simples mas execução não tão simples.

A aprendizagem foi constante ao nível dos fluxos e dados das Urgências, de séries temporais, de processamento de dados, de SQL, de análise e desenvolvimento de processos.

Foram valorizados e desenvolvidos tanto a autonomia e a polivalência, como o trabalho de equipa e o ambiente empresarial.

Referências

Adhikari, Ratnadip e R. K. Agrawal (2013). *An Introductory Study on Time Series Modeling and Forecasting*. LAP Lambert Academic Publishing. DOI: 10.13140/2.1.2771.8084.

Administração Central do Sistema de Saúde (2018). *Termos de Referência para Contratualização de Cuidados de Saúde no SNS para 2019*.

Batal, Holly et al. (2001). "Predicting Patient Visits to an Urgent Care Clinic Using Calendar Variables". Em: *Academic Emergency Medicine* 8.1, pp. 48–53.

Boyle, Justin, Melanie Jessup et al. (2011). "Predicting emergency department admissions". Em: *Emergency Medicine Journal* 29.5, pp. 358–365. DOI: 10.1136/emj.2010.103531.

Boyle, Justin, Marianne Wallis et al. (2008). "Regression Forecasting of Patient Admission Data". Em: *30th Annual International IEEE EMBS Conference* 38, pp. 19–22. DOI: 10.1109/IEMBS.2008.4650041.

Calegari, Rafael et al. (2016). "Forecasting Daily Volume and Acuity of Patients in the Emergency Department". Em: *Computational and Mathematical Methods in Medicine* 2016. DOI: 10.1155/2016/3863268.

Champion, Robert et al. (2007). "Forecasting emergency department presentations". Em: *Australian Health Review* 31.1, pp. 83–90. DOI: 10.1071/AH070083.

Díaz, J. et al. (2001). "A Model for Forecasting Emergency Hospital Admissions: Effect of Environmental Variables". Em: *Journal of environmental Health* 64.3, pp. 9–15.

Díaz-Hierro, José et al. (2012). "Evaluation of time-series models for forecasting demand for emergency health care services". Em: *Emergencias* 24.3, pp. 181–188.

Entidade Reguladora da Saúde (2011). *Estudo sobre a organização e desempenho das Unidades Locais de Saúde - Relatório Preliminar I*.

Grupo Português de Triagem (2015). *Uma Metodologia de Trabalho Coerente*. URL: http://www.grupoportuguestriagem.pt/ (acedido em 27/03/2019).

Harvey, Andrew C. (1993). *Time Series Models*. Harvester Wheatsheaf.

Heij, Christiaan et al. (2004). *Econometric Methods with Applications in Business and Economics*. Oxford University Press.

Jones, Spencer S. (2007). "Real-Time Demand Forecasting in the Emergency Department". Em: *AMIA Symposium Proceedings*, pp. 997–998.

Jones, Spencer S. et al. (2008). "Forecasting Daily Patient Volumes in the Emergency Department". Em: *Academic Emergency Medicine* 15.2, pp. 159–169. DOI: 10.1111/j.1553-2712.2007.00032.x.

Liboschik, Tobias, Konstantinos Fokianos e Roland Fried (2017). "tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models". Em: *Journal of Statistical Software* 82.5. DOI: 10.18637/jss.v082.i05.

Marcilio, Izabel, Shakoor Hajat e Nelson Gouveia (2013). "Forecasting Daily Emergency Department Visits Using Calendar Variables and Ambient Temperature Readings". Em: *Academic Emergency Medicine* 20.8, pp. 769–777. DOI: 10.1111/acem.12182.

McCarthy, Melissa L. et al. (2008). "The Challenge of Predicting Demand for Emergency Department Services". Em: *Clinical Practice* 15.4, pp. 337–346. DOI: 10.1111/j.1553-2712.2008.00083.x.

Microsoft (2019). *Business intelligence like never before*. URL: %5Chref%7Bhttps://powerbi.microsoft.com/en-us/%7D (acedido em 21/02/2019).

Montgomery, Douglas C., Elizabeth A. Peck e G. Geoffrey Vining (2006). *Introduction to Linear Regression Analysis*. Wiley-Interscience.

Murteira, Bento J. F., Daniel A. Müler e K. Feridun Turkman (1993). *Análise de Sucessões Cronológicas*. McGRAW-HILL.

Prologica (2018). *Prologica | Soluções de analítica para melhores decisões*. URL: https://www.prologica.pt/pt/ (acedido em 21/02/2019).

PTisp (2019). *Sabe o que é uma API (Application Programming Interface)?* URL: https://pplware.sapo.pt/high-tech/sabe-o-que-e-uma-api-application-programming-interface/ (acedido em 12/02/2019).

RStudio (2018). *RStudio*. URL: https://www.rstudio.com/ (acedido em 17/12/2018).

Sá, Armando Brito de (2002). "Urgência hospitalar e Cuidados de Saúde Primários: mitos e falácias". Em: *Revista Portuguesa de Clínica Geral*, pp. 347–348.

Serviço Nacional de Saúde (2017). *Entidades de Saúde*. URL: https://www.sns.gov.pt/institucional/entidades-de-saude/ (acedido em 15/03/2019).

The R Foundation (2018). *The R Project for Statistical Computing*. URL: https://www.r-project.org/ (acedido em 17/12/2018).

TWC Product and Technology LLC (2019). *Weather Forecast Reports*. URL: http://api.wunderground.com (acedido em 12/02/2019).

Zhao, Lei e Bernt Lie (2010). "Modeling and Simulation of Patient Flow in Hospitals for Resource Utilization". Em: *Simulation Notes Europe* 20.2. DOI: 10.11128/sne.20.tn.09976.