UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Hospitalizations due to Diabetes in Portugal: a time series analysis

Mónica Daniela Santos Fialho

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Prof.ª Doutora Marília Antunes

2020

ACKNOWLEDGEMENTS

À Professora Doutora Marília Antunes, pela disponibilidade na orientação do presente trabalho, bem como por todos os conselhos e ensinamentos transmitidos, o meu sincero obrigado.

À Administração Central do Sistema de Saúde, I.P., no âmbito da colaboração com a Faculdade de Ciências da Universidade de Lisboa, agradeço a cedência dos dados, sem os quais não teria sido possível realizar este trabalho. Devo também um obrigado aos meus colegas de mestrado, mas sobretudo ao David, por toda a ajuda ao longo do curso. Este agradecimento estende-se a colegas de trabalho e amigos, pelo apoio demonstrado em todos os momentos. Um obrigado especial à Ana e ao Osvaldo, que mesmo à distância sempre estiveram disponíveis para me ajudar.

Por fim, o meu muito obrigado à minha família e ao Alexandre, por tudo aquilo que representam para mim.

RESUMO

A diabetes é uma doença metabólica multifatorial, caracterizada por níveis elevados de glucose no sangue (hiperglicemia). Esta é uma condição crónica, resultante da progressiva destruição ou disfunção das células beta presentes no pâncreas, onde se dá a produção de insulina, uma hormona anabólica envolvida na absorção e metabolismo da glucose. Sintomas clássicos de diabetes incluem fome e sede excessivas (polifagia e polidipsia, respetivamente), vontade frequente de urinar (poliúria) e cansaço. Se o nível de glucose no sangue estiver continuamente acima do normal, podem ocorrer manifestações mais graves da doença, como sejam cetoacidose e coma hiperosmolar, com perigo de morte associado. A longo-prazo, indivíduos com diabetes têm um risco aumentado de complicações micro e macro vasculares, tais como retinopatia, nefropatia, neuropatia e doenças cardiovasculares. O controlo da doença é feito através de tratamento farmacológico, aliado a um estilo de vida saudável, procurando evitar ou retardar, tanto quanto possível, consequências graves. No entanto, uma baixa-autopercepção do risco de complicações e falhas no acompanhamento destes doentes ao nível dos cuidados de saúde primários contribuem para piores resultados clínicos, com eventual necessidade de cuidados de saúde hospitalares.

O presente trabalho teve por objetivo descrever e modelar uma série temporal de internamentos hospitalares por diabetes em Portugal, com ênfase na predição. Para tal, foram usados dados constantes da Base de dados de Morbilidade Hospitalar, cedida pela Administração Central do Sistema de Saúde (ACSS), I.P., do Ministério da Saúde. Foram selecionados todos os diagnósticos de diabetes como causa primária de admissão, codificados, até ao terceiro dígito, por 250 (diabetes *mellitus*), de acordo com a Classificação Internacional de Doenças (ICD), 9ª revisão, Modificação Clínica (ICD-9-CM) ou E10 (diabetes tipo 1), E11 (diabetes tipo 2), E13 (outro tipo de diabetes), segundo a 10ª revisão da ICD (ICD-10-CM/PCS). Cada um destes registos foi associado a um episódio específico, através de um número sequencial único entre bases de dados, selecionando-se aqueles com data de admissão entre 1 de janeiro de 2010 e 31 de dezembro de 2018 e internamento mínimo de um dia. Com base nestes dados, foi construída uma série temporal do número mensal internamentos por diabetes entre 2010 e 2018, num total de 108 observações. Um subconjunto destes dados, composto por observações entre janeiro de 2010 e dezembro de 2016, foi utilizado na identificação e estimação do modelo (conjunto de treino; 84 meses), e as restantes observações, entre janeiro de 2017 e dezembro de 2018, usadas exclusivamente para validação do modelo (conjunto de teste; 24 meses). Seguindo a metodologia de Box e Jenkins para modelos Autorregressivos e de Médias Móveis Integrados Sazonais (SARIMA), vários modelos foram identificados com base na análise gráfica das funções de autocorrelação e autocorrelação parcial e estimados por máxima verosimilhança. Na seleção do melhor modelo, foi considerado o critério de informação de Akaike (AIC), avaliando-se posteriormente a sua adequação aos dados

através da estatística de Ljung-Box e inspeção visual dos resíduos. A capacidade preditiva do modelo selecionado foi investigada por comparação entre as previsões obtidas e os dados do conjunto de teste, através de um procedimento de avaliação de origem móvel, em que novos valores são imputados ao modelo à medida que, supostamente, se tornam conhecidos. Neste caso, considerou-se quer a atualização do modelo, quer a sua recalibração, tendo por base uma janela fixa, onde se incluem todas as observações disponíveis até ao momento, ou móvel, composta pelas 84 observações mais recentes. Medidas como o erro absoluto médio (MAE), a raiz do erro quadrático médio (RMSE) e o erro percentual absoluto médio (MAPE) foram utilizadas para quantificar a precisão do modelo em cada contexto de predição, permitindo a sua comparação com um método de referência, um passeio aleatório sazonal, segundo o qual cada previsão iguala o valor da série no mesmo mês do ano anterior.

Entre janeiro de 2010 e dezembro de 2018, foram contabilizados em Portugal 73.050 episódios de internamento por diabetes (676 casos por mês, em média), o que representa 35% de todas as admissões hospitalares por esta causa. Este número resulta, na sua maioria, de admissões urgentes (79,5%). A distribuição por sexo mostrou-se relativamente equilibrada (52,5% de homens), tendo sido observado um maior número de internamentos entre indivíduos com idade igual ou superior a 60 anos e na região Norte do país. Globalmente, o número de episódios diminuiu 45% entre 2010 e 2018 (10.011 e 5.530 internamentos, respetivamente). Para além da tendência decrescente, foram também observadas flutuações sazonais, com um pico de casos nos meses de Inverno e números mais baixos no Verão. Tendo por base o número de internamentos por mês entre 2010 e 2016, foram identificados e estimados nove modelos candidatos para a série original e diferenciada, quer na componente regular, quer sazonal. Entre estes, o modelo mais parcimonioso, SARIMA$(1,1,2)\times(0,1,1)_{12}$ (AIC = 10,647), foi usado para prever o número mensal de internamentos em 2017 e 2018. Considerando a disponibilidade de novos dados a cada mês, foi avaliada a capacidade preditiva do modelo para os horizontes temporais de 1, 3, 6 e 12 meses. De uma forma geral, o modelo re-estimado teve um melhor desempenho do que o modelo atualizado, registando-se o menor erro médio em previsões a um mês obtidas por meio de um janela móvel (MAE = 39,5; RMSE = 47,4; MAPE = 7,8%). Independentemente de ser usada uma janela fixa ou móvel na recalibração do modelo, a capacidade preditiva deste piorou com o aumento do horizonte temporal para 3, 6 e 12 meses. Em todo o caso, quer por via da atualização, quer da recalibração do modelo, foi observado um erro relativo inferior a 10% num horizonte temporal até seis meses. Foi ainda calculado o MAPE para 2017 e 2018, considerando a re-estimação do modelo com janela móvel a cada 1, 3, 6 e 12 meses. Neste caso, previsões a três meses apresentaram a maior precisão, com um erro anual médio de 7,7%, muito próximo do obtido com previsões a um mês (MAPE = 7,8%). Da re-estimação do modelo a cada 12 meses resultou o maior erro de previsão (MAPE = 12,4%), representando, ainda assim, uma redução de 30% face ao modelo de referência (MAPE = 17,7%). A representação gráfica das previsões do modelo SARIMA mostrou que estas ficaram maioritariamente acima da série observada, sobretudo em 2018. Não obstante, com a exceção de fevereiro de 2017 e maio de 2018, todos os valores observados se situaram entre os limites obtidos para as previsões. Estes resultados suportam a aplicação de modelos SARIMA na previsão de internamentos por diabetes em Portugal a curto/médio prazo, permitindo que decisões ao nível da gestão hospitalar sejam tomadas atempadamente. Esta seria uma forma de melhorar

a capacidade de resposta dos serviços de saúde, sobretudo em períodos de maior fluxo de pacientes. Permitiria, igualmente, um uso mais eficiente do orçamento, pela adequação de recursos às reais necessidades dos pacientes, sem comprometer a qualidade dos cuidados prestados.

Não desvirtuando o estudo de um ponto de vista clínico e epidemiológico, este tem associadas algumas limitações metodológicas. Começar por referir que foram incluídos apenas os diagnósticos principais de diabetes e os tipos mais comuns da doença, subestimando quer o seu impacto, quer a procura de cuidados de saúde específicos por esta causa. Por outro lado, o uso de dados agregados a nível nacional inviabiliza o uso das previsões obtidas em contextos reais de prática clínica. No que concerne à modelação dos dados, o uso do AIC como critério de seleção pode ter levado a uma interpretação errónea da qualidade dos modelos, dada a sua aplicação a diferentes conjuntos de dados (série original e diferenciada).

Como trabalho futuro neste campo de investigação, conta-se a realização de uma análise espaciotemporal de internamentos hospitalares por diabetes, segundo métodos Bayesianos hierárquicos de mapeamento de doenças, considerando características sociodemográficas da população e indicadores de acesso a cuidados de saúde a nível regional.

Palavras-chave: Modelos SARIMA, Previsão, Avaliação de origem móvel, Diabetes, Internamentos

ABSTRACT

Diabetes is a chronic disease characterized by high blood sugar levels, as a result of the progressive destruction or dysfunction of the pancreatic β-cells that produce insulin, an anabolic hormone involved in cellular glucose uptake and metabolism. Poor self-management and inefficient monitoring at primary health care contribute to an inadequate glycaemic control, which often leads patients to seek hospital health care, while facing acute or long-term complications of diabetes. The objective of this study is then to describe and model a series of hospitalizations due to diabetes in Portugal, with an emphasis on prediction.

Episodes of hospital admissions occurred between 2010 and 2018 with main diagnosis of diabetes, coded, up to the third digit, by 250, according to International Classification of Diseases (ICD), 9th Revision, Clinical Modification (ICD-9-CM), or E10, E11, E13, based on the 10th revision of ICD (ICD-10-CM/PCS), and duration of at least one day were selected from the Hospital Morbidity Databases provided by the Central Administration of the Health System (ACSS), I.P. Following the Box-Jenkins approach for Seasonal Autoregressive Integrated Moving Average (SARIMA) modelling, a time series analysis on monthly hospitalizations in Portugal from January 2010 to December 2018 was conducted. Using data from 2010 to 2016 (84 observations), several models were identified as suitable and estimated by maximum likelihood. Akaike´s information criterion (AIC) was used to select the best model, whose adequacy was further investigated by residual analysis. For the selected model, 1, 3, 6 and 12-month forecasts were computed and compared against the observed series in 2017 and 2018, based on rolling-origin-update and rolling-origin-recalibration evaluation, with either a fixed (all available data) or rolling window (data from the last 84 months). The predictive ability of this model was assessed using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE), and compared with a benchmark method, namely a Seasonal Random Walk.

From 2010 to 2018, there were 73,050 hospitalizations due to diabetes in Portugal, representing 35% of all admissions for this cause. The series of monthly hospitalizations exhibits a decreasing trend and apparent seasonality, with a higher number of episodes observed in winter months. From nine candidate models, the SARIMA$(1,1,2) \times (0,1,1)_{12}$ (AIC = 10.647) was selected as the most parsimonious and used to predict hospitalizations in 2017 and 2018. For both rolling-origin-update and rolling-origin-recalibration, the relative error was lower than 10% for a forecast horizon up to six months. Overall, rolling-origin-recalibration performed better, with the lowest MAPE obtained with one-month forecasts, given either a fixed or a rolling window (8.2% and 7.8%, respectively). As the forecast timespan increased, up to 3, 6 and 12-months, the predictive accuracy of the model worsened. The average error for 2017 and 2018, obtained by using a rolling window to re-estimate the model every 1, 3, 6 and 12 months, revealed predictions at three months as the most

accurate (MAPE = 7,7%), followed by those at one month (MAPE = 7,8%). The highest error was obtained with 12-month forecasts (MAPE = 12,4%), still representing a 30% reduction in relation to the benchmark model (MAPE = 17,7%). The graphical representation of the forecasts showed that the selected model often overestimated the observed series, yet, all but two observations were in the 95% prediction interval.

The selected model was able to capture the seasonal patterns of the series, revealing a good predictive ability up to six months. These findings suggest that SARIMA models can be used to forecast hospitalizations due to diabetes at short/medium term with good accuracy, allowing for management decisions to be taken timely. Future work on this field of research includes a spatiotemporal analysis of hospitalizations due to diabetes in Portugal, following a Bayesian hierarchical disease mapping approach, while taking into account population socioeconomic characteristics and access to health care.

Keywords: SARIMA, Forecasting, Accuracy, Rolling-origin, Diabetes, Hospitalizations

Contents

List of Figures

List of Tables

Acronyms

ACF	Autocorrelation Function
ACSS	Central Administration of the Health System (Administração Central do Sistema de Saúde)
ADF	Augmented Dickey-Fuller
AIC	Akaike´s Information Criterion
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ARS	Regional Health Administration (Administração Regional de Saúde)
BIC	Bayesian Information Criterion
CI	Confidence Interval
CVD	Cardiovascular Diseases
HbA1c	Glycosylated Haemoglobin
ICD	International Classification of Diseases
IQR	Interquartile Range
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLE	Maximum Likelihood Estimates
MRAE	Mean Relative Absolute Error
MSE	Mean Square Error
NUTS	Nomenclature of Territorial Units for Statistics

PACF	Partial Autocorrelation Function
Q-Q	Quantile-Quantile
RelMAE	Relative Mean Absolute Error
RMSE	Root Mean Square Error
RMSPE	Root Mean Square Percentage Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SD	Standard Deviation
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
WHO	World Health Organization

Chapter 1

Introduction

As defined by the World Health Organization (WHO), the term diabetes describes a group of chronic metabolic disorders characterized by high blood sugar levels (hyperglycaemia), as a result of the progressive destruction or dysfunction of β-cells present in pancreatic islets, where insulin is produced [1]. In normal conditions, insulin promotes the cellular uptake of glucose so it can be used to produce energy [2]. As the body becomes unable to produce or make proper use of this hormone, glucose starts to accumulate in the blood and the classical symptoms of diabetes turn up. These include excessive thirst (polydipsia), hunger (polyphagia) and urination (polyuria), blurred vision, fatigue, and weight loss. Severe manifestations of the disease include ketoacidosis, in type 1 diabetes (T1D), and nonketotic hyperosmolar coma, in type 2 diabetes (T2D), possible life-threatening conditions. These are the major types of diabetes, accounting for more than 95% of all cases. Other types of diabetes are defined by the WHO as: hybrid forms of diabetes, other specific types, unclassified diabetes and hyperglycaemia first detected during pregnancy, including diabetes mellitus in pregnancy and gestational diabetes mellitus. With the exception of the latter condition, for which lower cutoffs for plasma glucose are defined, the diagnose of diabetes requires that at least one of the following criteria be met: fasting plasma glucose $\geqslant 126$ mg/dl; 2-hour post-load plasma glucose $\geqslant 200$ mg/dl; glycosylated haemoglobin (HbA1c, a measure of the average blood glucose levels over the past two to three months) $\geqslant 6.5\%$ [1].

Further classification of the disease relays on clinical features such as age of onset, presence of diabetes-related biomarkers (pancreatic autoantibodies), and level of dysfunction of β-cells [1, 3]. The mechanisms involved in loss of β-cells mass and/or function are themselves multifactorial, including genetic and epi-genetic factors, autoimmunity, insulin resistance, co-existence illnesses, inflammation, and environmental factors [1]. Still, despite diabetes different types and rate of progression, chronic hyperglycaemia confers a higher risk for micro and macrovascular complications [4]. Retinopathy, nephropathy and neuropathy are some of the long-term effects of diabetes, along with an increased risk for other diseases, such as peripheral artery disease cardio and cerebrovascular disease, cataracts, nonalcoholic fatty liver disease and infectious

diseases [1]. Compared to non-diabetic people, people with diabetes have a two-fold increase in the risk of cardiovascular diseases (CVD), explaining the highest morbidity and mortality of the disease [2]. Besides CVD, diabetes is a modifiable risk factor for some types of cancer and dementia [5].

Type 1 diabetes

In T1D, pancreatic β-cells are attacked by the immune system due to genetic susceptibility and environmental/behavioural conditions (e.g., viral infections, exposure to toxins, dietary factors) [2, 3]. Although not fully understood, the effect of environmental factors and their interaction with genetic factors is supported by the evidence that having highest-risk HLA alleles is neither a necessary nor sufficient condition for developing T1D [3, 4]. Regardless, there is no evidence that this type of the disease can be prevented [3].

The first stage of the disease reassembles to islet autoimmunity (i.e., the presence of antibodies to pancreatic islet antigens) [6] when, despite the progressive destruction of β-cells, euglycaemia (i.e., blood glucose at levels considered normal and healthy) is maintained due to pancreatic 'functional' reserve. Further attack to pancreas, with destruction of most β-cells, results in a decrease of insulin production followed by an increase in the concentration of blood glucose, until diagnosis [7]. At this point, patients present the classical symptoms of diabetes and insulin administration is required, on a daily basis, to assure their well-being and, ultimately, their survival [2].

This type of diabetes is usually recognized by having its onset in childhood, but it can occur later in life, although the classic symptoms may not be observed [2, 3]. In fact, about 50% of cases occur in adulthood and up to 50% of those might be misclassified as T2D at first [3]. Heterogeneity with respect to pathomorphology of the pancreatic islet, severity of auto-immune response and efficacy of therapy is also observed in these patients [6, 7]. Nonetheless, a faster rate of destruction of β-cells is common in children and adolescents, with ketoacidosis being the first manifestation of the disease in some cases [1]. As for hyperglycaemia, also hypoglycaemic episodes, which result in 4 to 10% of T1D-related deaths, must be prevented by carefully monitoring glucose levels. An adequate glycaemic control depends on the correct dose adjustments given the quantity of carbohydrates consumed, the practice of physical activity, as well as the co-occurrence of illness and stress. Other than that, even without a cure, people with T1D can live a healthy and long life [3].

At long term, maintaining 'near-to-normal' glucose blood levels — international guidelines suggest as targets for adult and paediatric patients values of HbA1c lower than 7.0% and 7.5%, respectively — reduces micro (e.g., retinopathy, neuropathy, nephropathy) and macrovascular (e.g., atherosclerosis, cerebral and coronary heart disease) complications of this disease [3] and preserves any β-cell mass or function, thus contributing to a better quality of life, itself a predictor of a better glycaemic control [3, 7].

Type 2 diabetes

The pathophysiology of T2D is based on the body's inability to trigger an adequate response to insulin, a phenomenon known as insulin resistance. In an attempt to lower the levels of glucose in blood, the pancreas increases the production of insulin which, ultimately, results in the failure of β-cells [2]. This form of the disease is the result of genetic and epigenetic influences, along with environmental and behavioural factors (e.g., unhealthy diet, physical inactivity, smoking and drinking habits) [8–10], justifying its link with overweight and obesity. Such conditions, or just a significant accumulation of fat in abdominal region, either cause or exacerbate insulin resistance [1].

Visceral adiposity and obesity contribute to decreased insulin sensibility, initially compensated by hypersecretion of this hormone in β-cells (euglycaemic hyperinsulinaemia). Over time, obese people stop responding to the increased production of insulin, which results in the elevation of blood glucose concentration (hyperglycaemic hyperinsulinaemia). At a certain point, with the ongoing deterioration of β-cells, the over-stimulated pancreas becomes unable to secrete enough insulin and hyperglycaemia turns evident (hyperglycaemic hypoinsulinaemia) with subsequent diagnosis of T2D. From abnormal insulin sensitivity to the moment of clinical diagnosis, several years may pass [7]. Weight loss improves insulin sensitivity, but with limited effects — the reverse of long-standing diabetes is a difficult achievement, even with large weight loss, as observed after bariatric surgery [4]. Other medical conditions, directly related to behavioural factors, may make people more prone to T2D. The list includes hypertension and dyslipidemia (hypercholesterolemia and hypertriglyceridemia) [11]. Hence, the management of the disease depends greatly on the adoption of a healthy lifestyle, in addition to antidiabetic mediation [2].

The clinical pattern of T2D is generally less obvious than T1D, and the precise moment of its onset is difficult to determine. Thus, people tend to keep undiagnosed for long periods and some complications are already present when they are diagnosed [2]. Moreover, most patients have other chronic conditions, which makes it difficult to control diabetes and increases mortality [12]. The clinical manifestations of the disease include fatigue, lethargy, recurrent infections, and visual impairment that, in many cases, motivates the medical appointment that propitiates the diagnosis. With severe loss of β-cells, the classical symptoms observed in T1D then occur [8].

Although hypoglycaemia is more common in T1D, it can occur in T2D, accounting for the underlying morbidity of this disease. For inpatients, episodes of hypoglycaemia, even non-severe, are associated both with increased length of stay and in-hospital mortality [13]. The risk for hypoglycaemia increases with diabetes duration, multi-morbidity and use of specific medicines, such as sulfonylureas [12, 14]. With respect to microvascular complications, people with T2D benefit from intensive glycaemic control. The same observation is not so evident when it comes to macrovascular events, such as cardiovascular disease and stroke [14]. Also, T2D is associated with a wide range of cancers (e.g., breast, endometrial, colorectal, liver) [2], mental and nervous system disorders and infections [7].

1.1 Epidemiology of diabetes

Type 2 diabetes is the most common form of the disease, representing 90 and 95% of the cases worldwide [1]. Type 1 diabetes follows, accounting for 5 to 10% of all diabetes cases [15]. Together, they affect over 460 million adults aged 20-79 years worldwide, 59 million in Europe [2], representing an important risk factor for vascular diseases and early mortality [16–18]. These values refer to both diagnosed and undiagnosed cases, with the latter estimated to account for 40.7% of all cases in Europe Region (as defined by the International Diabetes Federation) [2]. In Portugal, diabetes affects 9.8% of the population aged between 20 and 79 years, above the age-adjusted prevalence of diabetes in Europe (6.3%). It is the third highest value in this region, only surpassed by Germany and Turkey [2].

Worldwide, the prevalence of diabetes is slightly higher in men (9.6%), compared to women (9.0%), being expected to increase in both groups according to projections for 2030 and 2045 [2]. In Portugal, the results from the first National Health examination Survey (INSEF 2015) point out a greater difference between men (12.1%) and women (7.8%) [11]. Differences between age groups are also noticeable, with higher prevalence of diabetes at older ages. Along with age, ethnicity, obesity, and family history, with a highly increased risk for people with first-degree relatives with the disease, are major risk factors for T2D [8] — in Portugal, 68% of people with diabetes reported having a first-degree family member with diabetes [11]. On the other hand, lower prevalences of diabetes are observed for those having more education and being employed, even after age-standardization [11]. As opposed, due to its impact in other aspects of life, low socioeconomic status increases the risk of developing T2D [4].

In the group of children and adolescents (0-19 years), the prevalence of T1D in Portugal was 15% in 2018 [19]. In Europe, it affects almost 300,000 of those aged under 20 years, with about 31,000 new cases per year, more than in any other regions of World. Nonetheless, the incidence of T1D has been increasing worldwide, with changes in non-genetic, lifestyle related, factors as the most probable cause. More, there is evidence of an increasing prevalence of T2D in this age group, posing a great burden for families and society, as these individuals will present complications sooner in life [2].

Briefly, ageing populations and unhealthy lifestyles, including a poor diet (rich in sugars, fat and calories) and physical inactivity, have contributed to the increase in the prevalence of diabetes and its complications, being the clinical evolution of patients largely influenced by education, access to health care and co-occurrence of risk factors [1, 2]. Lifestyle modification, with the adoption of healthy habits early in life, is essential for primary prevention of T2D [20, 21], whereas control of risk factors and early intervention on disease complications can prevent the need of hospitalization, reducing costs and improving quality of life [2].

In 2018, diabetes was the main cause of 3.8% of all deaths in Portugal [19], with more than a quarter of the people who die in hospitals suffering from this disease [22]. Worldwide, estimates point to 4.2 million deaths in 2019 as a result of diabetes and its complications [2].

1.2 Importance of primary health care in diabetes management

To achieve better results when facing the disease, people benefit from support systems that combine educational strategies, including glucose self-management, dietary counselling as well as programmes of physical exercise, and psychological support [14, 23, 24]. Likewise, patient-centred care, conducted by multidisciplinary teams (e.g., doctor, nurse, nutritionist, pharmacist, exercise physiologists, psychologists, social workers) considering individual specific needs (patient empowerment, along with ongoing support) has proved to improve clinical outcomes. In this regard, primary health care is of paramount importance. One of its most valuable aspects is its broad spectrum of action, coordinated among patients, family, community and other health care providers, in which multiple health determinants and risk factors for diabetes are considered [5]. Unfortunately, many patients still do not have access to this model of care [2, 5]. Taken as example, only half of individuals are subjected to risk assessment for T2D in primary health care in Portugal Mainland, and regional discrepancies exist (from 23% in Regional Health Administration (ARS) of Algarve and ARS of Lisboa e Vale do Tejo, to 48% in ARS Norte). Plus, ARS of Algarve and Lisboa e Vale do Tejo are the regions with higher percentage of patients without general practitioner (11.7% and 13.1%, respectively) and greater ratio of patients by doctor (1964 and 1957, respectively) [19].

Thus, despite the access to new drugs and a better understanding of disease pathology, the control of diabetes remains unsatisfactory, in part due to poor self-management and inefficient monitoring by health services, and hospital admissions become more frequent than would be desirable [2, 19]. The low perception of the risk of diabetes-related complications in T2D also contributes for this outcome [25], which negatively impacts individuals, health systems and, ultimately, society [26, 27]. In face of that, hospital admissions by diabetes deserve careful analysis, with particular emphasis on hospitalizations, to which most costs with diabetes are attributable (53% of identified costs with diabetes in 2017, in mainland Portugal) [19]. In 2017, hospitalizations with diabetes as main diagnosis represented a cost of almost 15 million Euro, greatly exceeded by costs with hospitalizations associated to diabetes, although not entirely attributable to it (361 M) [19]. Indirect costs result from disability, absenteeism and early mortality due to diabetes [2].

1.3 Hospital admissions for diabetes in Portugal

In Portugal, data regarding hospital admissions are made available by the Central Administration of the Health System (Administração Central do Sistema de Saúde, ACSS), I.P. of the Portuguese Ministry of Health in the Hospital Morbidity Databases. Each episode has associated one or more diagnoses, allowing its analysis in view of the problems that led to or result from the admission on the health facility.

From 2010 to 2018, there were 208,882 hospital admissions due to diabetes (i.e., diabetes as the primary cause of admission) in Portugal, more frequently diabetes with ophthalmic complications (65.6%).

Overall, admissions due to diabetes show an increasing trend between 2010 and 2017 (13,647 and 34,458 cases, respectively), with a small decrease in 2018 (33,042 cases), and mimic the evolution of ambulatory episodes (61% of all admissions). The same pattern is shown by planned admissions (71% of all admissions), as these are mostly represented by ambulatory episodes. Opposite trend is observed for the emergencies, with a 40% decrease from 7,959 records in 2010 to 4,808 in 2018 (Figure 1.1).

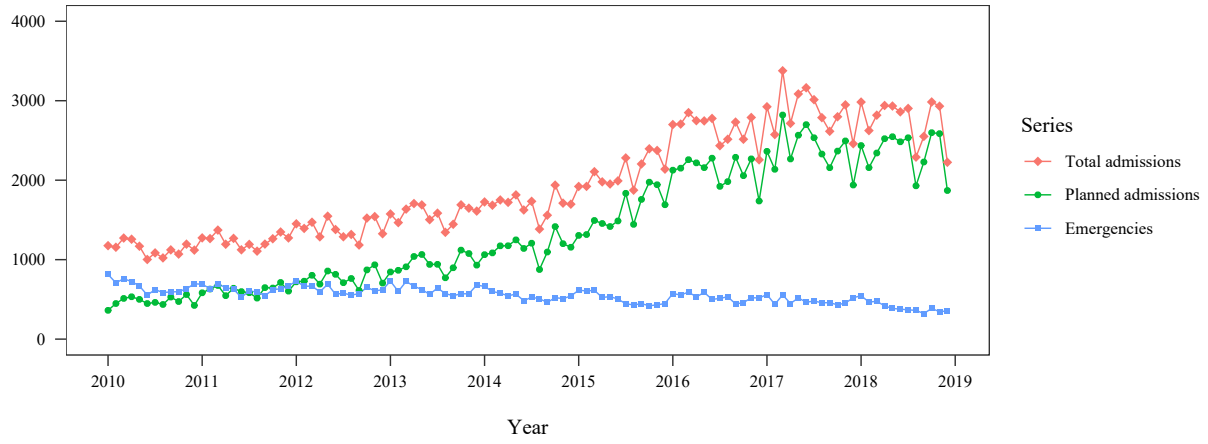

Figure 1.1: Series of hospital admissions due to diabetes in Portugal from 2010 to 2018: total admissions, planned admissions and emergencies.

1.4 Objectives and outline

A better understanding of the epidemiology of the disease is crucial to support medical decisions and allocate health resources, as the number of hospital admissions can provide useful information on the level and quality of assistance provided in primary health care to individuals with diabetes [19].

Facing the burden of this disease from individual, familiar and social perspectives, this study focus on the temporal evolution of hospitalizations due to diabetes in Portugal, aiming to describe and model a time series of monthly hospitalizations for this cause, with an emphasis on prediction.

The thesis is structured in six chapters. The first and current one introduces the research theme, to bring into focus the clinical and the epidemiological relevance of the study. Its main objectives were also enunciated, with the following chapter describing the series and variables under study (Chapter 2). In Chapter 3, a theoretical framework for time series is given, from general concepts to modelling and forecasting approaches. It follows the implementation of these methods, with the results of the time series analysis being presented in Chapter 4 and further discussed in Chapter 5, in the light of scientific evidence about the theme and statistical methodologies addressed, exploring both strengths and limitations of the work conducted. Finally, Chapter 6

includes the main conclusions of the study and their relevance from an epidemiological point of view, along with perspectives of future research.

Some of the results of this work have been presented in poster format at the meeting *Statistics on Health Decision Making: clinical trials*, in October 2020, and published as an extended abstract in the Journal of Statistics on Health Decision [28].

Chapter 2

Study description

The data used in this thesis were obtained from Hospital Morbidity Databases, provided by the ACSS, and relate to the period from January 2010 to December 2018. Section 2.1 refers to the procedure followed for the construction of the time series, while other variables of interest are described in Section 2.2.

2.1 Study series

Records of hospital admissions occurred from 2010 to 2018 (provisional data for 2017 and 2018) were firstly joined, and further filtered by code and type. Admissions with main diagnose coded, up to the third digit, by 250 (diabetes mellitus), according to the International Classification of Diseases (ICD), 9th Revision, Clinical Modification (ICD-9-CM) or E10 (diabetes type 1), E11 (diabetes type 2), E13 (other type of diabetes), following the 10th revision of ICD (ICD-10-CM/PCS), were selected (Tables 2.1 and 2.2, respectively). In turn, admissions caused by secondary diabetes and gestational diabetes, coded apart both in ICD-9-CM and ICD-10-CM/PCS, were purposely not included.

Each of the selected diagnosis was associated to a specific episode. Therefore, the previously obtained list was merged with episode data through a unique sequential number, used as identifier in the national database. Given the available information, episodes were selected by date of admission and length of stay in the health facility, thus keeping those starting between 1 January 2010 and 31 December 2018 and lasting at least one day, henceforth referred to as *hospitalizations due to diabetes*.

Given the month and year of admission of the patient to the health facility, a time series of monthly hospitalizations due to diabetes in Portugal between January 2010 and December 2018 was constructed. This series, in a total of 108 observations, was analysed according to Box-Jenkins approach, as detailed in Chapter 3.

Table 2.1: ICD-9-CM codes and description for diabetes.

Code	Description
250	Diabetes mellitus
250.0	Diabetes mellitus without mention of complication
250.1	Diabetes with ketoacidosis
250.2	Diabetes with hyperosmolarity
250.3	Diabetes with other coma
250.4	Diabetes with renal manifestations
250.5	Diabetes with ophthalmic manifestations
250.6	Diabetes with neurological manifestations
250.7	Diabetes with peripheral circulatory disorders
250.8	Diabetes with other specified manifestations
250.9	Diabetes with unspecified complication

Table 2.2: ICD-10-CM/PCS codes and description for diabetes.

Code	Description
E10	Type 1 diabetes mellitus
E10.1	Type 1 diabetes mellitus with ketoacidosis
E10.2	Type 1 diabetes mellitus with kidney complications
E10.3	Type 1 diabetes mellitus with ophthalmic complications
E10.4	Type 1 diabetes mellitus with neurological complications
E10.5	Type 1 diabetes mellitus with circulatory complications
E10.6	Type 1 diabetes mellitus with other specified complications
E10.8	Type 1 diabetes mellitus with unspecified complications
E10.9	Type 1 diabetes mellitus without complications
E11	Type 2 diabetes mellitus
E11.0	Type 2 diabetes mellitus with hyperosmolarity
E11.1	Type 2 diabetes mellitus with ketoacidosis
E11.2	Type 2 diabetes mellitus with kidney complications
E11.3	Type 2 diabetes mellitus with ophthalmic complications
E11.4	Type 2 diabetes mellitus with neurological complications
E11.5	Type 2 diabetes mellitus with circulatory complications
E11.6	Type 2 diabetes mellitus with other specified complications
E11.8	Type 2 diabetes mellitus with unspecified complications
E11.9	Type 2 diabetes mellitus without complications
E13	Other specified diabetes mellitus
E13.0	Other specified diabetes mellitus with hyperosmolarity
E13.1	Other specified diabetes mellitus with ketoacidosis
E13.2	Other specified diabetes mellitus with kidney complications
E13.3	Other specified diabetes mellitus with ophthalmic complications
E13.4	Other specified diabetes mellitus with neurological complications
E13.5	Other specified diabetes mellitus with circulatory complications
E13.6	Other specified diabetes mellitus with other specified complications
E13.8	Other specified diabetes mellitus with unspecified complications
E13.9	Other specified diabetes mellitus without complications

For model identification and estimation, a subset of the series composed by data from January 2010 to December 2016 (84 values) was used, leaving the remaining 24 observations (22% of the data available) for model validation. To assess the predictive ability of the selected model, forecasts were computed for the years 2017 and 2018, and compared against the observed values of the series. The chosen model was further evaluated by comparison with a benchmark method with respect to forecast accuracy.

2.2 Variables description

For the purpose of this study, variables concerning episodes (i.e., hospitalizations in medical facilities) and diagnostic data were retrieved from Hospital Morbidity Databases and listed bellow.

Diagnostic variables:

cod_diagnostico: Diagnosis code;

tipo_p_s: Type of diagnosis. Categorical variable with possible outcomes:

- Main diagnosis (diagnosis considered responsible for the patient's admission)
- Additional diagnostics (any diagnosis assigned to a patient in a given care episode, in addition to the main diagnosis)

Episode variables:

sexo: Patient sex. Categorical variable:

- Male
- Female
- Undefined

idade: Patient age, when admitted, in years;

distrito: Patient district of residence (two digit code);

concelho: Patient county of residence (two digit code);

freguesia: Patient parish of residence (two digit code);

data_entrada: Date of admission of the patient, in *dd-mm-yyyy* format;

data_saida: Date of discharge of the patient, in *dd-mm-yyyy* format;

dias_int: Length of stay of the patient in the health facility, in days;

dsp: Patient destination after discharge from a hospital service. Categorical variable:

- Unknown
- Discharge home
- Another institution (with hospitalization)
- Home care
- Discharge against medical opinion
- Specialized aftercare (tertiary)
- Deceased
- Palliative care at medical centre
- Post-hospital care
- Long-term hospital care

adm_tip: Nature or mode of admission of a patient to a health institution. Categorical variable:

- Planned admission
- Emergency
- Private Medicine
- Access plan to Ophthalmologic Surgery

n_ficticio_utente: Fictitious patient number;

versao_icd: Coding version (ICD-9 or ICD-10).

For each episode of interest, sociodemographic — sex, age, region of residence (Nomenclature of Territorial Units for Statistics (NUTS) regions of level 2 (NUTS 2), known from the county of residence) —, and clinical variables — mode of admission, length of hospitalization and patient destination after discharge — are detailed, with descriptive statistics being presented.

For a better perspective on regional disparities, annual estimates of population by region of residence were obtained from the website of Statistics Portugal (Instituto Nacional de Estatística) and used to calculate the number of hospitalizations due to diabetes per 100,000 inhabitants by NUTS 2, from 2010 to 2018.

Chapter 3

Statistical background on time series

The present chapter introduces theoretical concepts and methods on time series analysis, further applied in this thesis to the series of monthly hospitalizations due to diabetes. In Section 3.1 are defined stochastic processes, whereas Sections 3.2 and 3.3 describe specific models for time series. Box-Jenkins methods for model building are presented in Section 3.4, and, finally, forecasting approaches are described in Section 3.5.

3.1 Stochastic processes and time series

A stochastic process is a collection of random variables indexed by *t*, $\{Y_t\}$, in a parameter set *T* [29]. Whenever this set corresponds to ordered moments of time, $\{Y_t\}$ is defined for each *t* and the observed values at different time points constitute a time series [30]. Thus, a series of *N* observations generated over time, $(y_1, \dots, y_N)$, is a sample realization (from an infinite population) of a stochastic process [29].

One example of a time series would be the sequence of pH measures taken every day, from hour to hour ($t = 1, \dots, 24$). In this case, Y_t is the random variable that represents pH at time *t* and the set of measurements of each day constitutes a sample realization of the stochastic process. If data of many days were available (in a week, it would be seven realizations of the process), it would be possible to obtain the probability distribution of the variables that comprise the process, assuming that the pool is in similar conditions.

For simplicity of notation, and also for closer proximity to Box *et al.* [29], y_t will henceforth be used to represent the random variable Y_t and its observed value y_t, i.e., the time series that is a realization of the stochastic process and its observed values.

The mean function of the stochastic process represents the expected value of the marginal distributions of the random variables for each moment *t*,

$$E[y_t] = \mu_t.$$

If all the variables have equal mean, the mean function is constant and the process is said to be stable in the mean [30]. Similarly, it would be stable in the variance if this moment is constant over time,

$$\text{Var}\,[y_t] = \sigma^2.$$

Still to note that a process can be stable in the mean, but not in the variance, with the opposite being also true. Regardless, for most situations with practical interest, it is not possible to observe multiple realizations of the process. Given that limitation, in order to estimate 'momentary' characteristics of the process, such as its mean, one must assume that the marginal distribution of the random variables at any instant *t* is stable over time. These stability proprieties are the base of the concept of stationarity [30].

Events taking stable values over time, not showing any trends, are called stationary. On the other hand, phenomena showing non constant values over time, by means of trend, seasonality or other effects, are called nonstationary. A given time series, depending on the period of observation, can be stationary or not [30].

Conceptually, stationary processes represent a particular class of stochastic processes, characterized by the assumption of *statistical equilibrium* [29]. If the proprieties of some stochastic process remain constant over time, that is, if the joint probability distribution of *m* observations $y_{t_1}, y_{t_2}, \dots, y_{t_m}$ is identical to that of $y_{t_1+k}, y_{t_2+k}, \dots, y_{t_m+k}$, for any set of indices $\{t_1, t_2, \dots, t_m\}$ and lag *k*, that process is defined as *strictly stationary* [29, 31]. This implies that 1) all the variables y_t have identical marginal distributions and 2) for any set of variables, the finite-dimensional distributions depend only on the lags *k* by which they are separated [30].

As these assumptions can not be easily fulfilled, a concept of weak stationarity is usually considered, leading to a stochastic process with less rigid conditions (restricted to the first two moments) — *weakly stationary* (or second-order stationary) process — defined by:

1. $E\,[y_t] = \mu$, for all times t;
2. $\text{Var}\,[y_t] = \sigma_y^2 < \infty$, for all times t;
3. $\text{Cov}\,[y_t, y_{t-k}] = \gamma_k$ depends only on lag k for all times t.

To note, however, that strict stationarity does not imply weak stationarity, as the first does not assume finite variance. In either case, for univariate time series, when $m = 1$, the assumption of stationarity presupposes that the probability distribution of y_t, $p(y_t)$, is equal for all times *t*, so that the notation can be simplified as $p(y)$ [29]. It follows that the stochastic process has constant mean,

$$\mu = \int_{-\infty}^{\infty} y \cdot p(y)\, dy, \tag{3.1}$$

that defines the level about which the process fluctuates. It can be estimated by the sample mean,

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t. \tag{3.2}$$

The stochastic process has also constant variance,

$$\sigma_y^2 = E\left[(y_t - \mu)^2\right] = \int_{-\infty}^{\infty} (y-\mu)^2 \cdot p(y)\, dy, \tag{3.3}$$

that measures the dispersion of values about the series level. Like mean, the variance can be estimated by the sample variance of the time series [29],

$$\hat{\sigma}_y^2 = \frac{1}{N}\sum_{t=1}^{N} (y_t - \bar{y})^2. \tag{3.4}$$

In time series, the values of the variables y_t tend to be correlated, as the observations occur consecutively over time. It is, in fact, a particular characteristic of these processes, with values at a given moment affecting values later observed. This kind of dependence between two variables at different points in time, y_t and y_{t+k}, is described by the autocovariance and the autocorrelation functions.

Under the assumption of stationarity, the covariance between y_t and some value *k* lags (i.e., *k* intervals of time) apart must be the same for all *t*, that is, covariance must depend on time differences only, not the time point. This function is called *autocovariance* at lag *k* and is defined by

$$\gamma_k = \mathrm{Cov}\left[y_t, y_{t+k}\right] = E\left[(y_t - \mu)(y_{t+k} - \mu)\right], \tag{3.5}$$

with $\gamma_k = \gamma_{-k}$ and $\gamma_0 = \mathrm{Var}\left[y_t\right] = \sigma^2$. From that follows the definition of *autocorrelation* at lag *k*:

$$\begin{aligned} \rho_k &= \frac{E\left[(y_t - \mu)(y_{t+k} - \mu)\right]}{\sqrt{E\left[(y_t - \mu)^2\right] E\left[(y_{t+k} - \mu)^2\right]}} \\ &= \frac{E\left[(y_t - \mu)(y_{t+k} - \mu)\right]}{\sigma_y^2} \\ &= \frac{\gamma_k}{\gamma_0}, \end{aligned} \tag{3.6}$$

with $\rho_k = \rho_{-k}$ and $\rho_0 = 1$. These functions are related to each other by the variance σ_y^2 of the process, given that $\gamma_k = \rho_k \cdot \sigma_y^2$ [29].

The plots of γ_k and ρ_k versus lag *k* represent the autocovariance and autocorrelation functions of the stochastic process, respectively [29]. The graphical projection of the autocorrelation function (ACF) for different lags

(correlogram) is particularly useful to see the dependence structure of the process [32].

In practice, both autocovariance and autocorrelation functions are unknown and replaced by sample estimates. The sample autocorrelation function, r_k, can be obtained by

$$r_k = \hat{\rho_k} = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}, \tag{3.7}$$

where

$$\hat{\gamma}_k = \frac{1}{N}\sum_{t=1}^{N-k}(y_t - \bar{y})(y_{t+k} - \bar{y}), \tag{3.8}$$

for $k = 0, 1, \dots, K$, is the autocovariance estimate of the series. For large N, the vector of sample autocorrelations approximates to normal distribution with mean $\boldsymbol{\rho}$, the theoretical vector of autocorrelations, and approximate variance given by Bartlett's formula,

$$\text{Var}\,[r_k] \simeq \frac{1}{N}\sum_{u=-\infty}^{\infty}(\rho_u^2 + \rho_{u+k}\rho_{u-k} - 4\rho_k\rho_u\rho_{u-k} + 2\rho_u^2\rho_k^2). \tag{3.9}$$

For any process with autocorrelations $\rho_v = 0$ for $v > q$, this expression can be simplified [29, 30]. In that case, at large lags, the variance of the estimated autocorrelations is approximated by

$$\text{Var}\,[r_k] \simeq \frac{1}{N}(1 + 2\sum_{v=1}^{q}\rho_v^2) \qquad k > q.$$

The purpose of time series analysis is, indeed, to explore the dependence structure between different time points, naturally correlated [33]. In its essence, a time series model attempts to explain the correlation present in the data, improving their ability to predict future observations [32]. By accounting for some or all of such dependence, it is expected that the residuals obtained will be uncorrelated, with constant mean and variance, as if they were white noise [33]. This concept refers to a sequence of random variables with mean zero and constant variance, $w_t \sim WN(0, \sigma_w^2)$, representing a weakly stationary process, where all w_t are uncorrelated with autocovariance function defined as

$$\gamma_k = E\,[w_t w_{t+k}] = \begin{cases} \sigma_w^2, & k = 0 \\ 0, & k \neq 0. \end{cases} \tag{3.10}$$

Also defined as white noise, a sequence of random variables that, more than uncorrelated, are independent and identically distributed (iid), $w_t \sim \text{iid}(0, \sigma_w^2)$, represents the most simple example of a strict stationary process. If the variables follow a normal distribution, with both non-correlation and independence assumptions being fulfilled, the resulting process is defined as a Normal ou Gaussian white noise [29, 31]. Either in a strict or

weak sense, a white noise process has no memory, which is the same as saying that past values of the series provide no information for predicting future values [30].

From the sum of white noise terms, $w_t \sim \text{iid}(0, \sigma_w^2)$, results a process known as *Random Walk*:

$$y_t = y_{t-1} + w_t \qquad t = 1, 2, \ldots, \tag{3.11}$$

where $y_{t-1} = \sum_{i=1}^{t-1} w_i$, having as start point $y_1 = w_1$ [34]. This time series is a particular case of the process $y_t = \delta + y_{t-1} + w_t$, $t = 1, 2, ...$, when the *drift*, δ, equals 0, representing a stochastic process in which the value of the series at time t depends only on the value of the series at time $t - 1$ plus a random shock (w_t) [29, 34]. From a practical point of view, w_t can be interpreted as steps, back or forward, taken by some person and whose sequence (sum) define his/her position at time t [31].

From Equation 3.11, it can easily be obtained the mean,

$$\begin{aligned} \mu_t &= E[y_t] = E[w_1 + w_2 + \cdots + w_t] \\ &= E[w_1] + E[w_2] + \cdots + E[w_t] \\ &= 0, \end{aligned} \tag{3.12}$$

and the variance of the process $\{y_t : t = 1, 2, \ldots\}$,

$$\begin{aligned} \text{Var}[y_t] &= \text{Var}[w_1 + w_2 + \cdots + w_t] \\ &= \text{Var}[w_1] + \text{Var}[w_2] + \cdots + \text{Var}[w_t] \\ &= \sigma_w^2 + \sigma_w^2 + \cdots + \sigma_w^2 \\ &= t\sigma_w^2. \end{aligned} \tag{3.13}$$

This is a clear example of a stochastic process stable in the mean ($\mu_t = 0$ for all t), but not in the variance, as it increases linearly with time ($t\sigma_w^2$; Figure 3.1) [31].

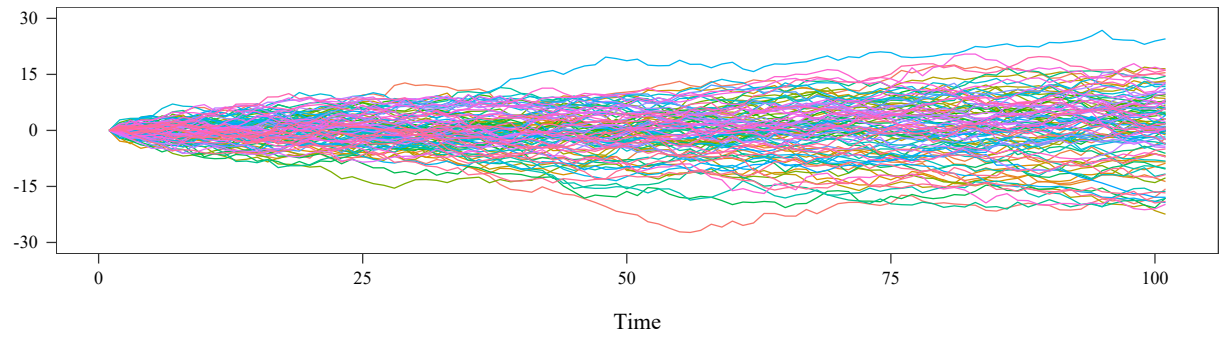

Figure 3.1: Simulation of one hundred steps of a random walk (100 realizations).

3.2 Models for stationary time series

Time series generated by a linear combination of independent, random, shocks can be represented by stationary models, of great importance in modelling a myriad of processes.

3.2.1 General linear process

A stochastic process represented as a weighted sum of actual and past values of white noise terms is a *general linear process*, defined by

$$\begin{aligned}\tilde{y}_t &= w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + ... \\ &= w_t + \sum_{i=1}^{\infty} \psi_i w_{t-i},\end{aligned} \tag{3.14}$$

with $\tilde{y}_t = y_t - \mu$ as the 'distance' of the process, if stationary, from its level $\mu = E[y_t]$. For this process, $E[y_t] = 0$ and $\gamma_k = \sigma_w^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$ [29].

Equation 3.14 can be rewritten in terms of the *backward shift operator, B*, $B^j y_t = y_{t-j}$, as

$$\begin{aligned}\tilde{y}_t &= w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + ... \\ &= \psi(B) w_t,\end{aligned} \tag{3.15}$$

where $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + ...$ is the *transfer function* of the linear filter that transforms the white noise process w_t into the process $\tilde{y}_t$ [29].

Concerning the right-hand side of Equation 3.14, $\tilde{y}_t$ represents a valid stationary process if the series ψ_i is either finite or infinite under the condition $\sum_{i=1}^{\infty} |\psi_i| < \infty$ [29]. This process can also be interpreted as a weighted linear combination of past values of the process $\{\tilde{y}_t\}$ plus a random white noise term w_t:

$$\begin{aligned}\tilde{y}_t &= \pi_1 \tilde{y}_{t-1} + \pi_2 \tilde{y}_{t-2} \cdots + w_t \\ &= \sum_{j=1}^{\infty} \pi_j \tilde{y}_{t-i} + w_t.\end{aligned} \tag{3.16}$$

To associate present observations with past series values in an interpretable manner, an invertibility condition must be fulfilled. The invertibility assumption for a general linear process requires the weights π_j to be entirely summable, that is $\sum_{j=0}^{\infty} |\pi_j| < \infty$, given $\pi(B) = \psi^{-1}(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$ [29].

In either forms (Equations 3.14 and 3.16), the practical application of the general linear process is limited due to its infinite number of parameters [29, 31].

3.2.2 Autoregressive process

Retaining from Equation 3.16 the first p parameters π of the expression, it is obtained the process

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \cdots + \phi_p \tilde{y}_{t-p} + w_t, \tag{3.17}$$

with the symbol ϕ now representing the set of weight parameters $(\phi_1, \ldots, \phi_p)$. Such process, denoted as *autoregressive* (AR) *process of order p*, AR(p) process, can also be written as

$$\phi(B)\tilde{y}_t = w_t, \tag{3.18}$$

where

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{3.19}$$

is the autoregressive operator of order p and $\phi(B) = 0$ is the characteristic equation. This AR process has $p + 2$ unknown parameters to be estimated, namely μ, $\phi_1, \phi_2, \ldots, \phi_p$, and σ_w^2 [29].

The employed name *autoregressive* derives from the fact that the actual value of the series, $\tilde{y}_t$, is the result of a weighted combination of its last p values plus a random value, w_t, in which is contemplated all the information at time t that could not be explained by previous values [31].

Since the series $\phi(B)$ is finite, an autoregressive process is invertible without any conditions over its parameters. Nevertheless, an AR process needs the autoregressive operator to have all its roots (the solutions of $\phi(B) = 0$) greater than one in absolute value (that is, outside the unit circle) to ensure stationarity [29]. For the general case, the following statements are necessary, but not sufficient, to satisfy this condition [31]:

$$\begin{cases} \phi_1 + \phi_2 + \cdots + \phi_p < 1 \\ |\phi_p| < 1. \end{cases} \tag{3.20}$$

When $p = 1$, the so called first-order AR process $\tilde{y}_t = \phi \tilde{y}_{t-1} + w_t$ just requires $|\phi| < 1$ to fulfil this requirement [29].

From multiplying Equation 3.17 by $\tilde{y}_{t-k}$, for $k \geqslant 0$, and taking expectations of the resulting values, follows that:

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p} \qquad k > 0, \tag{3.21}$$

which divided by γ_0 results in

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p} \qquad k > 0, \tag{3.22}$$

that is, the autocorrelation function of a stationary AR(p) process, generally represented by a combination of damped exponentials and damped sine waves [29, 31].

Taking $k = 1, 2, \ldots, p$ in Equation 3.22, with $\rho_0 = 1$, follows the sequence of equations known as *Yule-Walker equations*:

$$\begin{cases} \rho_1 = \phi_1 + \phi_2 \rho_1 + \cdots + \phi_p \rho_{p-1} \\ \rho_2 = \phi_1 \rho_1 + \phi_2 + \cdots + \phi_p \rho_{p-2} \\ \vdots \\ \rho_p = \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \cdots + \phi_p. \end{cases} \tag{3.23}$$

Estimates of the ϕ parameters can be obtained by replacing the theoretical autocorrelations ρ_k by the estimated autocorrelations r_k in the Yule-Walker equations [29].

The variance of an AR process, $\gamma_0 = \sigma_y^2$, can be expressed in terms of the parameters $\sigma_w^2, \phi_1, \phi_2, \ldots, \phi_p$ and autocorrelation values, as following [29, 31]:

$$\gamma_0 = \frac{\sigma_w^2}{1 - \phi_1 \rho_1 - \phi_2 \rho_2 - \cdots - \phi_p \rho_p}. \tag{3.24}$$

The order p of an AR process is generally unknown and needs to be defined from the sample data [29]. For that purpose, the autocorrelation function can not be said to provide useful information as it is infinitely extensive, that is, autocorrelations do not became zero after a specific number of lags, instead they tail off (Figure 3.2) [31]. The information provided by the autocorrelation function is, therefore, complemented by the partial autocorrelation function (PACF) at lag k, ϕ_{kk}, defined as

$$\phi_{kk} = \text{Corr}\left[y_t - \widehat{y}_t, y_{t-k} - \widehat{y}_{t-k}\right], \tag{3.25}$$

where $\widehat{y}_t = \phi_{k-1,1}\, y_{t-1} + \phi_{k-1,2}\, y_{t-2} + \cdots + \phi_{k-1,k-1}\, y_{t-k+1}$ and, equivalently, $\widehat{y}_{t-k} = \phi_{k-1,1}\, y_{t-k+1} + \phi_{k-1,2}\, y_{t-k+2} + \cdots + \phi_{k-1,k-1}\, y_{t-1}$ are the best linear predictors of y_t and y_{t-k}, respectively, based on the values $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$ [29]. Hence, the partial autocorrelation function quantifies the correlation between the residuals from these regressions or, simply, the correlation between y_t and y_{t-k} not accounting for the effect of the intermediary values $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$ [29, 31].

For an AR(p) process, the partial autocorrelation function cuts off after lag p, meaning that ϕ_{kk} are nonzero for all $k \leqslant p$ and zero for $k > p$ (Figure 3.2) [29]. Estimates for these values, $\widehat{\phi}_{kk}$, can be obtained by recursive methods, as the one proposed by Levinson (1947) and Durbin (1960) for either theoretical or sample partial

autocorrelations,

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_j}, \tag{3.26}$$

where $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}$ for $j = 1, 2, \ldots, k-1$ [31].

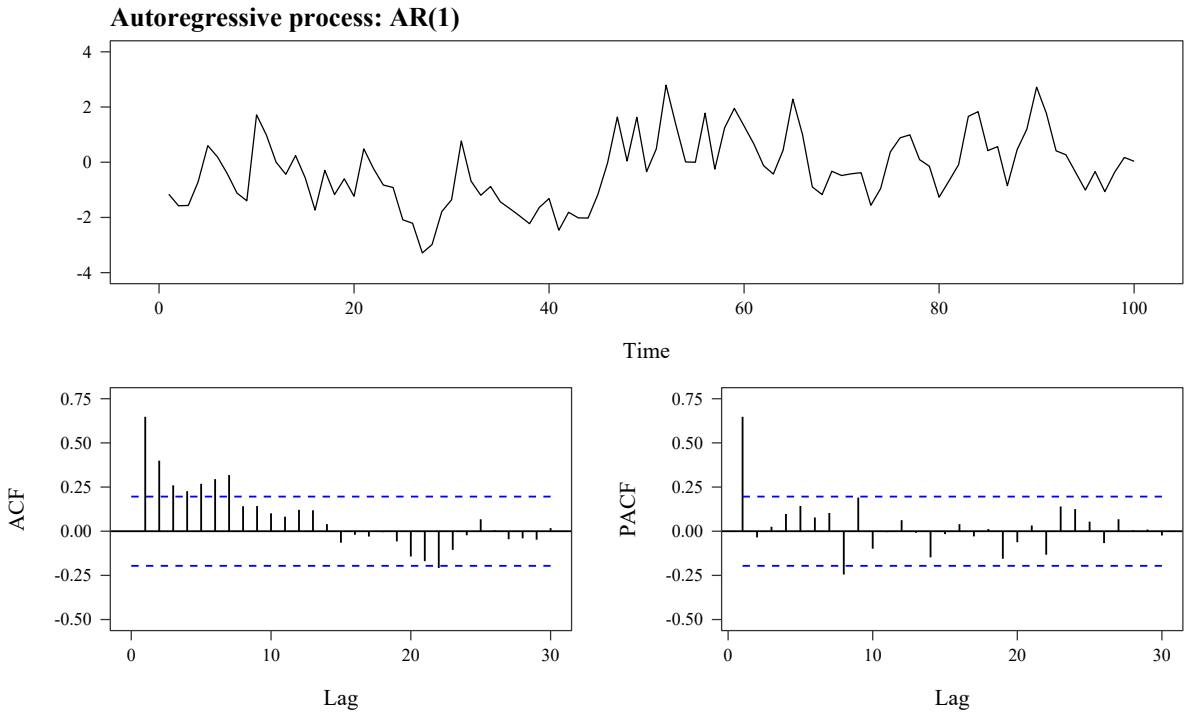

Figure 3.2: Simulation of an AR(1) process with parameter $\phi = 0.5$. Sequence plot (at top) and correlogram for the ACF and the PACF (at bottom). ACF, Autocorrelation Function; PACF, Partial Autocorrelation Function.

3.2.3 Moving Average process

Considering as nonzero only the first q weights ψ of the expression 3.14, the resulting process,

$$\tilde{y}_t = w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2} - \cdots - \theta_q w_{t-q}, \tag{3.27}$$

is referred to as *moving average* (MA) *process of order* q, or MA(q) process, that can be written, alternatively, as

$$\tilde{y}_t = \theta(B) w_t, \tag{3.28}$$

where

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \tag{3.29}$$

is the moving average operator of order q [29].

The weights $1, -\theta_1, -\theta_2, \ldots, \theta_q$ are applied to the white noise variables $w_t, w_{t-1}, w_{t-2}, \ldots, w_{t-q}$ to obtain y_t and then 'moved' over $w_{t+1}, w_t, w_{t-1}, w_{t-2}, \ldots, w_{t-q+1}$ to get y_{t+1}, and so on, justifying the name *moving average* [31].

The variance and the autocovariance function of the process $\tilde{y}_t$ are, respectively,

$$\sigma_y^2 = \gamma_0 = \sigma_w^2(1 + \theta_1^2 + \theta_2^2, \ldots, \sigma_q^2) \tag{3.30}$$

and

$$\gamma_k = \begin{cases} \sigma_w^2(-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q), & k = 1, 2, \ldots, q \\ 0, & k > q. \end{cases} \tag{3.31}$$

From that follows the autocorrelation function of the process, recognized by its cutoff after q lags,

$$\rho_k = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}, & k = 1, 2, \ldots, q \\ 0, & k > q. \end{cases} \tag{3.32}$$

Rephrasing, the autocorrelations of a moving average process of order q take the value zero beyond lag q, in a similar way as the partial autocorrelation function of an AR(p) process cuts off after lag p. Conversely, the partial autocorrelations of a MA(q) process show a behaviour similar to the autocorrelations of an AR(p) process (Figure 3.3). Estimates for $\theta_1, \theta_2, \ldots, \theta_q$ can be obtained by replacing the theoretical autocorrelations for their estimates, r_k, in Equation 3.32. The level of the process MA(q), μ, and the variance of the white noise process, σ_w^2, are also unknown, making a total of $q + 2$ parameters to be estimated from the sample data [29].

Moving average processes are subject to invertibility conditions, independent of stationarity requirements. To be invertible, the roots of $\theta(B)$ $(\theta(B) = 0)$ must all lie outside the unit circle, that is, they need to be greater than one in absolute value. No further conditions are required to ensure stationarity, since $\theta(B)$ is a finite series [29].

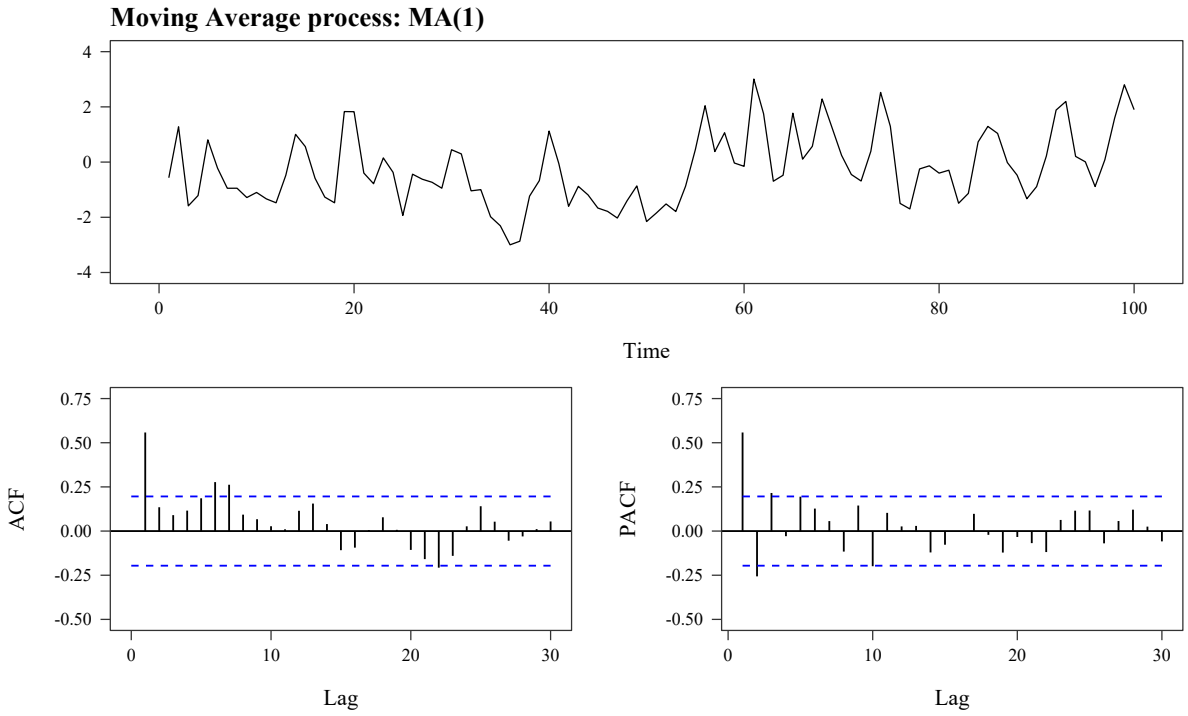

Figure 3.3: Simulation of a MA(1) process with parameter $\theta = 0.7$. Sequence plot (at top) and correlogram for the ACF and the PACF (at bottom). ACF, Autocorrelation Function; PACF, Partial Autocorrelation Function.

3.2.4 Mixed Autoregressive Moving Average process

In practice, some time series are better explained by both moving average and autoregressive terms. Such processes are represented by mixed *Autoregressive Moving Average* (ARMA) models with parameters *p* and *q*, ARMA(*p,q*):

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \cdots + \phi_p \tilde{y}_{t-p} + w_t - \theta_1 w_{t-1} - \cdots - \theta_q w_{t-q} \tag{3.33}$$

or, recurring to moving average and autoregressive operators,

$$\phi(B)\tilde{y}_t = \theta(B) w_t, \tag{3.34}$$

where $\phi(B)$ and $\theta(B)$ are the polynomial operators in *B* of orders *p* and *q* defined in Equations 3.19 and 3.29, respectively. Thus, an ARMA process has $p+q+2$ unknown parameters $(\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma_w^2)$ [29].

As $\tilde{y}_t = y_t - \mu$, the general ARMA process can be expressed in terms of the original series y_t as $\phi(B) y_t = \theta_0 + \theta(B) w_t$, where the constant θ_0 is a function of μ:

$$\theta_0 = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p). \tag{3.35}$$

Given the proprieties of the processes it integrates, the mixed process ARMA(*p,q*) takes as condition for stationarity that all the roots of $\phi(B) = 0$ lie outside the unit circle. The same requirement applies to the roots of $\theta(B) = 0$ so that $\phi(B)\tilde{y}_t = \theta(B)w_t$ defines an invertible process [29].

The autocovariance function of an ARMA(*p,q*) process is

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p} \qquad k > q, \tag{3.36}$$

from what follows its autocorrelation function

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p} \qquad k > q. \tag{3.37}$$

The 'look' of the ACF of a mixed process largely depends on the orders *p* and *q*: if $q - p < 0$, all the function will show a pattern of mixed damped exponentials and/or damped sine waves; if, instead, $q - p \geqslant 0$, the first $q - p + 1$ values $(\rho_0, \rho_1, \ldots, \rho_{q-p})$ will not follow such pattern. In turn, the partial autocorrelation function of an ARMA(*p,q*) process shows a pattern similar to the partial autocorrelation function of a MA(*q*) process, being characterized, once again, by damped exponentials and/or damped sine waves (Figure 3.4) [29].

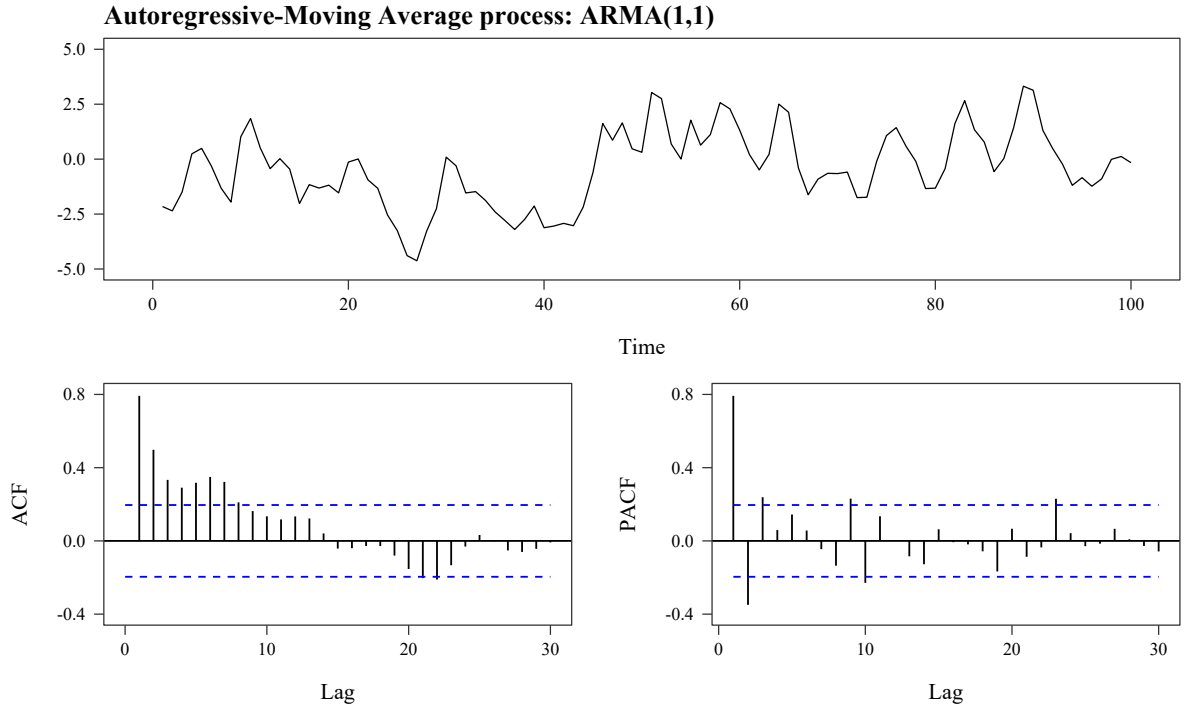

Figure 3.4: Simulation of an ARMA(1,1) process with parameters $\phi = 0.5$ and $\theta = 0.5$. Sequence plot (at top) and correlogram for the ACF and the PACF (at bottom). ACF, Autocorrelation Function; PACF, Partial Autocorrelation Function.

Table 3.1: Summary of properties of Autoregressive, Moving Average and Mixed processes.

	Autoregressive process	**Moving Average process**	**Mixed process**
Model in terms of previous $\tilde{y}'s$	$\phi(B)\tilde{y}_t = w_t$	$\theta^{-1}(B)\tilde{y}_t = w_t$	$\theta^{-1}(B)\phi(B)\tilde{y}_t = w_t$
Model in terms of previous $w's$	$\tilde{y}_t = \phi^{-1}(B)w_t$	$\tilde{y}_t = \theta(B)w_t$	$\tilde{y}_t = \phi^{-1}(B)\theta(B)w_t$
ψ weights	Infinite series	Finite series	Infinite series
π weights	Finite series	Infinite series	Infinite series
Stationarity condition	Roots of $\phi(B) = 0$ outside the unit circle	Always stationary	Roots of $\phi(B) = 0$ outside the unit circle
Invertibility condition	Always invertible	Roots of $\theta(B) = 0$ outside the unit circle	Roots of $\theta(B) = 0$ outside the unit circle
Autocorrelation function	Infinite (damped exponentials and/or damped sine waves)	Finite	Infinite (damped exponentials and/or damped sine waves after first $q - p$ lags)
	Tails off	Cuts off after lag q	Tails off
Partial autocorrelation function	Finite	Infinite (damped exponentials and/or damped sine waves)	Infinite (dominated by damped exponentials and/or damped sine waves after first $p - q$ lags)
	Cuts off after lag p	Tails off	Tails off

Source: Box, Jenkins and Reinsel [29].

3.3 Models for nonstationary time series

Many time series, in areas like economy and health, exhibit a nonstationary behaviour, presenting trend, seasonality, or both. To accommodate/describe such behaviour, a new class of models is considered, by assuming that some difference of process is, indeed, stationary. This class of models can be further extended to include seasonal terms.

3.3.1 Autoregressive Integrated Moving Average Process

As explained in the previous section, a mixed ARMA(*p,q*) model (Equation 3.34) is stationary if all the roots of $\phi(B) = 0$ are greater than one in absolute value. Otherwise, the process is nonstationary. If the roots take some value lower than one, it can be shown that the series has exponential growth, clearly incompatible with a stationarity assumption. The remaining case, when at least one of the roots of $\phi(B) = 0$ lies on the unit circle, seems to more closely describe the behaviour of nonstationary time series [29].

Given the general model

$$\phi(B)(1-B)^d\tilde{y}_t = \theta(B)w_t \tag{3.38}$$

or

$$\varphi(B)\tilde{y}_t = \theta(B)w_t,$$

where

- $\varphi(B)$ is a nonstationary autoregressive operator, with *d* roots of $\varphi(B) = 0$ on the unit circle (*d* unit roots);
- $\phi(B)$ is a stationary autoregressive operator, with all the roots of $\phi(B) = 0$ outside the unit circle;
- $\theta(B)$ is an invertible moving average operator, with all the roots of $\theta(B) = 0$ greater than one in absolute value [29],

becomes evident that such process is stationary if $d = 0$. Alternatively, making use of the *differencing operator* $\nabla = 1 - B$, and given that $\nabla^d\tilde{y}_t = \nabla^d y_t$, the previous equation can be written as

$$\phi(B)\nabla^d y_t = \theta(B)w_t, \tag{3.39}$$

denoting an Autoregressive Integrated Moving Average (ARIMA) process of order (*p,d,q*) obtained by integration (i.e. sum) of a stationary ARMA(*p,q*) process *d* times,

$$\phi(B)z_t = \theta(B)w_t, \tag{3.40}$$

where $z_t = \nabla^d y_t$, that is, the *d*-th order difference of the series y_t. Hence, if z_t is represented by a stationary ARMA(*p,q*) process, then y_t is said to follow an ARIMA(*p,d,q*) model (Figure 3.5) [29, 31].

In some contexts, when a deterministic component exists, it may be useful to extend the model (3.39) to the form

$$\phi(B)\nabla^d y_t = \theta_0 + \theta(B)w_t, \tag{3.41}$$

where θ_0 is a constant term representing a nonzero mean in the sense that

$$E\left[z_t\right] = E\left[\nabla^d z_t\right] = \mu_z = \frac{\theta_0}{1-\phi_1-\phi_2-\cdots-\phi_p}.$$

The inclusion of this constant gives the model the capacity to represent series with deterministic trends, as a function of time. Still, the assumption of a stochastic trend seems more adequate to most series, and, for that

reason, the mean is assumed to be zero unless clear evidence in contrary [29].

This general model can be expressed, and therefore interpreted, in terms of:

- actual (w_t) and previous shocks ($w_{t-j},\ j = 1, 2, ...$),
- previous values of the process ($y_{t-j},\ j = 1, 2, ...$) and actual shock (w_t), or, more conveniently,
- previous values of the process ($y_{t-j},\ j = 1, 2, ...$) and actual (w_t) and previous shocks ($w_{t-j},\ j = 1, 2, ...$).

For this last case, the difference equation form of the model, with $\theta_0 = 0$, is used:

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_{p+d} y_{t-p-d} + w_t - \theta_1 w_{t-1} - \cdots - \theta_q w_{t-q}. \tag{3.42}$$

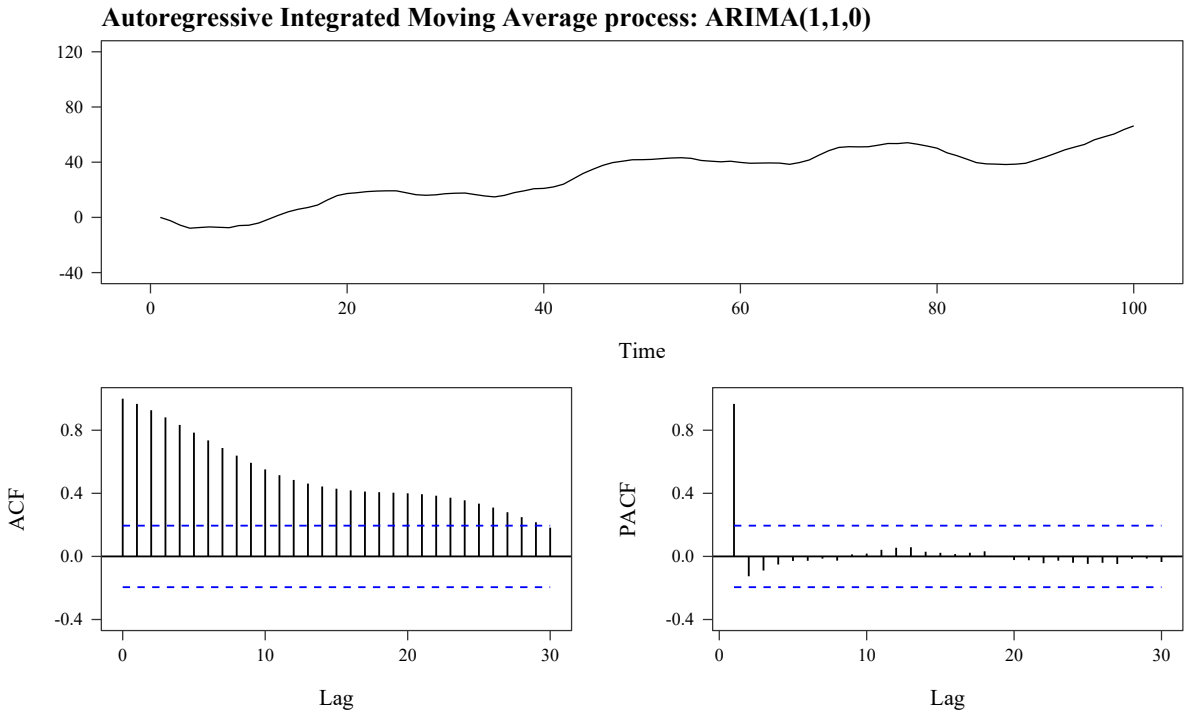

Figure 3.5: Simulation of an ARIMA(1,1,0) process with parameter $\phi = 0.9$. Sequence plot (at top) and correlogram for the ACF and the PACF (at bottom). ACF, Autocorrelation Function; PACF, Partial Autocorrelation Function.

3.3.2 Seasonal Autoregressive Integrated Moving Average process

Many time series present periodic fluctuations around the mean, as could be exemplified by the increasing sales of ice cream in the summer months. This effect is known as seasonality and must be explicitly incorporated into the model [33].

Seasonal time series are commonly analysed through its decomposition in *trend*, *seasonal* and *random* components [35–38], with exponential smoothing and seasonal *loess* (locally weighted scatterplot smoothing) as common methods, but some concerns can arise from that approach. While trend and seasonality from data can be properly fitted by a polynomial and a Fourier series, respectively, such methods can reveal some inflexibility when the objective is to predict future values of the series [29]. To deal with such components and the resulting departure of the series from the concept of stationarity, a seasonal process is considered.

Given $B^s y_t = y_{t-s}$ and $\nabla_s y_t = (1 - B^s) y_t = y_t - y_{t-s}$, where the nonstationary operator $1 - B^s$ has s unit roots for $e^{i(2\pi k/s)}$ ($k = 0, 1, \dots, s-1$) [29], a multiplicative seasonal ARIMA (SARIMA) process can be defined as

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \theta_q(B)\Theta_Q(B^s)w_t, \tag{3.43}$$

with nonseasonal orders *p, d, q*, seasonal orders *P, D, Q* and seasonal period *s*, where $\Phi(B^s)$ and $\Theta(B^s)$ are the seasonal AR and MA polynomials in B^s of degrees P and Q, respectively, satisfying stationarity and invertibility conditions [29, 31]. Hence, a series y_t is said to follow a seasonal ARIMA$(p, d, q) \times (P, D, Q)_s$ process if $z_t = \nabla^d\nabla_s^D y_t$ is represented by a stationary ARMA$(p, q) \times (P, Q)_s$ process, with seasonal period *s* (Figure 3.6).

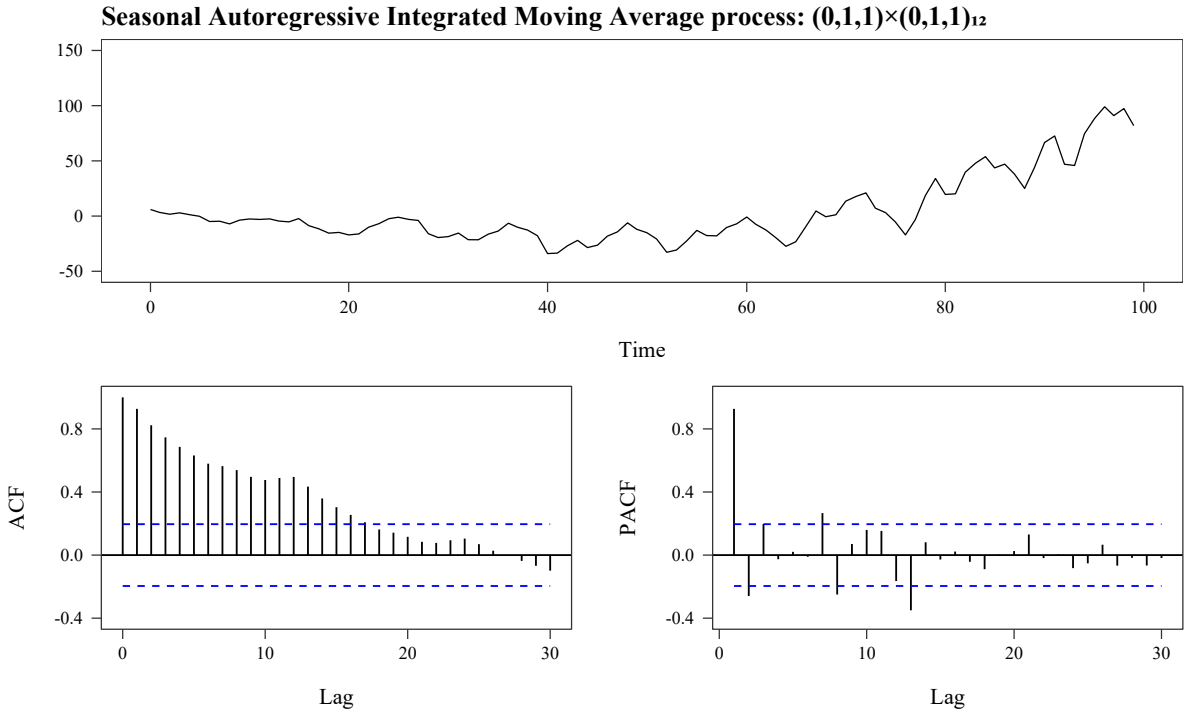

Figure 3.6: Simulation of a SARIMA(0,1,1)×(0,1,1)$_{12}$ process with parameters $\theta = 0.4$ and $\Theta = 0.6$. Sequence plot (at top) and correlogram for the ACF and the PACF (at bottom). ACF, Autocorrelation Function; PACF, Partial Autocorrelation Function.

3.4 Time series modelling

Box and Jenkins defined a three stage iterative method for model building — identification, estimation and diagnostic. This approach is one of the most popularized and will be further presented.

3.4.1 Model identification

Following the Box-Jenkins approach, time series analysis must start with the identification of an appropriate class of models to represent the process under study, given the general ARIMA family $\phi(B)\nabla^d y_t = \theta_0 + \theta(B)w_t$ [29].

As stated before, stationarity is an important feature of time series and to assess the validity of such assumption will be the first step in the analysis of a specific time series. Moreover, the presence of significant seasonality, that needs to be accounted for, must be investigated. Graphical methods are useful and commonly used tools in the preliminary identification of possible models to be fitted and checked later. Particular emphasis is given to sample autocorrelation and partial autocorrelation functions to assess the stationarity of y_t. If the estimated autocorrelation function, that follows the behaviour of the theoretical function, does not fall off quickly it could be taken as a signal that there may exist a root close to one. It suggests that the process under study should be treated as nonstationary in y_t, but eventually as stationary in $\nabla^d y_t$ with $d \geqslant 1$, reducing the process to a mixed ARMA model

$$\phi(B)z_t = \theta_0 + \theta(B)w_t,$$

where

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

For series with non-constant variance, possible transformations can be tested to stabilize it [33]. Power transformations, introduced by Box and Cox (1964) [31], stand out as one of the most popular, taking the form

$$g(y_t) = \begin{cases} \frac{y_t^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y_t), & \lambda = 0, \end{cases} \tag{3.44}$$

where λ could be considered an additional parameter to be estimated from the data. However, instead of a point estimation, a log-likelihood value is calculated for a range of selected values of λ, and the appropriate transformation chosen according to the parameter value that results in the maximum likelihood value [29].

This approach benefits from its simplicity, as the result can be easily interpreted, although it can not be applied to negative data. In that case, a preliminary step, where a positive constant is added to all of the observed values, must be taken before the transformation. An alternative would be to use another family of transformations [39].

Both transformation and differencing are intended to produce a series with constant scale and location [33]. Thus, the objective at this point is to identify the degree of differencing d, assuming that stationarity has been achieved when the estimated autocorrelations of y_t die out quickly. The inspection of the first 20 or so estimated autocorrelations of the original series ($d = 0$) and its first two differences, that is, $d = 1$ and $d = 2$, is recommended and usually enough, as most series will need no more than two differences to achieve stationarity. Further differencing, beyond the strictly necessary to achieve stationarity, will not improve the model, quite the opposite, it could introduce extra correlation into the series [29]. In addition, it could violate the assumption of invertibility and make it difficult to estimate model parameters [31].

While the initial evaluation of process stationarity is often made informally, based on characteristics of the series and its sample autocorrelation function, the decision on the need for differencing can be formally evaluated by testing for a unit root in the autoregressive operator of the model [31].

Considering the model $y_t = \phi y_{t-1} + w_t, t = 1, 2, ...,$ with $y_0 = 0$, the process $\{y_t\}$ is stationary if $|\phi| < 1$, but nonstationary when $\phi = 1$. To test the hypothesis of a unit root, one can use the Dickey-Fuller test, based on the the statistic

$$\hat{\tau} = \frac{\hat{\phi} - 1}{s_w \Big(\sum_{t=2}^{n} y_{t-1}^2 \Big)^{-1/2}},$$

where $\hat{\phi}$ is the conditional least-squares estimate of the parameter ϕ and $s_w^2 = (n-2)^{-1}(\sum_{t=2}^{n} y_t^2 - \hat{\phi} \sum_{t=2}^{n} y_{t-1} y_t)$ is the residual mean square. The alternative hypothesis is that the AR characteristic polynomial has no unit roots and so the process is stationary. As this is a one-sided test, the null hypothesis of $\phi = 1$ is rejected for small values of the test statistic, $\hat{\tau}$. These results can be extended for higher order models, AR($p + 1$), by using the augmented Dickey-Fuller (ADF) test. It is also valid for mixed ARIMA($p, 1, q$) or higher order differencing models. In that case, the model is approximated by an autoregressive model, whose order get to be estimated before the ADF test is applied. Other tests have also been considered for ARIMA models, such as Phillip and Perron tests or tests of the type of likelihood ratio. Notwithstanding, to a greater or lesser extent, unit root tests raise concerns regarding power for time series of short length [29].

Once defined the differencing order, the following step is to identify the AR and MA orders of the model for z_t. Suitable choices are identified based on the proprieties of the theoretical autocorrelation and partial autocorrelation functions previously described for moving average, autoregressive and mixed autoregressive moving average processes (Table 3.1). Given a MA(q) process, the autocorrelation function cuts off after lag

q, that is, ρ_k takes the value zero for lags greater than q, whereas the partial autocorrelation function tails off. Based on that, the identification of the order q of the model becomes possible through the inspection of the estimated autocorrelation function, whose behaviour tends to be identical to the theoretical function. Correspondingly, the order of an AR(p) process could be hypothesized considering the behaviour of the estimated partial autocorrelation function, as it cuts off after lag p ($\phi_{kk} = 0,\ k > p$), while the autocorrelation function tails off. For mixed ARMA(p,q) processes, both functions tail off, reassembling damped exponentials and/or damped sine waves [29].

One aspect to be noted about the estimated autocorrelations is that large covariance can exist between near values, in such a way that the estimated function does not perfectly match the behaviour of the theoretical function. Some large estimated autocorrelations can, thus, occur after lags q or p, justifying the need to further investigate some related models beyond the one suggested by the analysis of the ACF and the PACF [29]. Any indication on whether the autocorrelation and partial autocorrelations are effectively zero beyond specific lags q or p, respectively, is important.

For an hypothetical moving average process, with nonzero ρ up to lag q, the standard errors of estimated autocorrelations can be obtained by replacing theoretical autocorrelations for their estimates in Bartlett's approximation formula,

$$\hat{\sigma}\left[r_k\right] \simeq \sqrt{\frac{1}{n}\left(1 + 2\sum_{v=1}^{q} r_v^2\right)} \qquad k > q, \tag{3.45}$$

referred to as *large-lag standard error*, as it applies to lags greater than q [29]. For particular cases where $q = 0$, and thus $\rho = 0$ for all lags but lag 0, the series is perfectly random (white noise) and the standard errors for the estimated autocorrelations are simply

$$\hat{\sigma}\left[r_k\right] \simeq \frac{1}{\sqrt{n}} \qquad k > 0. \tag{3.46}$$

For time series of moderate size and theoretical autocorrelations equal to zero, the distribution of the correspondent estimated autocorrelations approximates to normal distribution, so that the statistic $r_k \,/\, \hat{\sigma}\left[r_k\right]$, to test $\rho = 0$, will approximate to a standard normal distribution [29]. It remains valid for partial autocorrelations, given the hypothesis of an AR(p) process. In that case, each estimated partial autocorrelation is divided by its standard error, defined as

$$\hat{\sigma}\left[\hat{\phi}_{kk}\right] \simeq \frac{1}{\sqrt{n}} \qquad k > p, \tag{3.47}$$

as shown by Quenouille (1949) [29].

Hence, based on the assumption that the process under analysis is white noise, it is possible to obtain limits

for autocorrelations and partial autocorrelations, usually plotted by dashed lines against their estimates as a visual clue to check whether the functions effectively cut off after lag *q* or *p*, respectively [29].

Other approaches may be considered for identification purposes, to complement the sample autocorrelation and partial autocorrelation functions which, in case of mixed processes, can give non-clear insights [29]. One of the most popular are the model selection criteria as the Akaike´s Information Criterion (AIC) proposed by Akaike (1974) and further normalized by the sample size *n*,

$$\begin{aligned} \text{AIC} &= \frac{-2\ln(\hat{L}) + 2r}{n} \\ &\approx \ln(\hat{\sigma}_w^2) + r\,\frac{2}{n} + \text{constant}, \end{aligned} \tag{3.48}$$

and the related Schwarz's Bayesian Information Criterion (BIC),

$$\begin{aligned} \text{BIC} &= \text{AIC} + r\Big(\ln(n) - 2\Big) \\ &= \frac{-2\ln(\hat{L}) + r\ln(n)}{n} \\ &\approx \ln(\hat{\sigma}_w^2) + r\,\frac{\ln(n)}{n}, \end{aligned} \tag{3.49}$$

where $\hat{L}$ is the maximized value of the likelihood function, $\hat{\sigma}_w^2$ is the maximum likelihood estimate of the error variance, σ_w^2, and $r = p + q + P + Q + c + 1$ is the number of estimated parameters, with $c = 1$ if the model includes a constant term, $c = 0$ otherwise [29]. When comparing several models, the one with the minimum value of AIC or BIC should be preferred. As the likelihood value naturally increases if a larger number of parameters is considered (more information imputed), the second term in Equations 3.48 and 3.49 intends to penalize the inclusion of additional parameters, leading to the choice of the most parsimonious model, that is, the best model with the minimal complexity. Such penalization is greater in BIC, so that the model chosen according to this criterion will have, at maximum, the same number of parameters as the model that would be chosen if the AIC was used [29].

Hurvich and Tsai (1989) proposed a new criterion, intended to eliminate the bias associated to AIC and defined, accordingly, as corrected AIC (AIC_c):

$$\text{AIC}_\text{c} = \left(n\text{AIC} + \frac{2r^2 + 2r}{n - r - 1}\right) / n, \tag{3.50}$$

where *r* represents the number of parameters to be estimated, and *n* is the effective sample size (for an ARIMA model it would be $n = N - d$). The authors have shown a better performance of AICc, compared to other criteria (e.g., AIC, BIC), when the value of *r* exceeds 10% of *n* ($r/n > 0.1$) [31].

Both AIC and BIC require models in analysis to be estimated by maximum likelihood, meaning that, for

ARMA models, multiple combinations of p and q need to be maximized, which may result in overfitting. This problem was addressed by Hannan and Rissanen (1982), that proposed a two-step approach to model selection, in which an AR model of high order, selected by using the AIC criterion, is firstly fitted, with the residuals obtained being used as estimates for the unobserved errors. The time series is then regressed by ordinary least squares on previous observed values and lagged residuals (from the model adjusted in first place) for multiple combinations of p and q. It follows that the model to be selected is the one with the lower value of BIC [29, 31].

3.4.2 Parameter estimation

Once an ARIMA model has been specified for the time series, parameter estimation follows. Since the time series is stationary after taking its d-th difference, the estimation of an ARIMA(p,d,q) process resumes to the estimation of a stationary ARMA(p,q) for the differenced time series. The same applies to seasonal models, an extension of ARIMA models [31]. The most common approaches will be addressed, but others can be applied, such as Bayes' theorem and bootstrapping [29, 31].

Method of moments

This is a simple method for parameter estimation, that consists in equating sample moments to their respective theoretical moments and solve the equation to obtain estimates for unknown parameters [31]. For an AR(p) process, estimates for $\phi_1, \phi_2, \dots, \phi_p$ can be obtained by solving the Yule-Walker equations (Equation 3.23) where the theoretical autocorrelations, ρ_k, are replaced by their estimates, r_k. For the error variance, estimates may be obtained from Equation 3.24, as $\sigma_w^2 = \gamma_0(1 - \phi_1\rho_1 - \phi_2\rho_2 - \cdots - \phi_p\rho_p)$, where the variance of the process γ_0 is replaced by its estimate. The same procedure can be followed for a MA(q) process, based on Equation 3.30, whereas Equation 3.32 could be used to obtain estimates for $\theta_1, \theta_2, \dots, \theta_q$. By replacing ρ_k by r_k for $k = 1, 2, \dots, q$, q nonlinear equations are obtained, each with multiple solutions, but only one is invertible. Thus, for models including a moving average term the method of moments is neither convenient nor efficient, producing poor estimates [29].

Maximum likelihood

The method of maximum likelihood has the use of all the information in the data as its most important feature. However, it requires working with the process joint probability density function [31].

For independent and identically distributed data, the joint probability density function, determined by the product of the marginal density function for each outcome observed, defines the probability of obtaining the observed data $\mathbf{y}$, given a specific set of parameters $\boldsymbol{\xi}$, fixed [29, 40]:

$$f(\mathbf{y};\boldsymbol{\xi}) = f(y_1,\dots,y_N;\boldsymbol{\xi}) = \prod_{t=1}^{N} f(y_t;\boldsymbol{\xi}).$$

Once the data were available, the question becomes what value of $\boldsymbol{\xi}$ could have originated the observations $\mathbf{y}$ actually obtained. The likelihood function addresses this question by taking the same form as $f(\mathbf{y};\boldsymbol{\xi})$, but assuming, instead, that $\mathbf{y}$ is fixed and $\boldsymbol{\xi}$ variable [29]. For an ARIMA model, it would be defined as a function of the $q+p+1$ unknown parameters, $\boldsymbol{\xi} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_w^2)$, given the observed data, $\mathbf{y}$:

$$L(\boldsymbol{\xi}|\mathbf{y}) = L(\boldsymbol{\xi}|y_1,\dots,y_N) = \prod_{t=1}^{N} f(y_t;\boldsymbol{\xi}).$$

Equivalently, the log-likelihood function, more convenient and often preferred in face of its additive proprieties, takes the form

$$\ln L(\boldsymbol{\xi}|\mathbf{y}) = \sum_{t=1}^{N} \ln f(y_t;\boldsymbol{\xi}).$$

The values of the parameters that most likely have originated the observations actually taken, that is, the values that maximize the likelihood and, consequently, the log-likelihood functions, are called *maximum likelihood estimates* (MLE) [29]. The properties of MLE can be extended to stationary processes, allowing the application of maximum likelihood to time series data, whose random variables y_t are not iid. Thus, the series $\mathbf{z}$ of length $n = N - d$ is used for estimation purposes. As previously noted, an ARIMA(*p,d,q*) model based on the original data y_t is equivalent to a stationary ARMA(*p,q*) model fitted to the differenced time series $z_t = \nabla^d y_t$ [29].

Conditional likelihood

Given $z_1,\dots,z_n$ observations from an ARMA(*p,q*), the log-likelihood for the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_w^2)$, conditional on the choice of the starting *p* values of y_t ($\mathbf{y}_*$) and *q* values of w_t ($\mathbf{w}_*$), prior to $t = 1$, would be defined as

$$\ell_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_w^2) = -\frac{n}{2}\ln(\sigma_w^2) - \frac{S_*(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_w^2}, \tag{3.51}$$

where

$$S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} w_t^2(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{z}_*, \mathbf{w}_*, \mathbf{z}) \tag{3.52}$$

is the *conditional sum-of-squares function*, with $w_t = z_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p} + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$. If the assumption $\mu = 0$, usually valid when $d > 0$, does not seem appropriate, z_t must be replaced by

$\tilde{z}_t = z_t - \mu$. In that case, μ may be estimated from the series mean, $\bar{z} = \sum_{t=1}^{n} z_t/n$, or, alternatively, specially for small sample sizes, included in $\boldsymbol{\xi}$ as a parameter to be estimated [29].

Given a fixed value of σ_w^2, ℓ_* is linear in $S_*(\boldsymbol{\phi}, \boldsymbol{\theta})$. Thus, maximizing the log-likelihood function very closely depends on minimizing the conditional sum-of-squares function, with the values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ obtained being denominated *conditional least-squares estimates* [29].

Unconditional likelihood

For an ARMA model, the exact or unconditional log-likelihood function is defined as

$$\ell(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_w^2) = f(\boldsymbol{\phi}, \boldsymbol{\theta}) - \frac{n}{2} \ln(\sigma_w^2) - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_w^2}, \tag{3.53}$$

where the *unconditional sum-of-squares function*,

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-\infty}^{n} [w_t | \mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\theta}]^2, \tag{3.54}$$

is based on the expectation of w_t conditional on the values of $\mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\theta}$. As $S(\boldsymbol{\phi}, \boldsymbol{\theta})$, $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ is independent of σ_w^2, but their contribution to $\ell(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_w^2)$ differs depending on the sample size. For series with small *n*, the log-likelihood is mostly determined by $f(\boldsymbol{\phi}, \boldsymbol{\theta})$. With increasing *n*, contours of the (log-)likelihood function are mainly defined by the unconditional sum-of-squares function, in such a way that MLE approximate to the estimates for $(\boldsymbol{\phi}, \boldsymbol{\theta})$ obtained by minimizing the *unconditional sum-of-squares function*, denoted as *unconditional (or exact) least-squares estimates* [29]. For large series, maximum likelihood and (conditional or unconditional) least squares estimators are identical and both approximately unbiased and normally distributed [31].

3.4.3 Model diagnostic

Once model identification and parameter estimation have been completed, the following step is to assess the adequacy of the fitted model to represent the time series under study. Usual approaches for diagnostic checking purposes include residual analysis and overfitting.

Residual analysis

Considering an ARMA model $\phi(B)z_t = \theta(B)w_t$ fitted to some time series, with parameters estimated by maximum likelihood $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$, the estimates for the error,

$$\widehat{w}_t = \widehat{\theta}^{-1}(B)\widehat{\phi}(B)z_t,$$

are known as *residuals*. If the fitted model is adequate to the data, it can be proved that

$$\widehat{w}_t = w_t + O\left(\frac{1}{\sqrt{n}}\right),$$

that is, the largest the sample, the closest $\widehat{w}_t$ becomes to white noise, w_t [29]. Thus, by taking white noise as model, it can be assessed through residuals whether the selected model is a good fit for the time series. If the residuals do not satisfy this assumption, a more appropriate model should be fitted, by returning to the stage of model identification. Thus, one should begin the model diagnostic check by visually inspecting the plot of residuals to investigate possible departures from randomness, as the pattern of the residuals over time can provide further insights about the (in)adequacy of the model. If the model fits the data, one will expect to see a random scatter around zero, with no trends or patterns. Yet, failure to identify patterns or trends in the plot of residuals is not alone indicative of a good fit of the model. In Figure 3.7 there are represented the standardized residuals of models AR(1) and MA(1) fitted to the same AR(1) process, simulated in Figure 3.2, and, as one could see, there are no obvious departures from randomness when the wrong model, MA(1), is applied to the series.

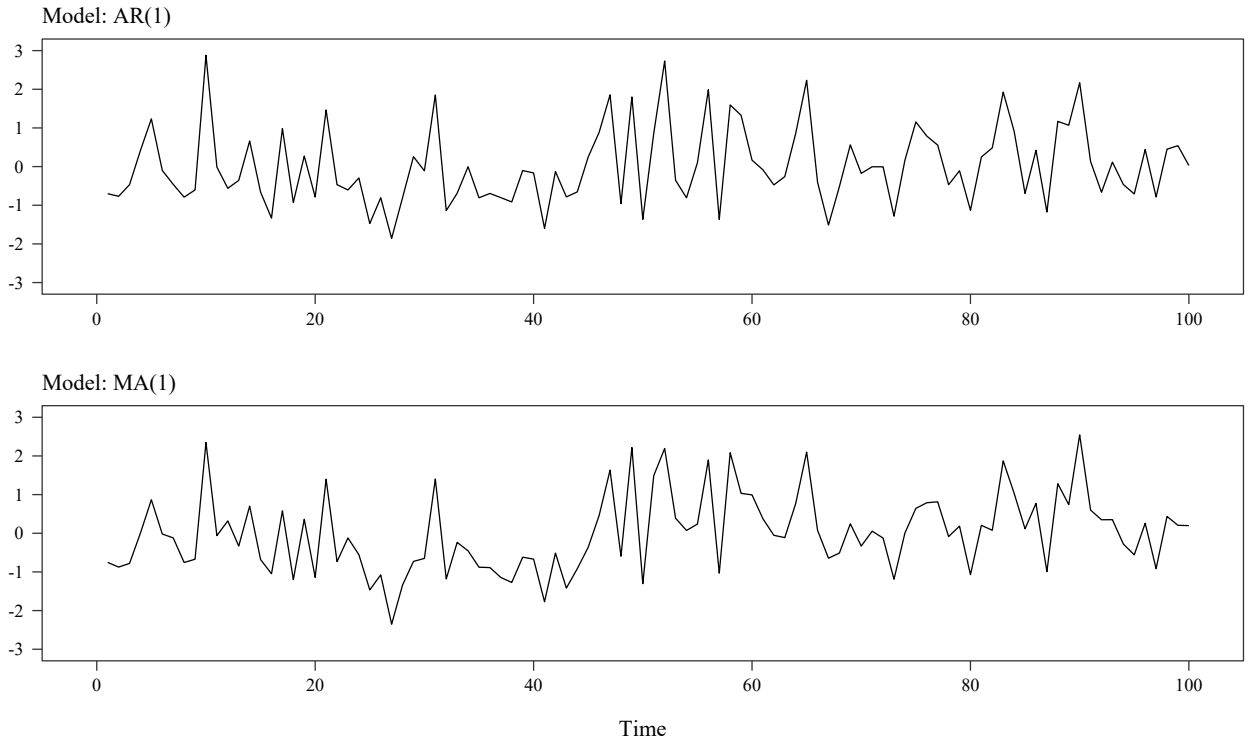

Figure 3.7: Standardized residuals of AR(1) and MA(1) models applied to an AR(1) process.

In complement, the quantile–quantile (Q–Q) plot applied to the residuals can point out the presence of outliers and provide insights about the adequacy of a normal approximation [29].

Autocorrelation function

To assess model adequacy, the autocorrelation function of the residuals is a useful tool, as it can point out apparent discrepancies from white noise [29].

Given a model perfectly identified, with known parameters, w_t's would be white noise with autocorrelations $r_k(w)$ uncorrelated and approximately normal, with zero mean and variance n^{-1}. In practice, one can only obtain $\hat{w}_t$, which, despite its valuable input about the nature of model inadequacy, has slightly different properties than w_t. Box and Pierce (1970) have shown that, at low lags, the $r_k(\hat{w})$'s can be extremely correlated, with variance considerably smaller than n^{-1}, but such effect dissipates for larger lags. Thus, infer on departures of $r_k(\hat{w})$ from their theoretical value of zero based on a standard error of $n^{-1/2}$ can be misleading for low lags, although its employment is reasonable for moderate to high lags [29]. If the residuals indicate some lack of fit, the model should be modified as suggested by the autocorrelations (e.g., extend an AR(1) to an AR(2) model) [31].

Portmanteau test

As a complement to the analysis of $r_k(\hat{w})'s$ individually, it is useful to know whether the first K autocorrelations of the residuals, as a whole, suggest lack of fit. Then, given a sufficiently large number of autocorrelations (so that the weights of ψ_j become negligible for $j > K$), $r_k(\hat{w})$, $k = 1, 2, \dots, K$, from an adequate ARIMA(p, d, q), the statistic

$$Q = n \sum_{k=1}^{K} r_k^2(\hat{w}),$$

where $n = N - d$, approximates to a χ^2 distribution with $K - p - q$ degrees of freedom. Fitting an inadequate model will inflate the value of Q. Therefore, a general 'portmanteau' test would reject the null hypothesis of model adequacy [29].

Ljung and Box (1978) later shown that, under the null hypothesis, this is not a satisfactory approximation for typical sample sizes, even for $n = 100$ [31]. A modified statistic was then proposed by the authors,

$$\tilde{Q} = n(n+2) \sum_{k=1}^{K} (n-k)^{-1} r_k^2(\hat{w}),$$

including a more accurate value for the variance of $r_k(w)$ [29]. Compared to Q, the Ljung-Box statistic is closer to the χ^2 distribution [31].

Cumulative Periodogram

When fitting some time series, mainly seasonal data, one must assure that specific characteristics, as nonrandom periodic oscillations, are taken into account. Such patterns can be detected in the periodogram of the residuals, defined as

$$I(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} w_t \cos(2\pi f_i t)\right)^2 + \left(\sum_{t=1}^{n} w_t \sin(2\pi f_i t)\right)^2\right],$$

with frequency $f_i = i/n$. A large value of $I(f_i)$ occurs when, for a given frequency, a pattern in the residuals correlates with a cosine or sine wave. Thus, for effective checking of periodic nonrandomness, the normalized cumulative periodogram

$$C(f_j) = \frac{\sum_{i=1}^{j} I(f_i)}{ns^2},$$

where s^2 is an estimate of σ_w^2, can be analysed. If the fitted model is adequate, then the series of w_t is white noise and the plot of $C(f_j)$ against f_j would display points along a straight line from the origin to $(0.5, 1)$. As opposed, nonrandom residuals obtained from not so well adjusted models would originate cumulative periodograms with recurrent deviations from such line [29].

Overfitting

Typically a problem, to be avoided when adjusting any model, overfitting reveals to be useful as a diagnostic tool in time series analysis. This technique consists of fitting a more elaborate model than the one previously identified, and believed as adequate. This new model contains additional parameters, to be estimated and, thus, check if they are really necessary. If a parameter estimate reveals not to be significantly different from zero, it does not prove that the identified model is the correct one, but it would support the choice of the simplified model. A smaller value of $\hat{\sigma}^2$ for this model would emphasize such conclusion, as one is led to believe that it is a better fit for the data [29].

In ARMA models, orders *p* and *q* could both be increased, but not simultaneously, that is, similar terms should not be added at the same time to both sides of the model in order to prevent parameter redundancy. The model $(1-cB)\phi(B)z_t = (1-cB)\theta(B)w_t$ would be as correct as $\phi(B)z_t = \theta(B)w_t$ for any arbitrary constant *c*, but lacks uniqueness and cancellation would be needed. Ideally, residual autocorrelations should be considered when deciding the direction to take [31].

3.5 Forecasting

When building a time series model, one of the main goals is surely to predict values of the process at future moments in time. In healthcare, as in business or finance, reliable forecasting provides valuable insights on what to expect, supporting decision making on assuring the best possible care to individuals.

Box-Jenkins ARIMA models, previously presented, may be used to forecast future values of some observed time series, taking as assumption that the model is correctly specified, that is, it is known exactly. Although it is not possible in practice to ensure this assumption, errors in the estimated parameters will not notably affect the forecasts for moderate to large time series [29].

3.5.1 Minimum mean square error forecasts

Given the values of a historical series up to time t, one would like to estimate the future observation

$$y_{t+l} = \varphi_1 y_{t+l-1} + \cdots + \varphi_{p+d} y_{t+l-p-d} + w_{t+l} - \theta_1 w_{t+l-1} - \cdots - \theta_q w_{t+l-q}. \tag{3.55}$$

The forecast of this value, $\hat{y}_t(l)$, is said to be made at time *origin* t for lead *time* l ($l > 0$). It could be defined as a linear function of actual and previous shocks $w_t, w_{t-1}, ...$, so that the best possible forecast is, supposedly,

$$\hat{y}_t(l) = \psi_l^* w_t + \psi_{l+1}^* w_{t-1} + \psi_{l+2}^* w_{t-2} + \cdots,$$

for which the weights $\psi_l^*, \psi_{l+1}^*, ...$ need to be determined [29]. Then, given the expression of the future observation y_{t+l} as an infinite weighted sum of shocks,

$$y_{t+l} = \sum_{j=0}^{\infty} \psi_j w_{t+l-j}, \tag{3.56}$$

the *mean square error* of such forecast is

$$E[y_{t+l} - \hat{y}_t(l)]^2 = \sigma_w^2 \left[(1 + \psi_1^2 + \cdots + \psi_{l-1}^2) + \sum_{j=0}^{\infty} (\psi_{l+j} - \psi_{l+j}^*)^2 \right], \tag{3.57}$$

minimized when $\psi_{l+j} = \psi_{l+j}^*$. It follows that

$$\begin{aligned} y_{t+l} &= (w_{t+l} + \psi_1 w_{t+l-1} + \cdots + \psi_{l-1} w_{t+1}) + (\psi_l w_t + \psi_{l+1} w_{t-1} + \cdots) \\ &= e_t(l) + \hat{y}_t(l), \end{aligned} \tag{3.58}$$

where

$$\begin{aligned} e_t(l) &= y_{t+l} - \hat{y}_t(l) \\ &= w_{t+l} + \psi_1 w_{t+l-1} + \cdots + \psi_{l-1} w_{t+1} \end{aligned} \tag{3.59}$$

is the l-steps-ahead forecast error, that is, the error of the forecast $\hat{y}_t(l)$ at *lead time* l, also called forecast horizon [29, 41].

From this point, it can be realized that the shocks w_t, so far presented as random independent variables, are, in fact, the *one-step-ahead forecast errors*:

$$\begin{aligned} e_t(1) &= y_{t+1} - \hat{y}_t(1) \\ &= w_{t+1}, \end{aligned} \tag{3.60}$$

which must be uncorrelated for a minimum mean square error forecast. Otherwise, each forecast could be predicted by previously ones and such dependence on the history of the process could be exploited in order to improve the forecast (in that case, $\hat{y}_t(l)$ would not be the best prediction for y_{t+l}) [29, 31]. This does not necessarily apply at longer lead times, with forecast errors made at the same lead time l from different time origins t being generally correlated. Also, forecast errors made at different lead times for the same origin are highly correlated. By fixing t, $\hat{y}_t(l)$ becomes a function of l — *forecast function* for origin t — with the existing correlation between forecast errors made at different lead times thus justifying the tendency for it to often lie either wholly above or below the values of the series [29].

As the relation established in Equation 3.60 implies, the variance of the forecast error at $l = 1$ is

$$V\Big(e_t(1)\Big) = \sigma_w^2. \tag{3.61}$$

For the general forecast error, $e_t(l)$, it is to note that $E_t[e_t(l)] = 0$, and so the forecast is unbiased. Furthermore, from Equation 3.59 results the variance of the forecast error,

$$V(l) = \text{Var}[e_t(l)] = \sigma_w^2 \left(1 + \sum_{j=1}^{l-1} \psi_j^2\right), \tag{3.62}$$

which can be interpreted as the expected value of $e_t^2(l)$, $E[y_{t+l} - \hat{y}_t(l)]^2$. Thus, turns out clear that the variance of the error increases as so increases the lead time l [29, 31].

An important fact can be further stated, by keeping the assumption that w_t are a sequence of independent random variables. As $E[w_{t+j|y_t, y_{t-1,\ldots}}] = 0$, $j > 0$, it follows from Equation 3.56 that

$$\begin{aligned}\hat{y}_t(l) &= \psi_l w_t + \psi_{l+1} w_{t-1} + \cdots \\ &= E_t[y_{t+l}],\end{aligned} \tag{3.63}$$

where $E_t[y_{t+l}]$ denotes the conditional expectation of y_{t+l} given the values of the process up to time t, $E_t[y_{t+l}|y_t, y_{t-1,\ldots}]$, to which equals the minimum mean square error forecast at origin t, for lead time l. This relation extends to any linear function of the forecasts, $\sum_{l=1}^{L} g_t \hat{y}_t(l)$, as it is also a minimum mean square error forecast of identical linear function of future observations, $\sum_{l=1}^{L} g_t y_{t+l}$ [29].

3.5.2 Forecasts and probability limits calculation

Given the relation established in Equation 3.63, the forecasts can be expressed in terms of difference equation as

$$[y_{t+l}] = \hat{y}_t(l) = \varphi_1[y_{t+l-1}] + \cdots + \varphi_{p+d}[y_{t+l-p-d}] - \theta_1[w_{t+l-1}] - \cdots - \theta_q[w_{t+l-q}] + [w_{t+l}], \tag{3.64}$$

where, by convention, the square brackets indicate that conditional expectations, at time t, must be taken, so that, $[w_{t+l}] = E_t[w_{t+l}]$ and $[y_{t+l}] = E_t[y_{t+l}]$ [29]. To calculate them, the following is to note:

$$[y_{t+j}] = \begin{cases} y_{t+j}, & j \leqslant 0 \\ \hat{y}_y(j), & j > 0 \end{cases} \qquad \text{and} \qquad [w_{t+j}] = \begin{cases} w_{t-j} = y_{t-j} - \hat{y}_{t-j-1}(1), & j \leqslant 0 \\ 0, & j > 0. \end{cases}$$

So, y_{t-j} and w_{t-j}, which already occurred and are available at time t, are used directly in the equation to obtain the forecasts, while y_{t+j} are replaced by their forecasts $\hat{y}_t(j)$, and w_{t+j} by zeros, as they have not yet occurred. Using Equation 3.64, the forecasts $\hat{y}_t(l)$ can then be calculated recursively in the order $\hat{y}_t(1), \hat{y}_t(2), \ldots$ as

$$\hat{y}_t(l) = \sum_{j=1}^{p+d} \varphi_j \hat{y}_t(l-j) - \sum_{j=l}^{q} \theta_j w_{t+l-j},$$

where $\hat{y}_t(-j) = [y_{t-j}]$ is the observed value y_{t-j} for $j \geqslant 0$ [29].

Based on the assumption that the variables w_t are normally distributed, then the probability distribution of y_{t+1}, conditional on the information available up to time t, $p(y_{t+1}|y_t, y_{t-1,\ldots})$, will also be normal with mean $\hat{y}_t(l)$ and standard deviation

$$\sigma(l) = \sigma_w \left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)^{1/2}. \tag{3.65}$$

It follows that $(y_{t+l} - \hat{y}_t(l)) \,/\, \sigma(l)$ will have a standard normal distribution. Thus, the probability limits for the forecast errors will be defined by $\hat{y}_t(l) \pm z_{1-\alpha/2} \cdot \sigma(l)$, where α is the level of significance adopted and $z_{1-\alpha/2}$ is the quantile of probability $1 - \alpha/2$ of the standard normal distribution [29]. Hence, given the information available until t, there is a probability of $1 - \alpha$ that the observed value of the process at time $t + l$ will be within these limits, that is

$$P\Big(\hat{y}_t(l) - z_{1-\alpha/2} \cdot \sigma(l) < y_{t+l} < \hat{y}_t(l) + z_{1-\alpha/2} \cdot \sigma(l)\Big).$$

In practice, σ_w is replaced by an estimate of the standard deviation of the process w_t, s_w, in Equation 3.65 to obtain $\sigma(l)$. In order to compute these limits, it will be further needed to calculate the weights $\psi_1, \psi_2, ...$ Knowing the values of φ and θ, one can then resort on the relation $\varphi(B)\psi(B) = \theta(B)$, that is,

$$(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_{p+d} B^{p+d})(1 + \psi_1 B + \psi_2 B^2 + \cdots) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q),$$

to recursively obtain the values of ψ as

$$\begin{cases} \psi_1 = \varphi_1 - \theta_1 \\ \psi_2 = \varphi_1 \psi_1 + \varphi_2 - \theta_2 \\ \vdots \\ \psi_j = \varphi_j \psi_{j-1} + \cdots + \varphi_{p+d} \psi_{j-p-d} - \theta_j, \end{cases} \tag{3.66}$$

with $\psi_0 = 1$, $\psi_j = 0$ for $j < 0$, and $\theta_j = 0$ for $j > q$ [29].

These same weights can be used to update the forecasts, once a new value becomes available. Considering the forecasts of y_{t+l+1} made for lead times $l + 1$ and l at time origins t and $t + 1$, respectively,

$$\begin{aligned} \hat{y}_t(l+1) &= \varphi_{l+1} w_t + \varphi_{l+2} w_{t-1} + \cdots \\ \hat{y}_{t+1}(l) &= \varphi_l w_{t+1} + \varphi_{l+1} w_t + \varphi_{l+2} w_{t-1} + \cdots, \end{aligned}$$

it follows

$$\hat{y}_{t+1}(l) = \hat{y}_t(l+1) + \varphi_l w_{t+1}, \tag{3.67}$$

meaning that the forecast at origin t can be updated by adding a multiple of the one-step-ahead forecast error, $w_{t+1} \equiv y_{t+1} - \hat{y}_t(l)$, and so become the forecast at origin $t + 1$ for the same value of the process. Hence, once the new value y_{t+1} is known, the forecasts for lead times $1, 2, \ldots, L$ at origin t can be updated through Equation 3.67 to obtain forecasts for $1, 2, \ldots, L - 1$ at $t + 1$. At this time origin, the L-step-ahead forecast, $\hat{y}_{t+1}(L)$, could be easily obtained using Equation 3.64 [29].

The conceptualizations presented above could be extended to seasonal models. Thus, taking as example the process SARIMA$(0,1,1)\times(0,1,1)_{12}$ simulated in subsection 3.3.2, $\nabla\nabla_{12}y_t = (1-\theta B)(1-\Theta B^{12})w_t$, the future observation y_{t+l} could be expressed as

$$y_{t+l} = y_{t+l-1} + y_{t+l-12} - y_{t+l-13} + w_{t+l} - \theta w_{t+l-1} - \Theta w_{t+l-12} - \theta\Theta w_{t+l-13},$$

with the minimum square error forecast at lead time l for origin time t, $\hat{y}_t(l)$, being given by

$$\hat{y}_t(l) = [y_{t+l-1} + y_{t+l-12} - y_{t+l-13} + w_{t+l} - \theta w_{t+l-1} - \Theta w_{t+l-12} - \theta\Theta w_{t+l-13}], \tag{3.68}$$

where, as one should recall,

$$[y_{t+l}] = E[y_{t+l}|y_t, y_{t-l}, \cdots; \theta, \Theta]$$

is the expectation of y_{t+l} at origin time t, conditional on the information available up to that moment and assuming that the parameters are known. As stated before, this assumption of exact parameters, although not true in practice, is acceptable given that small changes in the parameters due to estimation errors do not produce relevant changes in the forecasts [29].

For computational purposes, the calculation of the forecasts from the difference equation form of the model is as elegant as it is simple. Nonetheless, other points of view reveal more useful for the comprehension of their nature.

For the seasonal model referred above, it can be show that the forecasts satisfy the difference equation

$$(1-B)(1-B^{12})\hat{y}_t(l) = 0 \qquad l > 13, \tag{3.69}$$

with B here operating on lead time l. By writing l in the form $(r,m) = 12r + m$, $r \geqslant 0$ and $m > 0$, to represent a lead time of *r* years and *m* months (e.g., $l = 18 = (1,6)$), the solution of Equation 3.69 is given by

$$\hat{y}_t(l) = b_{0,m}^{(t)} + r b_1^{(t)} \qquad l > 0, \tag{3.70}$$

where the coefficients $b(t)$ are constants applied to all lead times for a given origin, but that are continuously changing with time t, allowing them to adapt to the specific part of the process under analysis. That being said, $b_{0,1}^{(t)}, b_{0,2}^{(t)}, \ldots, b_{0,12}^{(t)}, b_1^{(t)}$, determined from the 13 initial forecasts, represent 12 monthly and 1 yearly contributions [29].

Alternatively, $\hat{y}_t(l)$ can be represented as

$$\hat{y}_t(l) = \sum_{j=1}^{6} \left[b_{1j}^{(t)} \cos\left(\frac{2\pi jl}{12}\right) + b_{2j}^{(t)} \sin\left(\frac{2\pi jl}{12}\right) \right] + b_{16}^{(t)}(-1)^l + b_0^{(t)} + b_1^{*(t)} l,$$

providing a better understanding on the general pattern of the forecasts, perceived as a mixture of sinusoids at seasonal frequencies, along with a linear trend of slope $b_1^{*(t)}$, the monthly rate of change in the forecasts. As for the annual rate of change, it is represented by $b_1^{(t)} = 12 b_1^{*(t)}$ [29, 31].

Considering now the expression of the forecasts in Equation 3.70, the general updating formula takes two possible formulations, depending on the value of *m*.

So, if $m \neq s = 12$,

$$b_{0,m}^{(t+1)} + r b_1^{(t+1)} = b_{0,m+1}^{(t)} + r b_1^{(t)} + w_{t+1}(\lambda + r\lambda\Lambda),$$

where $\lambda = 1 - \theta$ and $\Lambda = 1 - \Theta$, results in the following updating formulas:

$$\begin{aligned} b_{0,m}^{(t+1)} &= b_{0,m+1}^{(t)} + \lambda w_{t+1} \\ b_1^{(t+1)} &= b_1^{(t)} + \lambda\Lambda w_{t+1}. \end{aligned} \tag{3.71}$$

If, instead, $m = s = 12$,

$$b_{0,12}^{(t+1)} + r b_1^{(t+1)} = b_{0,1}^{(t)} + (r+1) b_1^{(t)} + w_{t+1}(\lambda + \Lambda + r\lambda\Lambda),$$

the equations to update the forecasts are

$$\begin{aligned} b_{0,12}^{(t+1)} &= b_{0,1}^{(t)} + b_1^{(t)} + w_{t+1}(\lambda + \Lambda) \\ b_1^{(t+1)} &= b_1^{(t)} + \lambda\Lambda w_{t+1}. \end{aligned} \tag{3.72}$$

In this case, as for ARIMA models, the forecast $b_{0,m}^{(t+1)}$ is the updated version of $b_{0,m+1}^{(t)}$. Thus, if one defines January as the origin t, $b_{0,2}^{(t)}$ would be the estimate to March of the same year. Once in February, the update of this estimate will then be $b_{0,1}^{(t+1)}$ [29].

3.5.3 Forecasting accuracy

With important decisions being made based on forecasts, it becomes crucial to evaluate the performance of the selected model. More, measures of accuracy and reliability can help to decide between parameter sets or different models applied to the same time series. Being said that, the evaluation of the accuracy of the forecasts should follow the aforesaid steps in time series modelling [42].

Accuracy assessment should preferentially make use of out-of-sample tests instead of tests of goodness of fit on the past data, known as in-sample tests. Model identification and estimation are intended to adjust the forecasting method to the historical series. Though, overfitting may accentuate discrepancies between in and out of sample results as subtleties of the series in the past may not persist into the future, in the same way as peculiarities of future values may not have revealed themselves in the past, making them unlikely to anticipate through previous observations. It follows that, as one would expect, the forecast errors obtained out of sample generally exceed in-sample errors, also called forecasting residuals (difference between each observed value of the series and its fitted value), even for relatively short forecast horizons. More, the models that perform better in in-sample tests may not be the ones with the best out-of-sample forecasts [41].

The most obvious way to assess how accurate the obtained forecasts are would be to wait and compare them with actual values of the series as they occur in real time, but this has practical limitations as a long wait would be to expect before a reliable forecasting picture could be taken. In light of the above, *out-of-sample* evaluation has been widely applied. The first step is to divide the historical time series into a training set and a test set. The former is used exclusively to identify and estimate the model, whereas the test (held-out) data are reserved to evaluate the forecasting accuracy of the model [41].

As disadvantage there is the fact that not all the data available are used, although they would be useful as part of the trained data, specially if the data set is small, in order to achieve better results. By leaving apart some of the data when training the model, diversity is not as much as it could be and the errors obtained might represent features of the test set in particular, not observed in the remaining data [42].

In out-of-sample evaluation, either a single forecast origin — time point from which the forecast is made — or multiple forecasts origins can be used [41].

Fixed-origin evaluation

In fixed-origin evaluation, forecasts for lead times $l = 1, 2, \dots, L$ are generated from a unique origin t which, in this context, would be the last value of the training set (Figure 3.8.A). So, as only one forecast (and so, one forecast error) per lead time can be computed, errors can not be averaged for a single time series (averaging errors across lead times would be possible but would not make sense given their theoretical behaviour). Besides, by fixing the origin, the forecasts obtained are conditioned by particular characteristics of the series at this time point. Hence, the behaviour of the series at the forecast origin highly affects the results of the evaluation [42].

Rolling-origin evaluation

In the rolling-origin evaluation, also known as *n*-step-ahead evaluation, the forecast origin is updated and a new observation is added to the training data (Figure 3.8.B). Each update implies a revision of the forecasting equation, which can resume simply to the addition of a new observation to the training set — *Rolling-origin-*

update evaluation —, or may arise from the update of the imputed data plus recalibration (re-estimation) of the model — *rolling-origin-recalibration evaluation*. This last procedure desensitizes error measures to specific aspects of the original training set. Thus, although more computationally expensive, from a theoretical point of view it is preferable than simply updating [41]. Either way, one can assess the accuracy of the forecasts of a unique time series at each lead time by averaging the forecast errors [41, 42]. Compared to the fixed-origin evaluation, from which result N forecasts and respective errors, the rolling-origin procedure yields $N(N+1)\,/\,2$ forecasts, where N represents the length of the test set [41].

Implicitly, the data set used to obtain a forecast for a given horizon l becomes larger as the values of the test set are sequentially moved to the training set. Nonetheless, it is possible to maintain the training set of constant length by trimming the oldest value of the series at each update, in a procedure known as fixed-size, *rolling window* (Figure 3.8.C). By doing so, one can remove old data influence when re-estimating the model. Furthermore, this scheme allows a 'fair' comparison of forecasting accuracy between multiple periods of test, non-confounded by the use of different data sets to train the model [41].

When deciding the number of time points N to hold out from the series, one should have in mind what is the longest lead time required, L. Thus, the minimum N would be as large as L, with multiple forecasts computed for every horizon but the longest. If the series is not excessively small, the length of the test set could be increased so a minimum number of forecasts M can be obtained at lead time L. In that case, the length of the test set should equal $L + M - 1$. For short time series, it would be preferable to benefit from the rolling-origin and evaluate, at the worst scenario, one-step-ahead forecasts than truncate the data and leave too few observations to estimate the model [41].

Accuracy measures

Different accuracy measures can be computed and used to compare the forecasting performance (in practice, the magnitude of forecast errors) of different models. As detailed by Hyndman and Koehler [43], such measures can being classified in four major groups, namely:

Scale-dependent measures

Measures whose scale depends on the data are useful, and of proper use, only when comparing models applied to the same data set. From these measures, the most commonly used are based on absolute or squared forecast errors, $e_t(l) = y_{t+l} - \hat{y}_t(l)$.

The Mean Square Error (MSE),

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t(l)^2,$$

is widely used as accuracy measure, mainly because of its relevance from a theoretical point of view, but its application has been recommended against due to its sensibility to outliers [43].

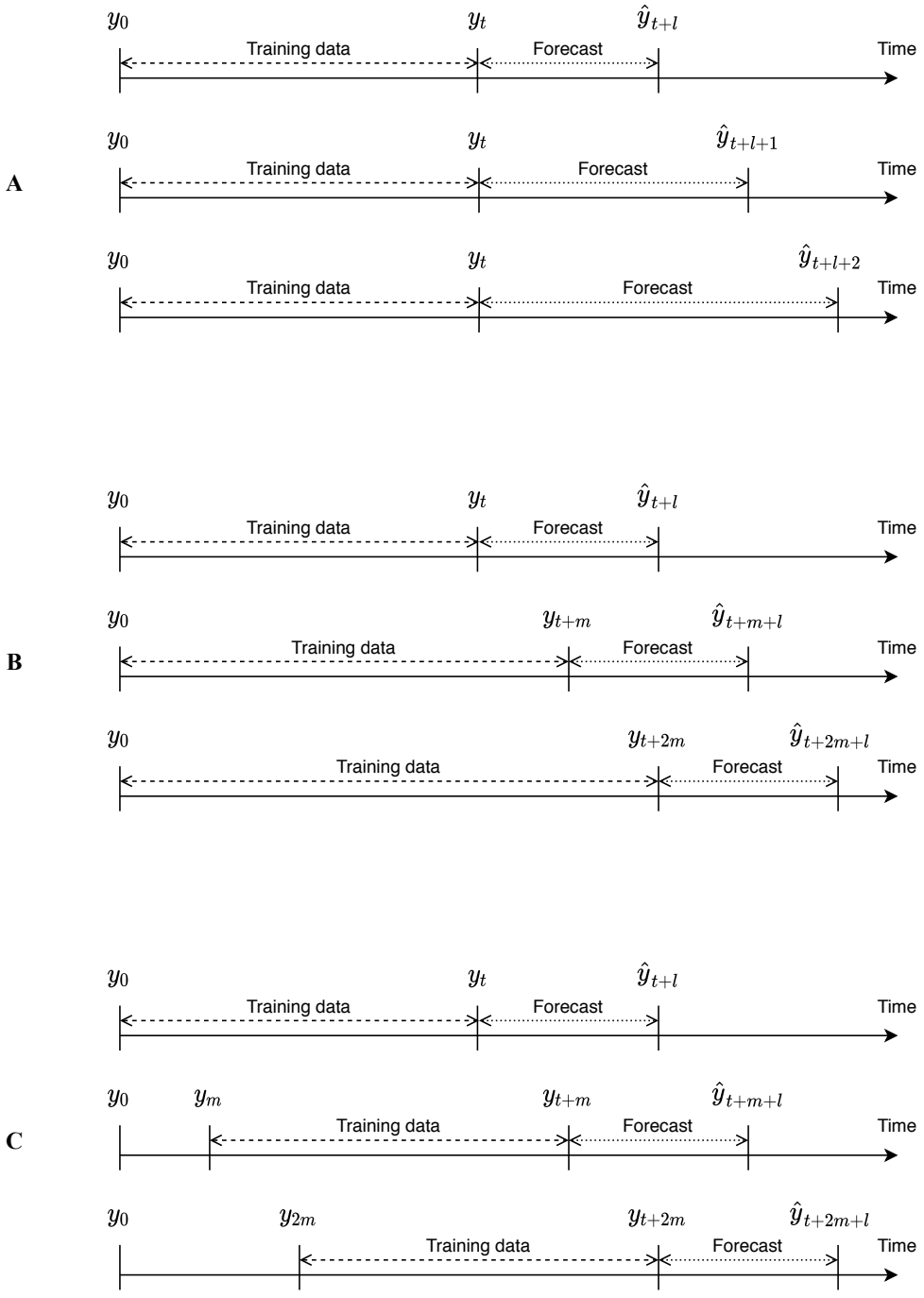

Figure 3.8: Schematic representation of (A) fixed-origin evaluation, (B) rolling-origin evaluation with fixed window and (C) rolling-origin evaluation with rolling window. Adapted from Becerra *et al.* [44].

The above also applies to the Root Mean Square Error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t(l)^2},$$

with the difference that this measure is on the same scale of the data, reason why it tends to be preferred to the MSE. Nevertheless, both measures are more sensitive to outliers than the Mean Absolute Error (MAE),

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t(l)|.$$

Measures based on percentage errors

Accuracy measures that do not depend on the scale of the data are mostly used to compare forecasting accuracy across different sets of data, but they can also be applied to series in the same scale.

This class of measures is mainly represented by the Mean Absolute Percentage Error (MAPE),

$$\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n} \left|\frac{e_t(l)}{y_{t+l}}\right|,$$

although other measures, similar to the scale-dependent ones aforementioned, can be computed. One example would be the Root Mean Square Percentage Error (RMSPE),

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(100\ \frac{e_t(l)}{y_{t+l}}\right)^2}.$$

Regardless of their formulation, measures based on percentage assume a natural zero, so it would make no sense to use them in evaluation of errors in data like temperature time series. Moreover, if the series does take the value of zero at some t, the MAPE turns infinite or undefined (if $\hat{y}_t(l) = y_{t+l} = 0$) [43]. Even if the probability of exact zeros occurring is very low, small values of the series (in the denominator) are associated with high percentage errors, so that the MAPE shows a highly skewed distribution whenever y_t approximates zero [42].

Measures based on relative errors

Relative errors provide a different perspective on the quality of the forecasts obtained, as their errors, $e_t(l)$, are compared to the ones obtained with a benchmark method, $e_t(l)^*$, when applied to the same set of data. Therefore, let $r_t(l) = e_t(l) \,/\, e_t(l)^*$ denote the relative error, from which one can define accuracy measures, such as the Mean Relative Absolute Error (MRAE),

$$\text{MRAE} = \frac{1}{n}\sum_{t=1}^{n} |r_t(l)|.$$

As forecasting benchmark methods, the naïve method, with the forecasts being equal the last observed value in the series, as well as the average method, where $\widehat{y}_t(l)$ is equal to the mean of the historical series, have been frequently used. For seasonal data, an extension of the naïve method, which gives forecasts equal to last value of the series adjusted for seasonality, has been considered [43]. Though, problems regarding zero values remain [42].

Relative measures

As an alternative to relative errors, relative measures can be calculated. For instance, given the MAE for the benchmark method, MAE*, the relative MAE (RelMAE) can be defined as

$$\text{RelMAE} = \frac{\text{MAE}}{\text{MAE*}},$$

and measures the improvement from the chosen forecast method in relation to the benchmark: when $\text{RelMAE} < 1$ the proposed method is said to perform better than the benchmark method; implicitly, $\text{RelMAE} > 1$ means that the chosen method is worse than the benchmark forecast method. By analogy, other relative measures can be obtained using either scale-dependent or percentage errors [43]. When the naïve method is used as benchmark, the relative RMSE is also known as Theil's U statistic [42].

Aside from their interpretability, relative measures can sidestep the problems with zeros presented by other measures. Yet, these measures has the shortcoming of depending on multiple forecasts per series and/or horizon [42, 43].

Chapter 4

Analysis of hospitalizations due to diabetes

From 2010 until 2018, there were 208,882 hospital admissions due to diabetes in Portugal, including hospitalizations, the object of study in this thesis. An exploratory analysis of these data is made in Section 4.1, followed by the application of the theoretical methodology on time series modelling and forecasting presented in the previous chapter to the series of monthly hospitalizations due to diabetes in Sections 4.2 and 4.3, respectively. Data analysis was conducted using the software R, version 3.6.3 [45], with a significance level of 0.05.

4.1 Exploratory analysis

Between January 2010 and December 2018, inclusive, 73,050 hospitalizations due to diabetes were accounted in Portugal (average of 676 cases per month or 8,117 cases per year), representing 35% of all admissions with diabetes as the main cause in the same period. From those, 72,698 (99.5%) are associated with a fictitious identification number, which makes it possible to determine how many episodes correspond to a particular person, that is, how many times each patient was hospitalized due to diabetes during the study period. Accordingly, a total of 48,308 people (51% men) were hospitalized just once, whereas 9,599 (55% men) were hospitalized at least two times (71% hospitalized twice), to which correspond 24,390 hospitalizations (2.5 episodes per person on average). The highest number of hospitalizations due to diabetes by a single person was 34. In total, 57,907 individuals (not accounting for those who would be associated with the 352 episodes without patient fictitious number) were hospitalized due to diabetes between 2010 and 2018, in Portugal.

Focusing on the series of hospitalizations, sex distribution is almost even, with men accounting for 52.5% of the cases, in a total of 38,375 hospitalizations, about 3,700 more than women (n = 34,674). The age of the patients ranges from 0 to 107 years, with people aged 60 or above representing almost two thirds of the total number of hospitalizations in the period under study (64.4%). Among men, the distribution of

cases is slightly different, as individuals between 40 and 79 years old are the most represented ones (68.4%). With regard to patients' region of residence at the moment of admission, Norte (15% of total cases in the district of Porto) and Área Metropolitana de Lisboa (24% of total cases in the district of Lisbon) are the most represented NUTS 2, being associated with 30.4% and 28.6% of the cases, respectively (Table 4.1). At this level, considering the number of hospitalizations per 100,000 inhabitants per year, different trends can be observed across locations. Autonomous regions of Açores and Madeira, with a very low number of cases for a large period, show an increase in the number of hospitalizations per 100,000 inhabitants in recent years, whereas in mainland Portugal a decreasing trend is observed (Appendix A, Figure A.1).

Table 4.1: Sociodemographic characteristics of patients hospitalized due to diabetes, at the moment of admission.

	Men (n = 38,375)	**Women (n = 34,674)**	**Total (N = 73,050)**
Age, % (n)			
0−19	8.2% (3,147)	8.9% (3,095)	8.5% (6,242)
20−39	7.7% (2,944)	8.3% (2,872)	8.0% (5,816)
40−59	22.5% (8,646)	15.2% (5,265)	19.0% (13,911)
60−79	45.9% (17,595)	40.1% (13,909)	43.1% (31,504)
$\geqslant$ 80	15.7% (6,043)	27.5% (9,533)	21.3% (15,577)
Region, % (n)			
Norte	29.6% (11,151)	31.2% (10,617)	30.4% (21,768)
Centro	24.4% (9,195)	25.1% (8,524)	24.7% (17,719)
Área Metropolitana de Lisboa	29.4% (11,076)	27.6% (9,396)	28.6% (20,473)
Alentejo	11.0% (4,137)	11.3% (3,857)	11.2% (7,994)
Algarve	4.2% (1,563)	3.7% (1,273)	4.0% (2,836)
Região Autónoma da Madeira	0.9% (325)	0.5% (170)	0.7% (495)
Região Autónoma dos Açores	0.5% (191)	0.5% (182)	0.5% (373)

Episodes with missing information regarding patient's sex not shown (n = 1; Age: $\geqslant$ 80 years; Region: Área Metropolitana de Lisboa).

Regarding the mode of admission, 79.5% (58,048 records) of the hospitalizations due to diabetes were classified as an emergency, while the remaining were planned admissions (n = 15,002). Generally speaking, men tend to be hospitalized for longer periods than women, also differing in the main causes of admission. For the former, the most common diagnosis is diabetes with circulatory complications (28.0%), whereas women are hospitalized mainly due to diabetes with other complications (than the ones detailed in other diagnosis) or diabetes without complications (18.4% and 18.2%, respectively). In turn, destination after discharge is similar between these groups, with patients returning home in more than 90% of the times. For the rest, about 4% of the patients died and 2.5% were discharged to another institution. Other destinations were also recorded, together representing 2.5% of the cases (Table 4.2).

When the analysis is made by year, a decreasing trend is observed, with the number of hospitalizations dropping by 45% from 2010 to 2018 (10,011 and 5,530 cases, respectively). The distribution of patients by age has remained relatively stable over the years, resulting in a global median of 67 years (interquartile range,

IQR: 52 – 78). Comparing men to women, the higher number of hospitalizations by the former is transversal to all years in the analysed period. The same is observed for the number of emergent admissions, with an accentuated difference, in percentage points, between emergencies and scheduled admissions from 2010 to 2018 (Table 4.3).

Table 4.2: Clinical characteristics of patients hospitalized due to diabetes, at the moment of admission.

	Men (n = 38,375)	**Women (n = 34,674)**	**Total (N = 73,050)**
Admission mode, % (n)			
Planned	22.1% (8,475)	18.8% (6,527)	20.5% (15,002)
Emergency	77.9% (29,900)	81.2% (28,147)	79.5% (58,048)
Diagnosis, % (n)			
Diabetes without complications	13.7% (5,246)	18.2% (6,321)	15.8% (11,567)
Diabetes with hyperosmolarity	3.8% (1,466)	7.2% (2,508)	5.4% (3,974)
Diabetes with ketoacidosis	14.1% (5,395)	17.8% (6,188)	15.9% (11,583)
Diabetes with other coma	1.1% (428)	1.6% (542)	1.3% (970)
Diabetes with kidney complications	10.7% (4,100)	11.1% (3,855)	10.9% (7,956)
Diabetes with ophthalmic complications	7.6% (2,929)	7.5% (2,603)	7.6% (5,532)
Diabetes with neurological complications	2.9% (1,096)	2.0% (705)	2.5% (1,801)
Diabetes with circulatory complications	28.0% (10,742)	14.9% (5,173)	21.8% (15,915)
Diabetes with other specified complications	17.3% (6,636)	18.4% (6,396)	17.8% (13,032)
Diabetes with unspecified complications	0.9% (337)	1.1% (383)	1.0% (720)
Days of hospitalization			
Range	1—545	1—437	1—545
Mean (SD)	12 (18)	10 (15)	11 (17)
Median (IQR)	7 (3—14)	6 (3—12)	7 (3—13)
Discharge destination, % (n)			
Discharge home	90.5% (34,722)	91.2% (31,606)	90.8% (66,328)
Another institution (with hospitalization)	2.9% (1,107)	2.1% (734)	2.5% (1,841)
Home care	0.4% (151)	0.5% (165)	0.4% (316)
Discharge against medical advice	0.9% (354)	0.7% (227)	0.8% (581)
Specialized aftercare (tertiary)	0.8% (316)	0.8% (270)	0.8% (586)
Palliative care at medical center	0.0% (15)	0.0% (16)	0.0% (31)
Post-hospital care	0.3% (122)	0.3% (110)	0.3% (232)
Long-term hospital care	0.1% (34)	0.1% (40)	0.1% (74)
Deceased	4.0% (1,554)	4.3% (1,506)	4.2% (3,061)

SD, Standard Deviation; IQR, Interquartile Range. Episodes with missing information regarding patient's sex not shown (n = 1; Mode of admission: emergency; Diagnosis: diabetes with kidney complications; Days of hospitalization: 19; Discharge destination: deceased).

Considering the number of hospitalizations by month, the main research series of this thesis, one can verify that hospitalizations due to diabetes dropped from 973 cases in January 2010 (alike March from the same year) to 400 in December 2018, with the minimum number of cases recorded in September 2018 (n = 365; Table 4.4).

Table 4.3: Descriptive statistics of hospitalizations due to diabetes by year.

	2010 (n = 10,011)	**2011 (n = 9,311)**	**2012 (n = 9,324)**	**2013 (n = 9,134)**	**2014 (n = 8,129)**	**2015 (n = 7,272)**	**2016 (n = 7,503)**	**2017 (n = 6,836)**	**2018 (n = 5,530)**
Monthly cases									
Range	705—973	664—888	670—912	670—878	599—803	520—755	550—690	505—670	365—609
Mean (SD)	834.2 (89.1)	775.9 (66.1)	777.0 (77.6)	761.2 (62.8)	677.4 (60.3)	606.0 (85.2)	625.2 (46.5)	569.7 (56.9)	460.8 (70.7)
Median (IQR)	815.5 (777.0—896.0)	771.0 (739.0—816.3)	759.0 (718.0—837.0)	758.0 (722.3—794.8)	664.0 (633.0—715.5)	577.0 (536.5—655.3)	631.5 (604.8—649.5)	551.0 (532.8—582.5)	445.5 (414.5—497.3)
Sex, n (%)									
Male	50.8% (5,081)	50.9% (4,740)	52.2% (4,864)	53.6% (4,894)	52.8% (4,293)	53.1% (3,864)	53.9% (4,046)	53.1% (3,630)	53.6% (2,963)
Female	49.2% (4,929)	49.1% (4,571)	47.8% (4,460)	46.4% (4,240)	47.2% (3,836)	46.9% (3,408)	46.1% (3,457)	46.9% (3,206)	46.4% (2,567)
Age									
Range	0—102	0—102	0—101	0—103	0—102	0—103	0—101	0—102	1—107
Mean (SD)	62 (21)	62 (21)	62 (22)	63 (22)	62 (22)	61 (23)	61 (23)	61 (23)	61 (23)
Median (IQR)	68 (53—78)	68 (53—78)	68 (52—78)	68 (54—78)	67 (52—78)	67 (50—78)	66 (50—78)	67 (50—78)	67 (49—78)
Admission mode, n (%)									
Planned	22.6% (2,267)	22.0% (2,051)	21.9% (2,041)	20.3% (1,850)	22.2% (1,805)	20.2% (1,467)	18.6% (1,399)	18.0% (1,232)	16.1% (890)
Emergency	77.4% (7,744)	78.0% (7,260)	78.1% (7,283)	79.7% (7,284)	77.8% (6,324)	79.8% (5,805)	81.4% (6,104)	82.0% (5,604)	83.9% (4,640)

SD, Standard Deviation; IQR, Interquartile Range.

Table 4.4: Monthly hospitalizations due to diabetes in Portugal from 2010 to 2018.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2010	973	887	973	923	851	705	808	729	765	781	793	823
2011	852	816	888	817	813	664	770	689	712	748	772	770
2012	912	856	831	722	855	737	695	670	706	829	781	730
2013	878	749	847	809	770	689	767	670	699	730	736	790
2014	803	726	747	685	712	630	655	610	599	670	658	634
2015	755	698	729	641	637	609	539	520	544	529	545	526
2016	666	644	688	634	690	617	629	613	552	550	640	580
2017	670	525	668	539	638	563	564	529	534	505	537	564
2018	609	522	555	489	452	446	413	419	365	445	415	400

Besides trend, some seasonality can be observed in the series, with differences being found between the number of hospitalizations due to diabetes per season ($\chi^2 = 422.65,\ p < 0.001$). A lower number of cases typically occurs in summer months, namely July, August, and September, accounting for 23% of all hospitalizations due to diabetes. As opposed, winter — January, February, and March — was the period of the year when the greatest number of hospitalizations occurred (28%). Annual peak occurs mainly in January, but also in March and May, while troughs are apparent between June and October. The average peak-to-trough amplitude (i.e., relative difference between the lowest and the highest observation at each year) was 27%, with a maximum of 40% in 2018 (Table 4.3 and Figure 4.1).

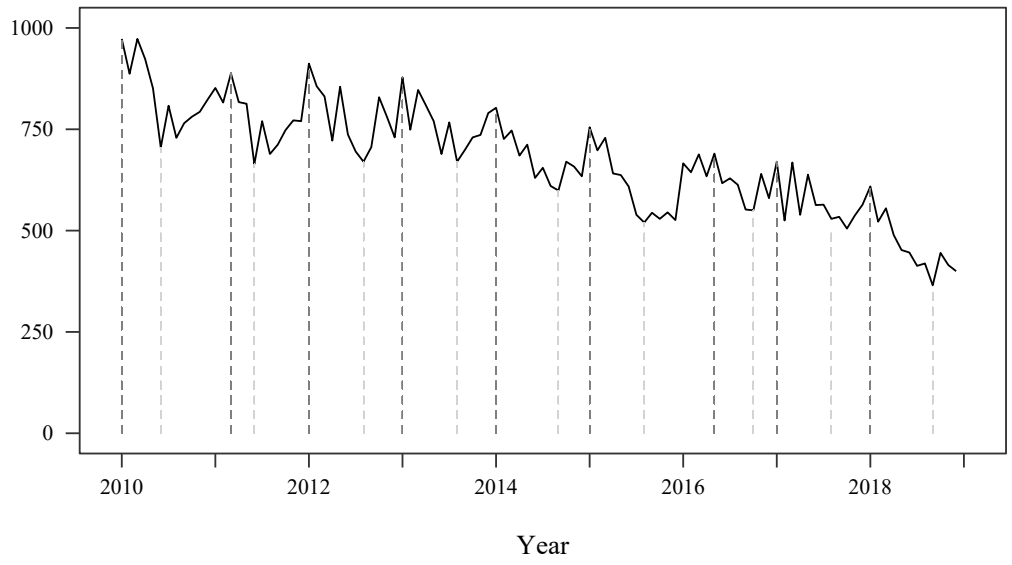

Figure 4.1: Monthly hospitalizations due to diabetes from 2010 to 2018. Vertical lines demonstrate seasonality, with peaks mainly in January and March (dark grey lines), and troughs between June and October (light grey lines). In 2010, January and March had the same number of hospitalizations, the highest of the year.

4.2 Model building and selection

When modelling the series of monthly hospitalizations due to diabetes, the last portion of the available data was reserved to further assess the performance of the selected model. Therefore, data from the first seven years, from January 2010 to December 2016, in a total of 84 months (60,684 hospitalizations), were used to build the model, leaving the data from the two remaining years (January 2017 to December 2018) for model evaluation (Figure 4.2).

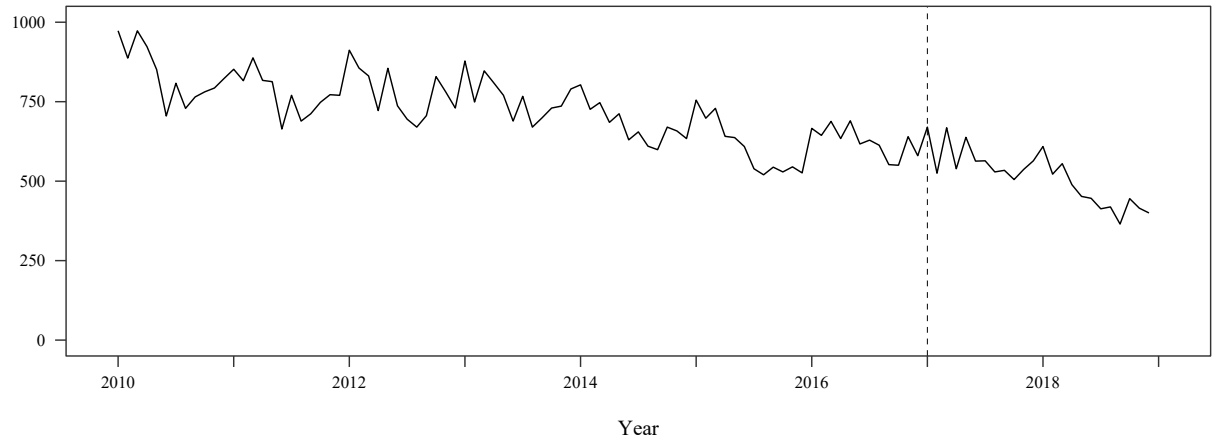

Figure 4.2: Monthly hospitalizations due to diabetes separated in training and test sets. Models were built using data from the first 84 months (training set, displayed at the left of the dashed vertical line), and evaluated against the remaining 24 months (test set, to the right of the dashed vertical line).

As stated for the entire series, the training data show a decreasing trend over the interval January 2010 – December 2018. By simply splitting this set in two, each with 42 observations, one can verify that the mean of the series decreases from the first half (January 2010 to June 2013: $E[y_t] = 794.95; \sigma_y = 78.43$) to the second half of the data (July 2013 to December 2016: $E[y_t] = 649.90; \sigma_y = 75.38$). Apart from this change in the level, the series does not show considerable variations in the dispersion of the observations over time. Nevertheless, Box-Cox approach (Equation 3.44) was followed to verify if some power transformation is suitable to these data (Figure 4.3).

The maximum likelihood estimate for λ is 0.4. Being close to 0.5, such value suggests that a square root of y_t is the adequate transformation to this series. However, the confidence interval (CI) for this parameter includes the value of 1, which itself suggests that no transformation is needed. For that reason, the original training series will be considered in further analysis.

By looking at the distribution of cases by month, as displayed in Figure 4.4, one can note some seasonal fluctuations over years, suggesting that a seasonal model should be investigated.

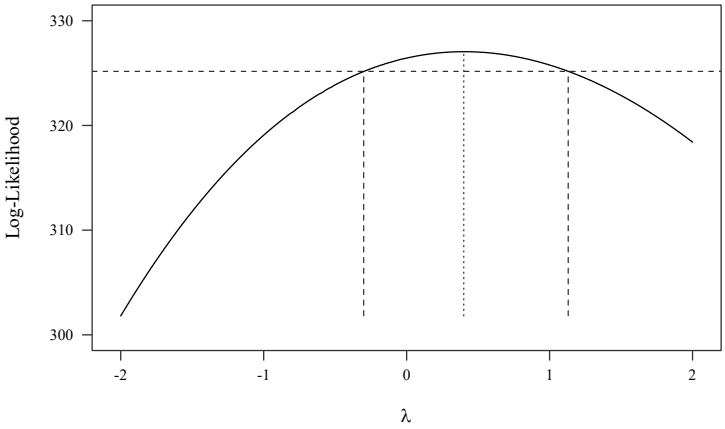

Figure 4.3: Box-Cox transformation applied to the series of monthly hospitalizations due to diabetes from 2010 to 2016. Reference lines denote the maximum likelihood estimate (dotted line) and the 95% CI for λ (dashed lines).

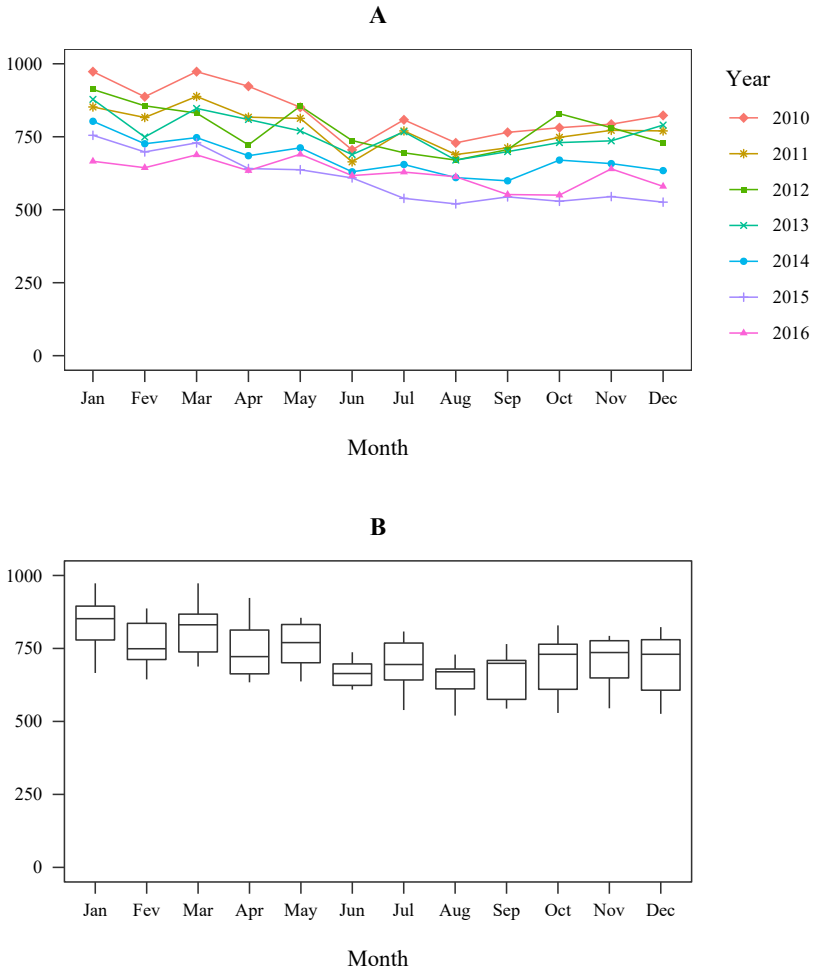

Figure 4.4: Distribution of hospitalizations by month from 2010 to 2016. (A) Evolution of monthly hospitalizations per year, and (B) distribution of monthly hospitalizations.

In view of the above, that is, the presence of trend and seasonality, as suggested by visual inspection of the data, first differences were taken, in an attempt to stabilize the mean (Figure 4.5).

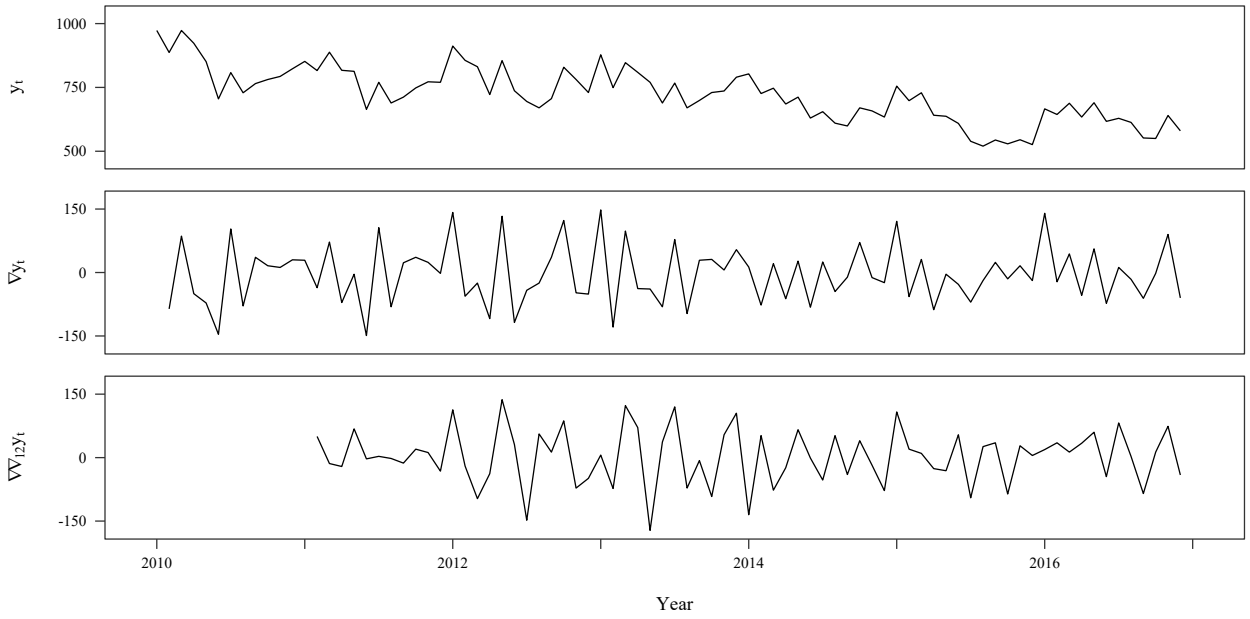

Figure 4.5: Series of monthly hospitalizations due to diabetes from 2010 to 2016. Original (y_t), monthly differenced (∇y_t), and monthly and yearly differenced ($\nabla\nabla_{12} y_t$) series are displayed.

The first plot represents the original (training) series of hospitalizations due to diabetes, y_t. As stated before, the mean does not seem to be constant. Differencing this series, ∇y_t, makes the data apparently stationary (second plot). Finally, considering both seasonal and regular differences, $\nabla\nabla_{12} y_t$, the resulting series also appears to be stationary. Further analysis of the data can help to decide if the series y_t is, indeed, nonstationary and some (non-seasonal and/or seasonal) differencing is needed.

Model identification

Firstly, to identify the appropriate model for the series of monthly hospitalizations due to diabetes, that is, to explore possible orders of the ARIMA model, the correlogram for the ACF and the PACF of the three series (y_t, ∇y_t, and $\nabla\nabla_{12} y_t$) were analysed. Given the hypothesis of seasonality, the first 30 lags were examined, thus allowing the analysis of two periods of 12 months (Figure 4.6).

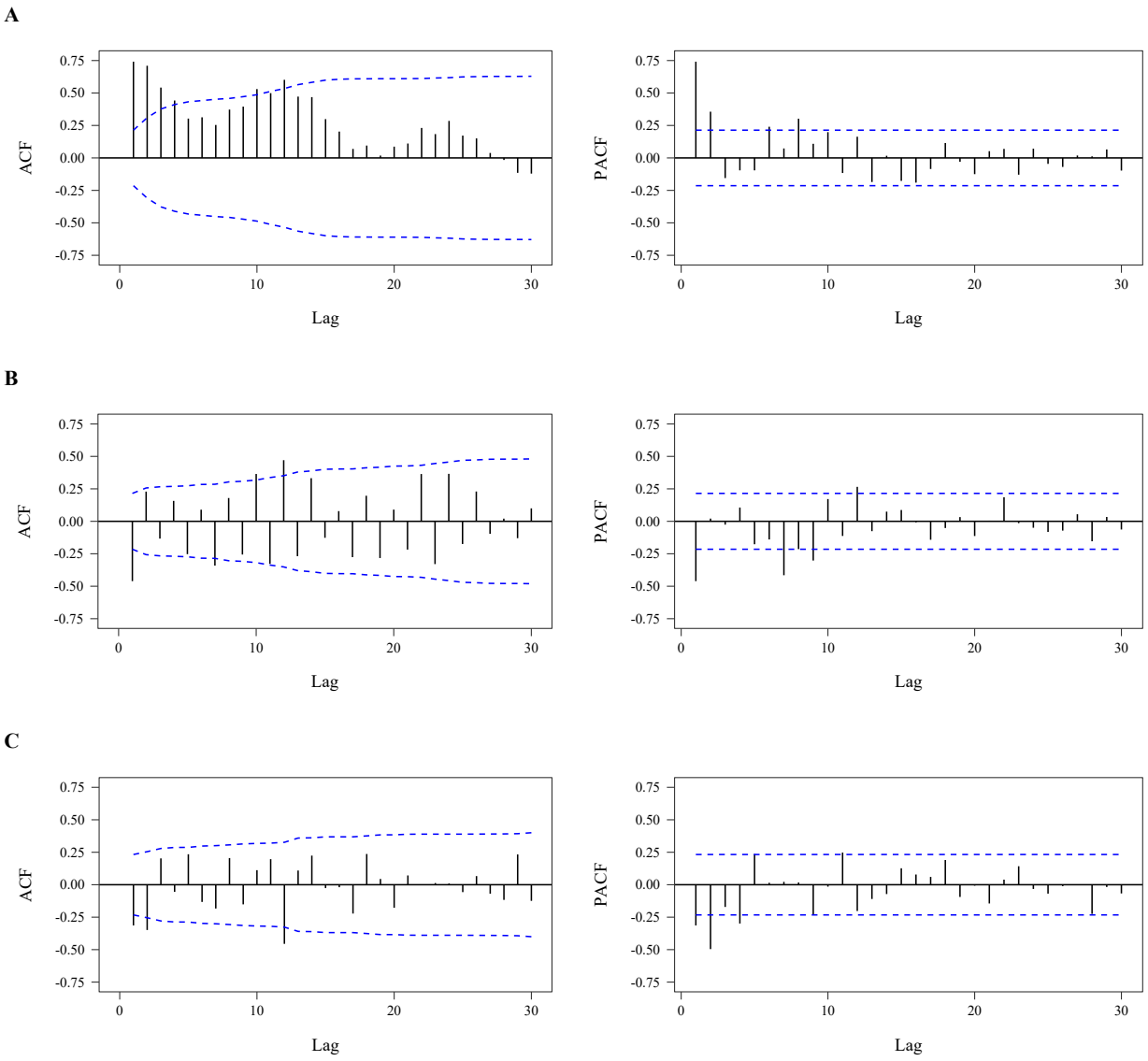

Figure 4.6: Correlogram for the ACF and the PACF for the (A) original, (B) monthly differenced and (C) monthly and yearly differenced series. The dashed blue lines represent the approximate 95% CI. ACF, Autocorrelation Function; PACF, Partial Autocorrelation Function.

Based on seasonal and non-seasonal components of both functions (Figure 4.6), the following interpretations can be made:

A. Original series, undifferenced (y_t): the ACF tails off, whereas the PACF seems to cut off at lag 2, even though there are significant autocorrelation values at higher lags in both functions. Still, the fact that the autocorrelations die out slowly and that the series itself shows a trend indicate that differencing is needed to make the process stationary. To verify this hypothesis, a formal unit root test was conducted. At first, an autoregressive approximation with order equal to 13 was identified

based on the AIC criterion. ADF tests for the model with a constant, and for the model with a trend plus a constant had non-significant results ($p = 0.876$ and $p = 0.750$, respectively). As such, the null hypothesis of a unit root (nonstationarity) was not rejected, supporting the need for differencing.

B. Series differenced with respect to months only (∇y_t): both the ACF and the PACF seem to cut off at lag 1, though significant values are observed at higher lags. These occur around lag 12, suggesting the presence of a seasonal unit root.

C. Series differenced with respect to months and years ($\nabla\nabla_{12} y_t$): autocorrelation values are highly reduced, cutting off after lag 2, whereas the partial autocorrelations tail off. An alternative would be to consider that the PACF cuts off at one of the first lags. Considering the seasonal component of the series, autocorrelation values cut off at lag 12, with no significant partial autocorrelations to be noted.

Given the above-mentioned aspects, nine candidate models were identified, of which one applies to the original series, two apply to the series differenced with respect to months (∇y_t), and the other six apply to the series differenced by month and year ($\nabla\nabla_{12} y_t$). Such models, and respective equations, are listed in Table 4.5.

Parameter estimation

All of these models, identified through graphical analysis of autocorrelation and partial autocorrelation functions, get to be estimated by maximum likelihood. Table 4.6 summarizes the results of this process, presenting estimates for all parameters, as well as the value and significance of the test, and information criteria (AIC, AICc, and BIC) for each model. In-sample MAPE are also presented.

When comparing these models, selection criteria were considered, though the (minimum) value of AIC prevailed for the decision on the model to choose. Thus, the preferred model was the $(1,1,2) \times (0,1,1)_{12}$. The decision on this model was supported by AICc, but not BIC. If the latter criterion was used, the model $(0,1,1) \times (0,1,1)_{12}$ would be chosen instead. Based on the relative error, one would conclude that the most complex model, $(4,1,2) \times (0,1,1)_{12}$, is the one that best adjusts to the training data.

No outliers were detected in the series during the estimation of the selected model,

$$\nabla\nabla_{12} y_t = -0.537\, \nabla\nabla_{12} y_{t-1} + w_t + 0.080\, w_{t-1} - 0.554\, w_{t-2} - 0.667\, w_{t-12} - 0.053\, w_{t-13} + 0.370\, w_{t-14},$$

which must further comply with tests and visual inspection of residuals. Therefore, residual analysis were performed to verify if the chosen model is indeed suitable to the series of monthly hospitalizations due to diabetes.

Table 4.5: Equations of candidate models for the series of monthly hospitalizations due to diabetes.

(2,0,0)	$(1-\phi_1 B-\phi_2 B^2)\tilde{y}_t = w_t$ $\tilde{y}_t = \phi_1\tilde{y}_{t-1}+\phi_2\tilde{y}_{t-2}+w_t$
(0,1,1)×(1,0,1)$_{12}$	$(1-\Phi B^{12})\nabla y_t = (1-\theta B)(1-\Theta B^{12})w_t$ $\nabla y_t = \Phi\nabla y_{t-12}+w_t-\theta w_{t-1}-\Theta w_{t-12}+\theta\Theta w_{t-13}$
(1,1,1)×(1,0,1)$_{12}$	$(1-\phi B)(1-\Phi B^{12})\nabla y_t = (1-\theta B)(1-\Theta B^{12})w_t$ $\nabla y_t = \phi\nabla y_{t-1}+\Phi\nabla y_{t-12}+\phi\Phi\nabla y_{t-13}+w_t-\theta w_{t-1}-\Theta w_{t-12}+\theta\Theta w_{t-13}$
(0,1,1)×(0,1,1)$_{12}$	$\nabla\nabla_{12} y_t = (1-\theta B)(1-\Theta B^{12})w_t$ $\nabla\nabla_{12} y_t = w_t-\theta w_{t-1}-\Theta w_{t-12}+\theta\Theta w_{t-13}$
(0,1,2)×(0,1,1)$_{12}$	$\nabla\nabla_{12} y_t = (1-\theta_1 B-\theta_2 B^2)(1-\Theta B^{12})w_t$ $\nabla\nabla_{12} y_t = w_t-\theta_1 w_{t-1}-\theta_2 w_{t-2}-\Theta w_{t-12}+\theta_1\Theta w_{t-13}+\theta_2\Theta w_{t-14}$
(1,1,1)×(0,1,1)$_{12}$	$(1-\phi B)\nabla\nabla_{12} y_t = (1-\theta B)(1-\Theta B^{12})w_t$ $\nabla\nabla_{12} y_t = \phi\nabla\nabla_{12} y_{t-1}+w_t-\theta w_{t-1}-\Theta w_{t-12}+\theta\Theta w_{t-13}$
(1,1,2)×(0,1,1)$_{12}$	$(1-\phi B)\nabla\nabla_{12} y_t = (1-\theta_1 B-\theta_2 B^2)(1-\Theta B^{12})w_t$ $\nabla\nabla_{12} y_t = \phi\nabla\nabla_{12} y_{t-1}+w_t-\theta_1 w_{t-1}-\theta_2 w_{t-2}-\Theta w_{t-12}+\theta_1\Theta w_{t-13}+\theta_2\Theta w_{t-14}$
(2,1,2)×(0,1,1)$_{12}$	$(1-\phi_1 B-\phi_2 B^2)\nabla\nabla_{12} y_t = (1-\theta_1 B-\theta_2 B^2)(1-\Theta B^{12})w_t$ $\nabla\nabla_{12} y_t = \phi_1\nabla\nabla_{12} y_{t-1}+\phi_2\nabla\nabla_{12} y_{t-2}+w_t-\theta_1 w_{t-1}-\theta_2 w_{t-2}-\Theta w_{t-12}+\theta_1\Theta w_{t-13}+\theta_2\Theta w_{t-14}$
(4,1,2)×(0,1,1)$_{12}$	$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)\nabla\nabla_{12} y_t = (1-\theta_1 B-\theta_2 B^2)(1-\Theta B^{12})w_t$ $\nabla\nabla_{12} y_t = \phi_1\nabla\nabla_{12} y_{t-1}+\phi_2\nabla\nabla_{12} y_{t-2}+\phi_3\nabla\nabla_{12} y_{t-3}+\phi_4\nabla\nabla_{12} y_{t-4}+w_t-\theta_1 w_{t-1}-\theta_2 w_{t-2}$ $-\Theta w_{t-12}+\theta_1\Theta w_{t-13}+\theta_2\Theta w_{t-14}$

Table 4.6: Summary of candidate models for the series of monthly hospitalizations due to diabetes, including coefficient estimates, information criteria and in-sample error.

Model	Parameter	Estimate	SE	t-value	p-value	σ^2	AIC	AICc	BIC	MAPE
(2,0,0)						3,662	11.157	11.160	12.273	6.831
	Mean (μ)	733.668	56.474	12.991	< 0.001					
	AR 1 (Φ_1)	0.486	0.101	4.830	< 0.001					
	AR 2 (Θ_1)	0.411	0.102	4.014	< 0.001					
(0,1,1)×(1,0,1)$_{12}$						1,973	10.701	10.707	10.818	4.971
	MA 1 (θ_1)	0.583	0.108	5.400	< 0.001					
	SAR 1 (Φ_1)	0.973	0.038	25.945	< 0.001					
	SMA 1 (Θ_1)	0.728	0.181	4.018	< 0.001					
(1,1,1)×(1,0,1)$_{12}$						1,939	10.712	10.721	10.858	4.909
	AR 1 (ϕ_1)	0.244	0.220	1.110	0.271					
	MA 1 (θ_1)	0.773	0.169	4.588	< 0.001					
	SAR 1 (Φ_1)	0.975	0.037	26.286	< 0.001					
	SMA 1 (Θ_1)	0.739	0.186	3.983	< 0.001					
(0,1,1)×(0,1,1)$_{12}$						2,014	10.660	10.665	10.756	4.458
	MA 1 (θ_1)	0.631	0.100	6.277	< 0.001					
	SMA 1 (Θ_1)	0.722	0.167	4.325	< 0.001					
(0,1,2)×(0,1,1)$_{12}$						1,976	10.669	10.677	10.796	4.465
	MA 1 (θ_1)	0.498	0.147	3.381	0.001					
	MA 2 (θ_2)	0.196	0.168	1.166	0.248					
	SMA 1 (Θ_1)	0.718	0.164	4.390	< 0.001					
(1,1,1)×(0,1,1)$_{12}$						1,984	10.678	10.687	10.806	4.435
	AR 1 (ϕ_1)	0.161	0.193	0.833	0.408					
	MA 1 (θ_1)	0.729	0.143	5.110	< 0.001					
	SMA 1 (Θ_1)	0.731	0.171	4.288	< 0.001					
(1,1,2)×(0,1,1)$_{12}$						1,918	10.647	10.660	10.807	4.463
	AR 1 (ϕ_1)	-0.537	0.188	-2.861	0.006					
	MA 1 (θ_1)	-0.080	0.167	-0.476	0.635					
	MA 2 (θ_2)	0.554	0.103	5.389	< 0.001					
	SMA 1 (Θ_1)	0.667	0.152	4.404	< 0.001					
(2,1,2)×(0,1,1)$_{12}$						1,909	10.660	10.678	10.851	4.452
	AR 1 (ϕ_1)	-0.623	0.209	-2.974	0.004					
	AR 2 (ϕ_2)	-0.216	0.188	-1.146	0.256					
	MA 1 (θ_1)	-0.136	0.199	-0.682	0.498					
	MA 2 (θ_2)	0.401	0.181	2.215	0.030					
	SMA 1 (Θ_1)	0.639	0.147	4.361	< 0.001					
(4,1,2)×(0,1,1)$_{12}$						1,766	10.685	10.718	10.940	4.014
	AR 1 (ϕ_1)	-0.250	0.120	-2.085	0.041					
	AR 2 (ϕ_2)	-1.274	0.107	-11.909	< 0.001					
	AR 3 (ϕ_3)	-0.416	0.101	-4.116	< 0.001					
	AR 4 (ϕ_4)	-0.459	0.118	-3.887	< 0.001					
	MA 1 (θ_1)	0.269	0.069	3.891	< 0.001					
	MA 2 (θ_2)	-1.000	0.158	-6.311	< 0.001					
	SMA 1 (Θ_1)	0.617	0.147	4.209	< 0.001					

SE, Standard Error; AIC, Akaike´s Information Criterion; AICc, Akaike´s Information Criterion corrected; BIC, Bayesian Information Criterion; MAPE, Mean Absolute Percentage Error; MA, Moving Average; SAR, Seasonal Autoregressive; SMA, Seasonal Moving Average, AR, Autoregressive.

Model diagnostic

In Figure 4.7, the time plot of standardized residuals does not reveal any abnormal patterns. Similarly, the normalized cumulative periodogram of the residuals does not show departures from linearity (the 0.05 probability limits were not crossed) which would be a sign of periodic nonrandomness. Considering either the histogram (of residuals) or the normal Q-Q plot (of standardized residuals), the normality of the residuals seems acceptable, without apparent outliers. By inspecting the ACF plot one can check that none of the autocorrelations are significant, suggesting the independence of the residuals. Finally, the Ljung-Box statistic was used to test the hypothesis of model adequacy. Up to the first 40 autocorrelations, the tests are all out the 0.05 level, thus reinforcing the evidence that the selected model is a good fit for the data.

As a whole, the results presented above support the model $(1, 1, 2) \times (0, 1, 1)_{12}$ as the most appropriated forecasting model for the series of monthly hospitalizations due to diabetes.

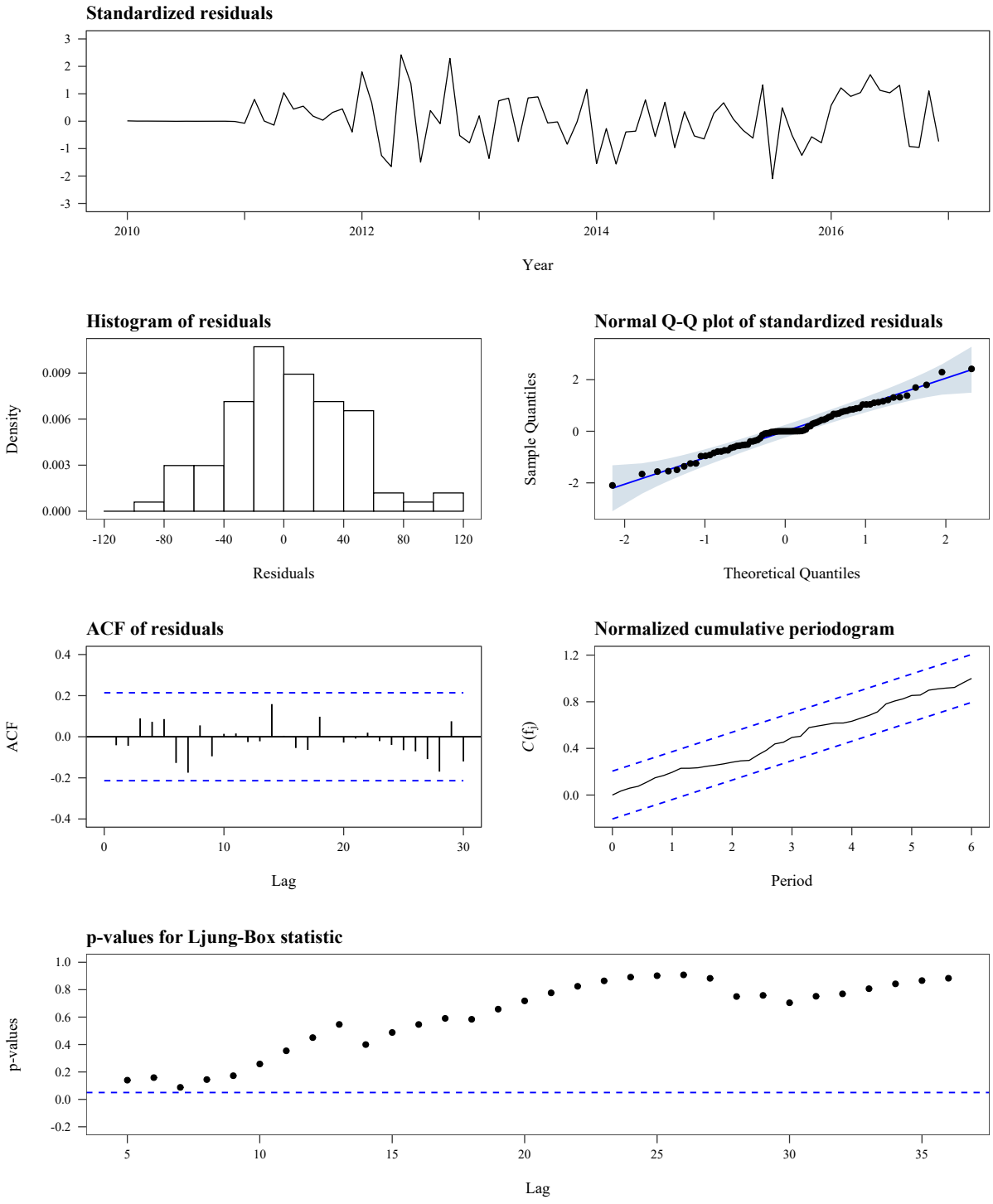

Figure 4.7: Graphical check of the residuals for the model $(1,1,2)\times(0,1,1)_{12}$. In the Q-Q plot (blue box), the ACF and the periodogram (blue dashed lines) there are represented 95% CI. In the plot of p-values for the Ljung-Box statistic, the blue dashed line represents the 0.05 level. Q-Q, Quantile-Quantile; ACF, Autocorrelation Function.

4.3 Forecasting

The ultimate test to model adequacy would be its ability to forecast, so the performance of the selected model was evaluated in terms of out-of-sample errors, by using the 24 months from the test set. This evaluation was made by lead time (1, 3, 6 and 12-months), but also by calendar year, across different lead times, considering either the recalibration of the model or just its update as soon as new observations become available and integrate the set of training data.

Table 4.7 presents accuracy measures, namely MAE, RMSE and MAPE, obtained using the SARIMA model $(1,1,2)\times(0,1,1)_{12}$ to predict hospitalizations due to diabetes for the selected forecast horizons. New observations, starting in January 2017, were included in the training set, one month at a time, and used along with previous ones in the prediction of the number of future hospitalizations. For the rolling-origin-update evaluation, the minimum error was obtained for three-month-ahead forecasts, with a MAE of 40.4, corresponding to a MAPE of 8.3%. The relative error remained under 10% until a 6-month forecast horizon, having reached a maximum of 16% at 12 months. When a rolling-origin-recalibration evaluation was conducted, the lowest MAPE was obtained with one-month-ahead forecasts, regardless of whether a fixed or a rolling window (84 months) is used to re-estimate the model (8.2% and 7.8%, respectively). In both cases, as the forecast timespan increased, up to 3, 6 and 12-months, the predictive accuracy of the model worsened. Forecasts obtained by using a rolling window were more accurate when made up to five months in advance, but not for a lead time of six or more months. For the latter, fixed window forecasts resulted in lower errors. Overall, the model $(1,1,2)\times(0,1,1)_{12}$ performed well, with a MAPE lower than 10% when forecasting up to six months, regardless of the model to be updated or recalibrated (detailed results, for a forecast horizon of 1 to 12 months, can be seen in Appendix B, Table B.1).

Table 4.7: Forecasting accuracy of the model (1,1,2)×(0,1,1)$_{12}$ for different lead times, based on rolling-origin-update and rolling-origin-recalibration evaluation.

	Rolling-origin-update			**Rolling-origin-recalibration**					
	Fixed window			*Fixed window*			*Rolling window*		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
$l=1$	44.9	54.8	9.1	41.1	48.7	8.2	39.5	47.4	7.8
$l=3$	40.4	49.1	8.3	41.3	48.4	8.4	40.8	47.8	8.3
$l=6$	45.0	56.8	9.8	44.0	55.9	9.5	44.9	56.4	9.7
$l=12$	69.1	81.4	16.0	68.2	79.8	15.8	70.0	81.5	16.2

l, lead time; MAE, Mean Absolute Error; RMSE, Root Mean Square Error; MAPE, Mean Absolute Percentage Error.

While these forecasts were obtained following one month increments, in practice observations may not be available with such frequency. It may also be true that such a frequent update of the training data set adds only a small improvement in terms of forecast accuracy. So, Table 4.8 presents the MAPE for the years 2017 and 2018 considering the re-estimation of the model every 1, 3, 6 and 12-months, that is, 12, 4, 2 and 1 times per year, respectively. At the same time, the performance of a seasonal random walk was investigated for

$l = 12$, with forecasts equal to the last value from the same season, that is, to the observed value in the same month from the previous year. This model was defined as benchmark method and compared against the selected SARIMA model.

Table 4.8: MAPE for the years 2017 and 2018 for SARIMA and Benchmark models. The model SARIMA(1,1,2)×(0,1,1)$_{12}$ was re-estimated every 1, 3, 6 and 12-months, that is, 12, 4, 2 and 1-time per year, given a rolling window of 84 months. The benchmark model (Seasonal Random Walk) was updated once each year, with forecasts being obtained for the next 12 months.

	SARIMA				**Benchmark**
	1 month	3 months	6 months	12 months	12 months
2017	7.5	7.2	6.1	5.7	10.3
2018	8.1	8.3	11.0	19.1	25.1
Average	7.8	7.7	8.6	12.4	17.7

SARIMA, Seasonal Autoregressive Integrated Moving Average.

When the SARIMA model with rolling window was used as forecasting method, the average MAPE for 2017 and 2018 ranged from 7.7% to 12.4%, for the forecasts obtained every 3 and 12 months, respectively. Different results were found for each year independently. In 2017, it is to note the decreasing error from the smallest (l = 1) to the largest (l = 12) forecast horizon, whereas in 2018 the values of MAPE increase as the lead time goes from 1 to 12 months. In short, regardless of how often the model is re-estimated, the errors obtained in 2017 were all lower than those in 2018. The same applies to the benchmark model, from which resulted an average MAPE of 17.7%, for l = 12. At this forecast horizon, the comparison between the two models shows that the SARIMA performed better, with a reduction of 30% in the MAPE in relation to the benchmark.

In Figure 4.9 are represented the forecasts and respective 95% limits from the SARIMA and the benchmark models. For the former, the projections are mostly above the series, especially in 2018, as one can see in Figure 4.8, where negative values of the error, $e_t(l) = y_{t+l} - \hat{y}_t(l)$, indicate an overestimation of the series. The number of forecasts higher than the observed values increased with the forecast horizon, in such a way that when forecasts are made for the next 12 months all estimated values for 2018 are higher than those observed. Despite, all the observed values, except those for February 2017 and May 2018, are in the 95% prediction interval for the forecasts. For the benchmark model, wider limits were observed. Still, there are four forecasts outside the prediction interval, all in 2018 (May, July, September, December). In this case, all the predictions, with the exception of January 2017, are higher than the observations, reflecting the decreasing trend of the series.

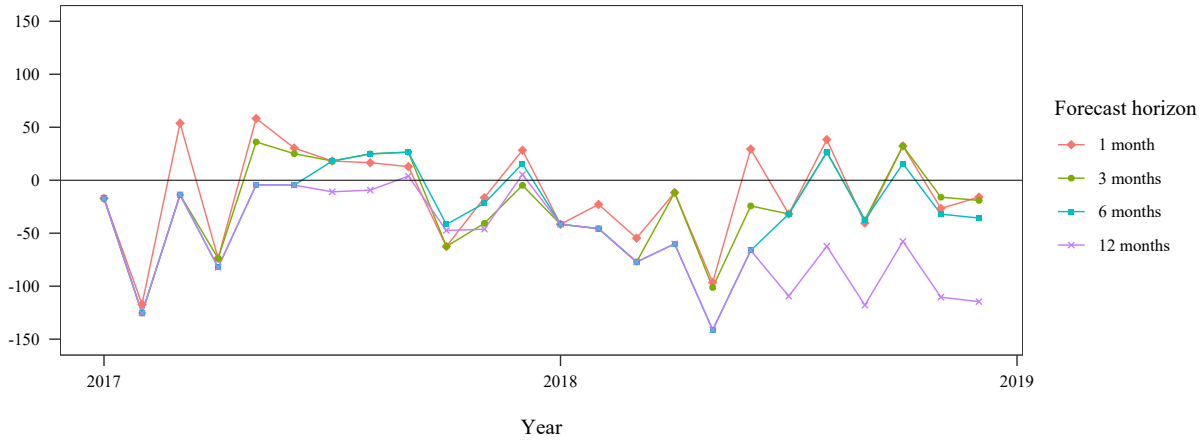

Figure 4.8: Forecast errors of the model $(1,1,2)\times(0,1,1)_{12}$ in the test set when the model was re-estimated at every 1, 3, 6 or 12-months. The error takes negative values when the forecast overestimates the observed series and positive values when the opposite occurs.

In addition, the selected model has proved able to predict turning points as, for more than half of the evaluation period, upward and downward changes in forecasted values are in agreement with changes in observed values of the series. In this respect, model recalibration at every month is the method with the lower number of variations coincident with the series (14 out of 24). On the other hand, 3 and 6-month forecasts reveal as the most accurate (17 out of 24). Regardless of the forecast horizon, discrepancies in turning points between predicted and observed values are more pronounced in the second half of each year.

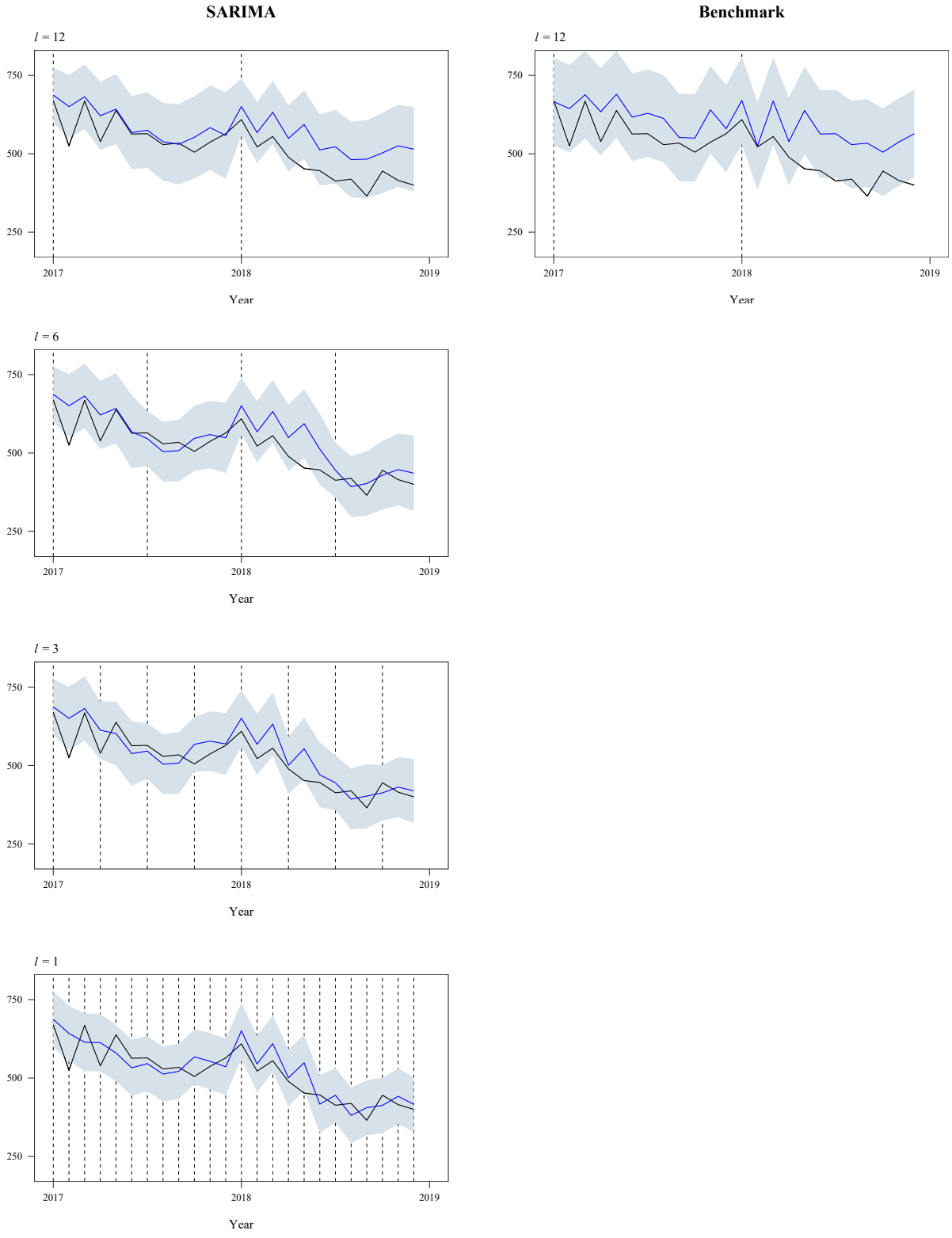

Figure 4.9: Forecasts of monthly hospitalizations due to diabetes for 2017 and 2018. Observed values (black line) and forecasts (blue line; blue box = 95% prediction interval) for SARIMA(1,1,2)×(0,1,1)$_{12}$ and Benchmark (Seasonal Random Walk) models. Vertical lines indicate time points at which the model was updated (Benchmark) or re-estimated (SARIMA). l, lead time.

Chapter 5

Discussion

As one of the major types of non-communicable diseases, along with CVD, cancer and chronic respiratory diseases, diabetes represents a tremendous challenge for health systems [46]. In Portugal, the National Programme for the Prevention and Control of Diabetes from the Ministry of Health, presented in 2008, aimed to reduce diabetes prevalence, morbidity and mortality, by reducing the incidence of complications, or delaying their onset. It intended, among others, to reduce the number of episodes of hospitalization due to diabetes complications, and, more specifically, hospitalizations due to ketoacidosis, severe hypoglycaemia and hyperosmolarity [47].

The present work focus precisely on the consequences of diabetes, aiming to describe and model the temporal evolution of hospitalizations due to diabetes in Portugal, in a total of 73,050 episodes from 2010 to 2018. Over this period, the series of monthly hospitalizations exhibits a decreasing trend, with apparent seasonality. A higher number of hospitalizations was recorded in winter, with the annual peak occurring most frequently in January, whereas summer months accounted for the fewest number of hospitalizations due to diabetes during the study period. The same was observed by Gomes, Fonseca, and Freitas [48], while studying the seasonal variation of hospitalizations with primary diagnosis of T2D with hyperosmolarity, in mainland Portugal. The study of the factors explaining the seasonal variability observed in the data is out of the scope of this thesis, but it is worth mentioning that this pattern is in line with seasonal fluctuation in the levels of blood glucose [49, 50] and HbA1c [50–52], with the peak observed in the first months of the year reflecting, quite possibly, the excesses committed during Christmas festivities and consequent weight gain [53, 54], along with a diminished practice of physical activity [55].

Box-Jenkins approach was further applied to the data in order to identify the SARIMA model that best fitted the series of monthly hospitalizations due to diabetes. Nine candidate models were trained using data from January 2010 to December 2016 (84 months), of which the one with the lowest AIC, $(1,1,2) \times (0,1,1)_{12}$, was selected and used to forecast hospitalizations up to 12 months ahead over a period of 24 months, corre-

spondent to the years 2017 and 2018. This model was firstly subjected to rolling-origin-update and rolling-origin-recalibration evaluations by successively incrementing the training set by one month. By doing that, one aimed to verify if it would be worthwhile to re-estimate the model instead of just updating it, while the accuracy of the model was assessed by averaging forecast errors for each lead time.

Following the recalibration procedure, fixed and rolling windows were further used to investigate if there was any benefit in keeping the oldest observations or, on the contrary, more accurate predictions would be obtained if these were dropped when new ones become available. The obtained results suggest that, for longer horizons, the model takes advantage on the use of more data, given a fixed window, to produce more accurate forecasts. Still, regardless of whether a fixed or a rolling window was used, the predictive accuracy of the model worsened as the forecast timespan increased, with the minimum MAPE being obtained for one-month-ahead forecasts (8.2% and 7.8%, respectively). This was expected in all forecasting scenarios given the propagation of errors, that the model cannot account for, across the time window. Nonetheless, for the method of rolling-origin-update, one-month forecasts did not present the lowest error. This can possibly be justified by the constant imputation of new data, considerably different from previous, that the model was unable to properly accommodate as it was not re-estimated, thus reflecting in poorer forecasts. Overall, rolling-origin-recalibration performed better, but for both methods the SARIMA model was able to forecast hospitalizations due to diabetes with good accuracy up to six months in advance, that is, with a relative error lower than 10%. Similar results were obtained by Villani *et al.* [56]. The authors verify that a SARIMA model can accurately forecast monthly prehospital caseload of acute diabetic emergencies, namely hypoglycaemia and hyperglycaemia.

The performance of the model was also evaluated across lead times, while forecasting the number of hospitalizations due to diabetes for 2017 and 2018. Different schemes, in which new values were available every month, or only at every 3, 6 or 12 months, were considered. After forecasting for the first horizon, new data were included in the training set and used to re-estimate the model before the next horizon get to be predicted, simulating real forecasting scenario, where data are imputed to the model as soon as they become available.

For the year 2017, MAPE was lower than 10% in all cases, with the particularity of decreasing from the shortest to the longest forecast horizon. It would be expected that the recalibration of the model with the greatest frequency improved its accuracy, but a pernicious effect was observed instead. For instance, the series behaviour at the beginning of 2017 was quite different from the pattern observed in 2016. As an example, the number of hospitalizations in February 2017, month from a season typically associated with higher values, was lower than the lowest number of monthly hospitalization in 2016, observed in October. So, at each recalibration time point, the model was imputed with data that did not agreed with previous data from which the model was identified. Such disagreement was reflected in the forecasts computed, being more notorious with the greater frequency of re-estimation. Facing this results, it could be questioned if the length of the training set had influenced forecasting accuracy. As referred by Schweigler *et al.* [57], if it is too short, parameter estimates can be imprecise. In the other hand, if too long, the model could not be

able to adapt to recent behaviour and new patterns. In the present work, the two years of the validation set correspond to 22% of the available data, allowing to test the model over two periods of 12 months and so evaluate how it captures the seasonality of the series. Without a formal rule about the dimension of the test set, the percentage of data retained approximated to the value of 20% can be regarded as acceptable [44].

As for 2018, in turn, model's accuracy improved when it got to be re-estimated more often, as expected. Notwithstanding, the magnitude of the error largely increased when forecasts for 12 months were obtained. Happens, possibly, that the use of information from an unusual year, combined with the behaviour, sharply decreasing, of the series in the year to be forecast resulted in poor predictions. In average, a MAPE of 12.4% was obtained for that horizon, representing an improvement (30% reduction, even more if the model was re-estimated more than once a year) in relation to the benchmark method, a seasonal random walk. In this case, the model needs to be updated just once a year, since each forecast value equals the observed value for the same month in the previous year [58]. The fact that it is a simple method that requires no parameter estimation, and so can be easily implemented in health services by non-statistician, justifies its choice as benchmark model. The graphical representation of the forecasts shows that this method overestimates the series, in a more notorious way than the SARIMA. For the latter, this was specially evident in the first months of either 2017, when the series took extreme low values — that would be to expect in warmer months according to the past —, and 2018, as the model did not foresee the abrupt decline of the series. Despite, this model was able to capture the seasonality of the series, predicting a higher number of hospitalizations in coldest months, while ensuring a good forecasting accuracy when re-estimated at least twice a year. Hence, if monthly update is not viable in clinical practice, quarterly or semi-annual recalibration would stand as a good alternative.

These findings support the use of SARIMA models to forecast hospitalizations due to diabetes at short/medium term, allowing management decisions to be taken timely. Being alerted for high demand periods, health managers can plan according to patient flow, thus improving quality of care, while making a more efficient use of the budget, as the allocation of resources becomes closer to real needs [44, 59, 60].

As a general class of linear models, Box-Jenkins (S)ARIMA models are capable of modelling most time series and have been widely used in health forecasting, in the prediction of admissions to the Emergency Department [44, 57, 60–66], new admission / discharged inpatients [67, 68], hospital daily outpatient visits [69], patient volume in Hospital Medicine [70], patient volume at a primary health care clinic [71], surgical case volume [72], length of stay, discharge and readmission rate [73], prehospital acute diabetic emergencies [56], demand for red blood cell transfusion [74], incidence and mortality rate for prostate cancer [75]. This family of models is often compared against other forecasting methods, including the exponential smoothing, equally popular [44, 56, 60, 64, 69, 70, 73, 74, 76–78], and neural networks [64, 68, 74, 79].

Time series models, despite being less informative than regression models, can provide greater forecast accuracy. Nonetheless, it is not guaranteed that time series models produce the best results in all contexts, as

evidenced by Ordu, Demir, and Tofallis [76]. The authors developed, in collaboration with finance, strategy and planning directors, a forecasting modelling framework for all acute services of a hospital in England, including all the specialities in outpatient, inpatient and emergency and accident departments, having conclude that the best predictions arise from different methods and horizons (daily, weekly and monthly forecasts). Their findings support the importance of exploring several options when selecting a forecasting model.

The present work did not seek to make a formal comparison between forecasting methods, but rather an exploratory analysis of different forecasting procedures, applied over data on hospitalizations due to diabetes. Soyiri and Reidpath [59] refer that such condition-specific forecasts can better prepare health care providers compared to aggregated forecasts, as would be predictions for overall hospitalizations in this particular case. Gershon *et al.* [66] and Becerra *et al.* [44], for example, based their work on respiratory diseases. Olsavszky *et al.* [80] also focused on specific conditions, having predicted hospitalizations for each one of the top 10 causes of death, including diabetes and respiratory diseases, while testing different modelling approaches through automated time series machine learning. In common there is the application of forecasting techniques over regional data. Other authors, like Jones *et al.* [64] and Schweigler *et al.* [57], used data from different health facilities, independently. Such a strategy would be welcomed by health managers, given clear differences between hospitals in their organizational structure, as well as in the social, environmental and climatic context that they make part of [70]. Of note, in this regard, that a greater forecast error would be to expect at this level since, as denoted by Jenkins [81], the higher the level of disaggregation, the higher the randomness and, therefore, the unpredictability of the series. So, it is possible that hospitals in regions with more pronounced seasonal effects, as, for example, more extreme winters, would benefit the most from the use of forecasting models.

While the awareness for seasonal patterns is in itself advantageous when facing periods of higher demand of medical services, it could also be useful to separate elective from non-elective cases when building the forecasting model. According to Zinouri, Taaffe, and Neyens [72], this could be a way to balance workload by scheduling non-emergent cases in periods (days or months) associated to a lower patient flow. This was, in fact, made by Ordu, Demir, and Tofallis [76], which modelled separately elective and non-elective admissions, but also first referrals and follow-ups. In the present work, the number of emergencies far exceeded that of planned admissions such that one can question whether planning and coordination strategies from health care providers could benefit, to some degree, from this distinction in data used to predict hospitalizations due to diabetes.

There are acknowledged limitations in the work conducted, concerning both the nature of the data and the statistical methods applied. Firstly, only main diagnoses of diabetes and the most common types of the disease were included, thus underestimating its impact, as well as the demand for diabetes-specific health care. Moreover, the use of aggregated data at national level makes unviable the use of the obtained forecasts in real clinical practice. Concerning the modelling process, the use of AIC to decide on the forecasting model to choose, while reflecting a compromise between goodness of fit and complexity, may have led to

a misinterpretation of the quality of the models since different data sets (differenced/undifferenced data) were used to train them. Another approach, while computationally more expensive, would be to select all formulated models, as done by Villani *et al.* [56], or at least some of them with lower AIC, and choose the one with the greatest predictive ability based on MAPE. Other authors, like Earnest *et al.* [75], based their decision on in-sample fit. Yet, the model with the lowest error in the training data is not guaranteed to have the best forecasting accuracy. Finally, it should be noted that model evaluation was based on provisional data. Still, and although it can not be assured that the forecast errors would be as diminutive if definitive data were used, this work shows that SARIMA models are capable to predict diabetes-related conditions with good accuracy.

Future work on this field of research includes a spatiotemporal analysis of hospitalizations due to diabetes in Portugal, following a Bayesian hierarchical disease mapping approach. In this process, population socioeconomic characteristics and access to health care at regional level will be taken into account, thus supporting the implementation of community-based interventions intended to reduce diabetes complications and the consequent need for medical care at the hospital.

Chapter 6

Conclusion

Diabetes is a chronic, complex disease, posing a considerable burden on individuals, health systems, and society in general. Facing the increasing prevalence of this disease and its consequences, this work aimed to describe and model the temporal evolution of hospitalizations due to diabetes in Portugal from 2010 to 2018. On the first point, it was shown that, despite the increasing trend in the admissions for diabetes over these years, the number of episodes requiring hospitalization revealed a decreasing pattern, which can be interpreted optimistically as a result of national policies on diabetes prevention and control, intended to reduce disease morbidity and mortality. Seasonal fluctuations were also observed in the series of monthly hospitalizations for diabetes-related complications, highlighting a greater number of cases in the winter months. Although the recognition of these patterns can be helpful in high demand periods, reliable and accurate forecasts could better guide planning and organization by health care providers, thus addressing patient needs in a more effective way. In this regard, the validity and accuracy of SARIMA models in forecasting monthly hospitalizations due to diabetes were assessed following Box-Jenkins approach and out-of-sample evaluation. In summary, the results of this study reveal that SARIMA models, central in the field of time series analysis, perform well in this context, suggesting that they can be used to predict hospitalizations far enough to allow for an adequate allocation of medical resources with good accuracy. Whether these forecasts would be useful in clinical settings, supporting decision-making, should be further investigated.

References

[1] Classification of diabetes mellitus. Geneva: World Health Organization, 2019.

[2] IDF Diabetes Atlas. 9th ed. International Diabetes Federation, 2019.

[3] Dimeglio LA, Evans-molina C, Oram RA. Type 1 diabetes. Lancet 2018; 391(10138):2449–62.

[4] Skyler JS *et al.* Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes. 2017; 66(2):241–55.

[5] Zeytinoglu M, Huang ES. Diabetes: A Primary Health Care Approach. In: Andrade JP, Pinto FJ, Arnett DK, eds. Prevention of Cardiovascular Diseases - From Current Evidence to Clinical Practice. Cham, Switzerland: Springer, 2015. p. 91-9.

[6] Rewers M, Ludvigsson J. Environmental risk factors for type 1 diabetes. Lancet. 2016; 387(10035):2340–48.

[7] Zaccardi F, Webb DR, Yates T, Davies MJ. Pathophysiology of type 1 and type 2 diabetes mellitus: A 90-year perspective. Postgrad. Med. J. 2016; 92(1084):63–9.

[8] Moini J. Type 2 diabetes. In: Moini J, eds. Epidemiology of Diabetes. Amsterdam, Netherlands: Elsevier, 2019. Chapter 7. Available from: https://doi.org/10.1016/B978-0-12-816864-6.00007-9 [Accessed on: 2020 Mar 31].

[9] Zhou T, Liu X, Liu Y, Li X. Meta-analytic evaluation for the spatio-temporal patterns of the associations between common risk factors and type 2 diabetes in mainland China. Medicine. 2019; 98(20):e15581.

[10] Fallahzadeh H, Ostovarfar M, Lotfi MH. Population attributable risk of risk factors for type 2 diabetes; Bayesian methods. Diabetes Metab. Syndr. Clin. Res. Rev. 2019; 13(2):1365–8.

[11] Barreto M *et al.* Prevalence, awareness, treatment and control of diabetes in Portugal: Results from the first National Health examination Survey (INSEF 2015). Diabetes Res. Clin. Pract. 2018; 140:271–8.

[12] Chiang JI *et al.* Associations between multimorbidity, all-cause mortality and glycaemia in people with type 2 diabetes: A systematic review. PLoS One. 2018; 13(12):1–19.

[13] Lake A, Arthur A, Byrne C, Davenport K, Yamamoto JM, Murphy HR. The effect of hypoglycaemia during hospital admission on health-related outcomes for people with diabetes: a systematic review and meta-analysis. Diabet. Med. 2019; 36(11):1349–59.

[14] Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet. 2017; 389(10085):2239–51.

[15] Moini J. Type 1 diabetes. In: Moini J, eds. Epidemiology of Diabetes. Amsterdam, Netherlands: Elsevier, 2019. Chapter 6. Available from: https://doi.org/10.1016/B978-0-12-816864-6.00007-9 [Accessed on: 2020 Mar 31].

[16] Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet. 2010; 375(9733):2215–22.

[17] Raghavan S *et al.* Diabetes mellitus–related all-cause and cardiovascular mortality in a national cohort of adults. J. Am. Heart Assoc. 2019; 8(4).

[18] Lee YB *et al.* Risk of early mortality and cardiovascular disease in type 1 diabetes: A comparison with type 2 diabetes, a nationwide study. Cardiovasc. Diabetol. 2019; 18(1):1–17.

[19] Ministério da Saúde. Direção-Geral da Saúde. Programa Nacional para a Diabetes 2019. Desafios e Estratégias. Lisboa: Direção-Geral da Saúde, 2019.

[20] Uusitupa M, Khan T, Viguiliouk E, Kahleova H. Prevention of Type 2 Diabetes by Lifestyle Changes: A Systematic Review and Meta-Analysis. Nutrients. 2019; 11(2611):1–22.

[21] Glechner A *et al.* Effects of lifestyle changes on adults with prediabetes: A systematic review and meta-analysis. Prim. Care Diabetes 2018; 12(5):393–408.

[22] Observatório da Diabetes. Diabetes Factos e Números - O Ano de 2015 - Relatório Anual do Observatório Nacional de Diabetes. Lisboa: Sociedade Portuguesa de Diabetologia, 2016.

[23] Świątoniowska N, Sarzyńska K, Szymańska-Chabowska A, Jankowska-Polańska B. The role of education in type 2 diabetes treatment. Diabetes Res. Clin. Pract. 2019; 151:237–46.

[24] Cruz-Cobo C, Santi-Cano MJ. Efficacy of Diabetes Education in Adults With Diabetes Mellitus Type 2 in Primary Care: A Systematic Review. J. Nurs. Scholarsh. 2020; 52(2):155–63.

[25] Rouyard T, Kent S, Baskerville R, Leal J, Gray A. Perceptions of risks for diabetes-related complications in Type 2 diabetes populations: a systematic review. Diabet. Med. 2017; 34(4):467–77.

[26] Koro CE, Bowlin SJ, Bourgeois N, Fedder DO. Glycemic control from 1988 to 2000 among U.S. adults diagnosed with type 2 diabetes: a preliminary report. Diabetes Care. 2004 Jan; 27(1):17–20.

[27] Nogueira ML. Custos com a Diabetes. 2015. Available from: http://www.spd.pt/index.php/custos-com-a-diabetes-mainmenu-105 [Accessed on: 2018 Dec 12].

[28] Fialho M, Antunes M. Prediction of the number of hospitalizations due to diabetes in Portugal: a time series analysis. J. Stat. Heal. Decis. 2020; 2(2):15–6. Available from: https://doi.org/10.34624/jshd.v2i2.21051.

[29] Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. 5th ed. Hoboken, New Jersey: Wiley, 2016.

[30] Alonso AM, García-Martos C. Time Series Analysis. Madrid, 2012.

[31] Cryer JD, Chan KS. Time Series Analysis With Applications in R. 2nd ed. New York: Springer, 2008.

[32] Quantcademy. Serial Correlation in Time Series Analysis. Available from: https://www.quantstart.com/articles/Serial-Correlation-in-Time-Series-Analysis/ [Accessed on: 2020 Apr 26].

[33] NIST/SEMATECH. Introduction to Time Series Analysis. 2020. Available from: http://www.itl.nist.gov/div898/handbook/ [Accessed on: 2020 Jul 11].

[34] Shumway RH, Stoffe DS. Time Series Analysis and Its Applications With R Examples. 4th ed. Springer, 2017.

[35] Dangor Z *et al.* Temporal association in hospitalizations for tuberculosis, invasive pneumococcal disease and influenza virus illness in South African children. PLoS One. 2014; 9(3):e91464.

[36] Ke G *et al.* Epidemiological analysis of hemorrhagic fever with renal syndrome in China with the seasonal-trend decomposition method and the exponential smoothing model. Sci. Rep. 2016 Dec; 6:39350.

[37] Rojo J, Rivero R, Romero-Morte J, Fernández-González F, Pérez-Badia R. Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing. Int. J. Biometeorol. 2017 Feb; 61(2):335–48.

[38] Polwiang S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017). BMC Infect. Dis. 2020 Mar; 20(1):208.

[39] Yeo IK, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000; 87(4):954–9.

[40] Zivot E. Estimation of ARMA Models. 2005.

[41] Tashman LJ. Out-of-sample tests of forecasting accuracy: An analysis and review. Int. J. Forecast. 2000; 16(4):437–50.

[42] Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf. Sci. 2012; 191:192–213.

[43] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int. J. Forecast. 2006; 22(4):679–88.

[44] Becerra M, Jerez A, Aballay B, Garcés HO, Fuentes A. Forecasting emergency admissions due to respiratory diseases in high variability scenarios using time series: A case study in Chile. Sci. Total Environ. 2020; 706:134978.

[45] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2020. Available from: https://www.r-project.org/.

[46] World Health Organization. Noncommunicable diseases. 2018. [Accessed on: 2020 Jun 27].

[47] Ministério da Saúde. Direcção-Geral da Saúde. Direcção de Serviços de Cuidados de Saúde. Programa Nacional de Prevenção e Controlo da Diabetes. Lisbon, 2008.

[48] Gomes C, Fonseca D, Freitas A. Seasonal variation of diabetes with hyperosmolarity hospitalizations and its characteristics in mainland Portugal. Prim. Care Diabetes. 2020; 14(5):445–7.

[49] Kershenbaum A, Kershenbaum A, Tarabeia J, Stein N, Lavi I, Rennert G. Unraveling seasonality in population averages: An examination of seasonal variation in glucose levels in diabetes patients using a large population-based data set. Chronobiol. Int. 2011; 28(4):352–60.

[50] Gikas A, Sotiropoulos A, Pastromas V, Papazafiropoulou A, Apostolou O, Pappas S. Seasonal variation in fasting glucose and HbA1c in patients with type 2 diabetes. Prim. Care Diabetes. 2009; 3(2):111–4.

[51] Kim YJ *et al.* Seasonal variation in hemoglobin a1c in Korean patients with type 2 diabetes mellitus. J. Korean Med. Sci. 2014; 29(4):550–5.

[52] Pinho Pereira MTR, Lira D, Bacelar C, Oliveira JC, De Carvalho AC. Seasonal variation of haemoglobin A1c in a Portuguese adult population. Arch. Endocrinol. Metab. 2015; 59(3):231–5.

[53] Zorbas C *et al.* The Relationship Between Feasting Periods and Weight Gain: a Systematic Scoping Review. Curr. Obes. Rep. 2020; 9(1):39–62.

[54] Turicchi J *et al.* Weekly, seasonal and holiday body weight fluctuation patterns among individuals engaged in a European multi-centre behavioural weight loss maintenance intervention. PLoS One. 2020; 15(4):1–19.

[55] Cepeda M *et al.* Seasonality of physical activity, sedentary behavior, and sleep in a middle-aged and elderly population: The Rotterdam study. Maturitas. 2018; 110:41–50.

[56] Villani M, Earnest A, Nanayakkara N, Smith K, De Courten B, Zoungas S. Time series modelling to forecast prehospital EMS demand for diabetic emergencies. BMC Health Serv. Res. 2017; 17(1):1–9.

[57] Schweigler LM, Desmond JS, McCarthy ML, Bukowski KJ, Ionides EL, Younger JG. Forecasting models of emergency department crowding. Acad. Emerg. Med. 2009; 16(4):301–8.

[58] Hyndman R, Athanasopoulos G. Forecasting: Principles and Practice. 2nd ed. Melbourne, Australia: OTexts, 2018. Available from: https://otexts.com/fpp2/ [Accessed on: 2020 Mar 10].

[59] Soyiri IN, Reidpath DD. An overview of health forecasting. Environ. Health Prev. Med. 2013; 18(1):1–9.

[60] Carvalho-Silva M, Monteiro MT, Sá-Soares F de, Dória-Nóbrega S. Assessment of forecasting models for patients arrival at Emergency Department. Oper. Res. Heal. Care. 2018; 18:112–8.

[61] Kadri F, Harrou F, Chaabane S, Tahon C. Time series modelling and forecasting of emergency department overcrowding. J. Med. Syst. 2014; 38(107):1–20.

[62] Juang WC, Huang SJ, Huang FD, Cheng PW, Wann SR. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. BMJ Open. 2017; 7(11):1–8.

[63] Marcilio I, Hajat S, Gouveia N. Forecasting daily emergency department visits using calendar variables and ambient temperature readings. Acad. Emerg. Med. 2013; 20(8):769–77.

[64] Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ, Snow GL. Forecasting daily patient volumes in the emergency department. Acad. Emerg. Med. 2008; 15(2):159–70.

[65] Sun Y, Heng BH, Seow YT, Seow E. Forecasting daily attendances at an emergency department to aid resource planning. BMC Emerg. Med. 2009; 9.

[66] Gershon A, Thiruchelvam D, Moineddin R, Zhao XY, Hwee J, To T. Forecasting hospitalization and emergency department visit rates for chronic obstructive pulmonary disease a time-series analysis. Ann. Am. Thorac. Soc. 2017; 14(6):867–73.

[67] Zhu T, Luo L, Zhang X, Shi Y, Shen W. Time-Series Approaches for Forecasting the Number of Hospital Daily Discharged Inpatients. IEEE J. Biomed. Heal. Informatics 2017; 21(2):515–26.

[68] Zhou L, Zhao P, Wu D, Cheng C, Huang H. Time series model for forecasting the number of new admission inpatients. BMC Med. Inform. Decis. Mak. 2018; 18(1):1–11.

[69] Luo L, Luo L, Zhang X, He X. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. BMC Health Serv. Res. 2017; 17(1):1–13.

[70] Kim K, Lee C, Leary KJO. Predicting Patient Volumes in Hospital Medicine: A Comparative Study of Different Time Series Forecasting Methods. 2014.

[71] Abdel-Aal RE, Mangoud AM. Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis. Comput. Methods Programs Biomed. 1998; 56(3):235–47.

[72] Zinouri N, Taaffe KM, Neyens DM. Modelling and forecasting daily surgical case volume using time series analysis. Heal. Syst. 2018; 7(2):111–9.

[73] Weiss TW, Ashton CM, Wray NP. Forecasting areawide hospital utilization: A comparison of five univariate time series techniques. Heal. Serv. Manag. Res. 1993; 6(3):178–90.

[74] Pereira A. Performance of time-series methods in forecasting the demand for red blood cell transfusion. Transfus. Pract. 2004; 44:739–46.

[75] Earnest A, Evans SM, Sampurno F, Millar J. Forecasting annual incidence and mortality rate for prostate cancer in Australia until 2022 using autoregressive integrated moving average (ARIMA) models. BMJ Open. 2019; 9(8):1–7.

[76] Ordu M, Demir E, Tofallis C. A comprehensive modelling framework to forecast the demand for all hospital services. Int. J. Health Plann. Manage. 2019; 34(2):e1257–71.

[77] Bergs J, Heerinckx P, Verelst S. Knowing what to expect, forecasting monthly emergency department visits: A time-series analysis. Int. Emerg. Nurs. 2014; 22(2):112–5.

[78] Jones SS *et al.* A multivariate time series approach to modeling and forecasting demand in the emergency department. J. Biomed. Inform. 2009; 42(1):123–39.

[79] Jilani T, Housley G, Figueredo G, Tang PS, Hatton J, Shaw D. Short and Long term predictions of Hospital emergency department attendances. Int. J. Med. Inform. 2019; 129:167–74.

[80] Olsavszky V, Dosius M, Benecke J, Vladescu C. Time series analysis and forecasting with automated machine learning on a national ICD-10 database. Int. J. Environ. Res. Public Health. 2020; 17(14):1–17.

[81] Jenkins GM. Some practical aspects of forecasting in organizations. J. Forecast. 1982; 1(1):3–21.

Appendices

A Diabetes hospitalizations by region

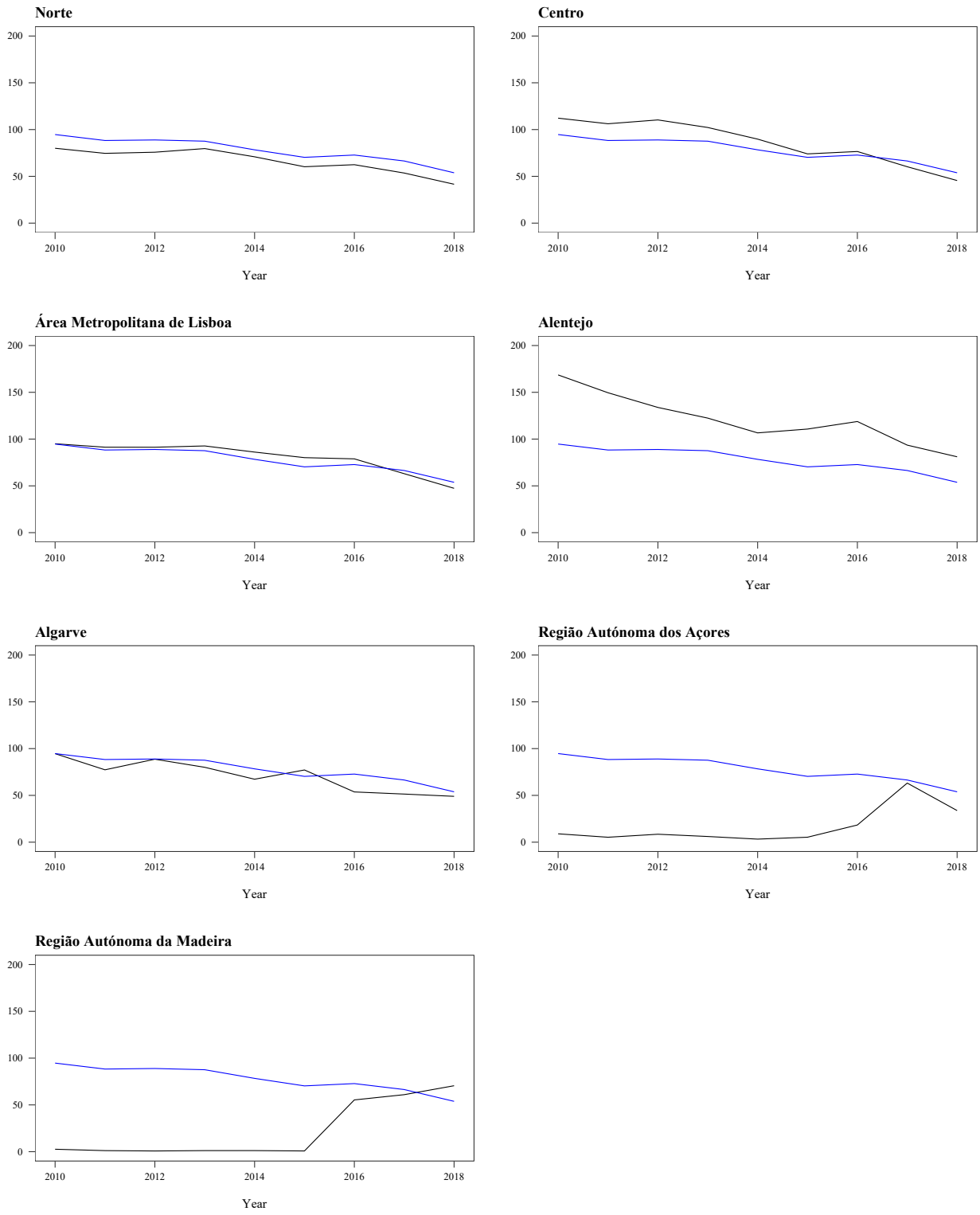

Figure A.1: Evolution of hospitalizations due to diabetes per 100,000 inhabitants by region in Portugal. The blue line represents the hospitalizations due to diabetes per 100,000 inhabitants in Portugal, for all population, over the years.

B Forecasting accuracy by lead time

Table B.1: Forecasting accuracy of the model $(1,1,2)\times(0,1,1)_{12}$ for 1 to 12-months-ahead, based on rolling-origin-update and rolling-origin-recalibration evaluation.

	Rolling-origin-update			**Rolling-origin-recalibration**					
	Fixed window			*Fixed window*			*Rolling window*		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
$l = 1$	44.9	54.8	9.1	41.1	48.7	8.2	39.5	47.4	7.8
$l = 2$	39.3	48.1	8.0	38.4	46.9	7.7	38.3	46.9	7.6
$l = 3$	40.4	49.1	8.3	41.3	48.4	8.4	40.8	47.8	8.3
$l = 4$	42.9	51.1	8.9	42.8	50.9	8.8	42.5	50.9	8.7
$l = 5$	43.7	54.6	9.4	43.9	54.3	9.4	43.9	54.6	9.3
$l = 6$	45.0	56.8	9.8	44.0	55.9	9.5	44.9	56.4	9.7
$l = 7$	50.0	62.2	11.1	50.3	62.2	11.1	51.9	63.5	11.3
$l = 8$	55.5	68.2	12.4	54.6	67.1	12.2	56.2	68.2	12.5
$l = 9$	58.2	71.0	13.3	58.3	70.8	13.3	59.4	71.8	13.5
$l = 10$	67.9	77.4	15.3	67.7	76.5	15.3	69.7	77.6	15.7
$l = 11$	66.1	76.9	15.3	65.4	76.6	15.1	66.8	77.9	15.4
$l = 12$	69.1	81.4	16.0	68.2	79.8	15.8	70.0	81.5	16.2

l, lead time; MAE, Mean Absolute Error; RMSE, Root Mean Square Error; MAPE, Mean Absolute Percentage Error.