UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Métodos de regressão na análise de nascimentos prematuros

Maria Francisca Monteiro Martins Antão

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado pelo
Professor Doutor Rui Martins

2022

Agradecimentos

À minha família, em especial aos meus avós que serão sempre um exemplo, agradeço toda a educação, força e apoio no meu percurso pessoal e académico.

Ao meu orientador Prof Dr. Rui Martins agradeço os conselhos e a compreensão que contribuíram para o melhoramento deste trabalho.

Um último agradecimento a todos os professores, colegas e amigos com os quais me cruzei e acompanharam o meu percurso académico neste últimos anos na FCUL.

Resumo

Sendo a prematuridade uma das principais causas associadas à mortalidade neonatal, em Portugal e no Mundo, torna-se fundamental estudar a evolução da prevalência de nascimentos prematuros em Portugal, num horizonte temporal de médio prazo (1989-2019).

Através da aplicação da Metodologia *Box Jenkins* foi possível escolher um modelo ARIMA adequado à modelação dos nascimentos Prematuros, de Baixo Peso e de Muito Baixo Peso. Para tal, seguiram-se as etapas propostas em estudos de análises de séries temporais: definição do problema, recolha de informação, análise exploratória dos dados, escolha do melhor modelo, avaliação dos resultados e previsão. A fase final de previsão com o modelo ARIMA consistiu num conjunto de previsões feitas dentro e fora do horizonte temporal considerado. O poder preditivo do modelo estimado para cada uma das séries foi, posteriormente, comparado com 2 modelos de previsão alternativos: o Modelo das Médias Móveis e o Modelo de Alisamento Exponencial.

Palavras-chave: Séries Temporais, Nascimentos Prematuros, ARIMA, Métodos de Previsão

Abstract

Prematurity is one of the main causes associated with neonatal mortality, not only in Portugal but globally, so it is essential to study the prevalence evolution of preterm births in Portugal in the medium term (1989-2019).

Through the application of Box Jenkins Methodology, it was possible to choose an ARIMA model suitable to Preterm, Low-Weight and Extremely Low-Weight birth data. To make that attainable there were followed the steps proposed in time series studies: problem definition, gather information, data exploratory analysis, identify the best model, evaluation results and forecasting. The final step was applied within and outside the considered time series horizon. The predictive power of the estimated model for each time series was subsequently compared with 2 alternative forecasting models: the Moving Average Model and Exponential Smoothing Model.

Keywords: Temporal Series , Preterm birth, ARIMA model, Forecasting Methods

Índice

Índice de Tabelas

Índice de Figuras

Índice de Anexos

Lista de Acrónimos e Abreviaturas

AE	Método de Alisamento Exponencial
AIC	Critério de Informação *Akaike*
AR	*Modelo Autoregressivo*
ARIMA	*Modelo Autoregressivo Integrado e de Médias Móveis*
ARMA	*Modelo Autoregressivo e de Médias Móveis*
BIC	Critério Bayesiano de Informação *Akaike*
FAC	Função de autocorrelação
FACP	Função de autocorrelação parcial
IG	Idade gestacional
INE	Instituto Nacional de Estatística
MM	Modelo das Médias Móveis
NBP	Nascimentos de Baixo peso
NMBP	Nascimentos de Muito Baixo Peso
NP	Nascimentos Prematuros
PN	Peso de nascimento
RN	Recém-nascido
RNMBP	Recém-Nascido de Muito Baixo Peso
ST	Série temporal

Capítulo 1. Introdução

1.1. Organização

No capítulo 1 é apresentada a introdução da dissertação e a organização da mesma, a definição do problema de investigação, a justificação do estudo e os principais objetivos que irão orientar a componente empírica.

O capítulo 2 (Enquadramento Teórico) encontra-se dividido em duas partes: a *Parte I* dedicada ao tema da Prematuridade e a *Parte II* às Séries Temporais. Na *Parte I* introduz-se a temática da Prematuridade, discute-se a sua evolução histórica no contexto europeu, contextualiza-se os conceitos de sobrevivência/mortalidade/viabilidade, analisa-se a evolução dos principais fatores de risco em Portugal e, por fim, discute-se a importância das previsões em estudos deste tipo. A Parte II é inteiramente dedicada às Séries Temporais e apresenta o suporte teórico usado no estudo de Séries Temporais. É iniciada pela introdução do conceito de série temporal e, em seguida, subdivide-se a análise em dois tipos de processos: processos estacionários (dos quais se descreve a função de autocorrelação/autocovariância e os modelos AR, MA e ARMA) e processos não estacionários (dos quais se descreve a técnica para tornar a média/variância estacionária e o modelo ARIMA). Por fim, explicitam-se as medidas de desempenho e os modelos alternativos de previsão, em particular o Método de Alisamento Exponencial e o Modelo das Médias Móveis.

O capítulo 3 descreve a metodologia escolhida, a Metodologia *Box Jenkins*, aplicada nas suas 3 principais etapas: identificação do modelo, estimação dos parâmetros e avaliação do diagnóstico. Acresce ainda informação sobre a caracterização dos dados utilizados e o software para análise.

No capítulo 4 é feita a apresentação e discussão dos Resultados obtidos e no capítulo 5 uma síntese dos principais Resultados e uma reflexão sobre as limitações e recomendações de questões futuras de investigação.

No final do presente trabalho encontra-se uma secção de Anexos constituída por *outputs* e tabelas que sustentam a análise do capítulo dos Resultados e acrescentam informação relevante.

1.2. Definição do problema e principais objetivos

Sendo a prematuridade uma das principais causas associadas à mortalidade neonatal, em Portugal e no Mundo, torna-se fundamental estudar a evolução da prevalência associada a médio/longo prazo em Portugal. Assim, o presente estudo tem como finalidade contribuir para o conhecimento da prevalência e evolução dos nascimentos Prematuros/ Baixo Peso/Muito Baixo Peso em Portugal num período de 30 anos. A pertinência do estudo justifica-se pelo reduzido número de estudos nacionais, a partir do ano de 2012, e pela aplicação inovadora de modelos de séries temporais na previsão de dados desta natureza.

No presente estudo destacam-se duas finalidades: a modelação, cuja base corresponde à construção de um modelo estatístico que permita descrever o mais adequadamente possível a evolução das 3 séries anteriores, e a previsão, cujo fim pressupõe a obtenção de valores futuros para o nº de nascimentos de cada uma das categorias anteriores. Contudo, torna-se fundamental um conjunto de decisões prévias,

entre as quais: decidir a natureza da previsão, limitar o horizonte temporal, detalhar o âmbito geográfico e escolher a população alvo.

Tendo em conta as finalidades do estudo, referidas anteriormente, os objetivos delineados nesta investigação são:

- *Produzir informação que permita caracterizar e modelar o fenómeno demográfico dos nascimentos de nados vivos prematuros/baixo peso/muito baixo peso em território nacional, no horizonte temporal de 1989 a 2019;*

- *Prever o nº de nascimentos Prematuros, de Baixo peso e de Muito baixo peso em Portugal no horizonte temporal de curto prazo (4 anos) e comparar os resultados obtidos com os valores registados;*

- *Avaliar e comparar a capacidade de ajustamento de 3 diferentes modelos preditivos: Modelo ARIMA, Modelo de Médias Móveis e Método de Alisamento Exponencial;*

Em síntese, para que os objetivos anteriores possam ser alcançados, o trabalho seguirá o modelo de passos a serem dados em estudos de análises de séries temporais:

1) *Definição do problema;*

2) *Recolha de informação;*

3) *Análise Exploratória dos dados;*

4) *Escolha do melhor modelo;*

5) *Avaliação dos Resultados;*

6) *Previsão.*

Capítulo 2. Enquadramento Teórico

2.1. Parte I

2.1.1. Nascimentos Prematuros e de Baixo Peso

A Organização Mundial da Saúde (OMS) estima que nasçam todos os anos 15 milhões de nados vivos prematuros e define prematuridade como o nascimento que ocorre antes das 37 semanas de gestação completas, subdividindo o conceito pela idade gestacional em que ocorre: prematuro extremo (<28 semanas), muito prematuro (28 a 31 semanas) e prematuro moderado (32 a 36 semanas).

O nascimento prematuro está também associado ao baixo peso à nascença, podendo ser:

- ***Nascimento de Baixo Peso*:**]1500g,2500g[
- ***Nascimento de Muito Baixo Peso ou de Extremo Baixo Peso*:**]500g,1500g].

As subdivisões por idade gestacional e por peso à nascença são importantes porque a decrescente idade gestacional e o peso à nascença estão associados a uma maior taxa de mortalidade e a uma maior necessidade de cuidados intensivos (World Health Organization [WHO], 2012).

Têm sido diversas as iniciativas e estudos epidemiológicos no contexto português, em especial nos últimos 30 anos, dedicadas ao estudo da prematuridade. Destacam-se as seguintes:

- Registo Nacional de Muito Baixo Peso (RNMBP): Foi uma iniciativa pioneira da Sociedade Portuguesa de Neonatologia (SPN) e da Sociedade Portuguesa de Pediatria (SPP) que se iniciou em 1994 com a participação de 13 unidades hospitalares, tendo como principal objetivo determinar a prevalência de RNMBP em Portugal. Procura também a consolidação de uma rede de estudos neonatais, o estimular da investigação e a determinação do contributo do RNMBP para a mortalidade neonatal. Integra atualmente todas as unidades neonatais do país e é um instrumento fundamental no registo e contabilização neonatal em Portugal.
- A monografia *Nascer Prematuro em Portugal*, um estudo multicêntrico nacional referente aos anos 1996-2000, galardoado com o Prémio Bial da medicina clínica (Peixoto et al., 2002). Neste estudo é feito um balanço e contabilização dos resultados do registo nacional do RNMBP indicado anteriormente no qual se constata que, no período considerado, a prevalência de RNMBP aumentou de 0,81% em 1996 até 0,94% em 2000 tendo atingido um máximo de 1,01% em 1999.

As causas associadas ao nascimento prematuro são ainda desconhecidas (WHO, 2012), sendo que em muitos casos o parto prematuro ocorre sem que seja identificada uma causa subjacente (Beck et al., 2009). Cerca de 45-50% dos nascimentos são espontâneos sem que tenha ocorrido rutura de membranas, 30% resultam dessa rutura e os restantes 15-20% ocorrem por indicação médica devido a complicações maternas ou fetais (Beck et al., 2009). Sabe-se também o nascimento prematuro e o baixo peso ao nascer têm uma etiologia multifatorial e entre os **fatores de risco** mencionados na literatura estão: fatores de

risco epidemiológico (fatores maternos, paternos e fetais), ambientais (socioeconómicos, stress, infeções) e genéticos (Murphy, 2007), dos quais se destacam:

- Gravidez gemelar/múltipla, idade materna, histórico de partos prematuros, tabagismo/alcoolismo, descolamento prematuro da placenta, infeções durante a gestação, entre outros.

A monitorização dos cuidados de saúde prestados no início da gravidez podem reduzir o risco de um parto prematuro, daí a importância da identificação, monitorização dos fatores de risco associados. Para que se possa dar significado e interpretar a evolução no nº de nascimentos prematuros/baixo peso importa, primeiramente, descrever como tem sido a evolução dos dois principais fatores de risco:

- Gravidez múltipla

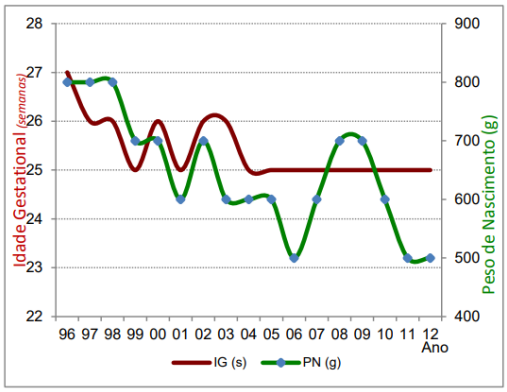
Os nascimentos resultantes de partos gemelares têm uma elevada frequência de prematuridade, baixo peso e muito baixo peso ao nascer (Peixoto et al., 2002). A idade gestacional média reportada é de 37 semanas, 50% dos nados vivos resultantes de partos gemelares registam peso ao nascer inferior a 2500g e 10% inferior a 1500g. Por outro lado, as gravidezes múltiplas de gémeos triplos/quádruplos, apesar de residuais, registam uma idade gestacional de 33 semanas, 90% têm peso ao nascer inferior a 2500g e 25% inferior a 1500g. O número de nados vivos resultantes de partos gemelares aumentou de 2,4% para 3,0% do total de nascimentos, entre 2001 e 2019 (Anexo B). A proporção de nados vivos gemelares é mais evidente nas mães de grupos etários superiores, sendo que aumentou de 1,7% para 4,3% entre 2001 e 2019 no grupo etário de mães acima dos 40 anos (Instituto Nacional de Estatística [INE], 2005).

- Idade materna

A transição para a parentalidade tende a acontecer cada vez mais tarde e regista-se o aumento da idade média das mulheres aquando do nascimento do primeiro filho, sendo esta tendência estritamente crescente desce 1983. (Anexo C). Este facto reflete-se num aumento da maternidade tardia (mães com idade superior a 35 anos) que tem vindo a aumentar desde o início da década de 2000 e, por comparação, a maternidade precoce apresenta-se em declínio (Anexo D). O grupo etário das mães entre os 30 e os 34 anos mantém-se entre 2001 e 2019 como aquele em que nascem mais nados vivos prematuros/baixo peso (Anexo F). Contudo, a maior subida regista-se no grupo etário dos 35 a 39 anos e > 40 anos.

Por outro lado, as complicações associadas aos nascimentos prematuros e aos nascimentos de baixo peso são a principal causa associada à mortalidade neonatal e a uma maior probabilidade de desenvolvimento de sequelas graves quando comparados com os nascimentos a termo. Aproximadamente 35% das 3,1 milhões de mortes neonatais que ocorrem anualmente em todo o mundo decorrem de complicações associadas a nascimentos prematuros/baixo peso/muito baixo peso.

O conceito de **viabilidade**, intrinsecamente ligado ao conceito de prematuridade, define-se como a idade gestacional a partir da qual o RN tem > 50% de hipóteses de sobrevida e em que pelo menos 50% dos sobreviventes ficam sem sequelas a longo prazo. O limite da viabilidade e a capacidade de intervenção com sucesso regista-se em idades gestacionais e pesos de nascimento cada vez menores. Em Portugal, no ano de 1996 situava-se nas 27 semanas e 800g e em 2012 nas 25 semanas e 500g de peso ao nascer, sendo que desde 2005 se regista uma estagnação na idade gestacional relativa ao limite de viabilidade.



Fonte: Mimoso, G., & Almeida, A. (2018)

Figura 2.1. Evolução do limiar de viabilidade de 1996 a 2012

A OMS classificou em 2013 o nascimento prematuro como uma prioridade a nível de políticas de saúde pública, tendo definido um conjunto de políticas de prevenção, gestão e acompanhamento posterior dos nascimentos prematuros. A nível internacional um dos 8 objetivos de desenvolvimentos do milénio definidos pelas Nações Unidas foi a redução em dois terços da taxa de mortalidade neonatal até ao ano de 2015 e sendo o nascimento prematuro uma das principais causas associadas à mortalidade neonatal, a sua redução contribuiu significativamente para o alcançar dessa meta.

Em Portugal, a **taxa de mortalidade** associada à percentagem de nascimentos com peso <1500g decresceu de 27% em 1996 para 15% em 2012 e a percentagem de nascimentos com peso <1000g sofreu um decréscimo ainda maior de 54% em 1994 para 29% em 2012. A redução do número de óbitos neonatais surge como consequência da crescente sobrevivência associada aos nascimentos prematuros. A taxa de sobrevivência global dos nados vivos com peso à nascença > 1000g subiu de 72,7% em 1996 para 79,8% em 2000 e para valores de peso <1000g de 45,3% em 1996 para 60,3% em 2000 (Figura 2.2). Dados mais recentes da SPP indicam que as taxas de sobrevida para nascimentos com peso inferior a 1500g têm aumentado para mais de 85% e atualmente um nado vivo nascido com menos de 1000g tem uma probabilidade de sobreviver de mais de 95%.

Assim, a diminuição da taxa de mortalidade e o aumento da taxa de sobrevivência registadas nesta população tem contribuído para o aumento do número do número de nascimentos prematuros e de baixo peso/muito baixo peso. A este facto associa-se a capacidade de sucesso em idades gestacionais cada vez menores às quais correspondem pesos à nascença proporcionalmente decrescentes (Figura 2.2).

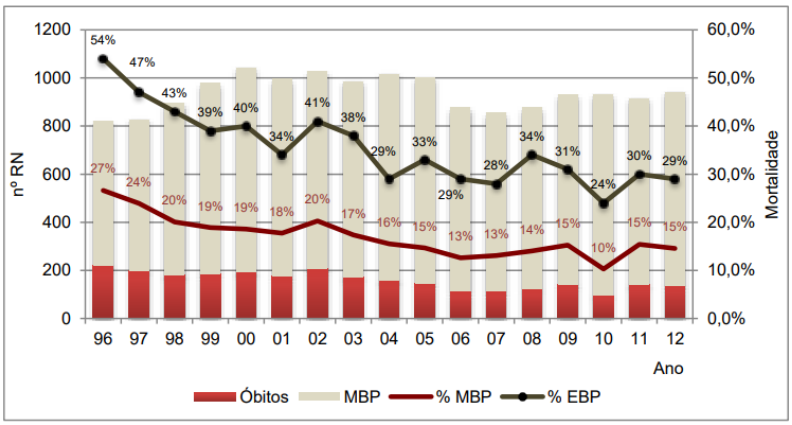


Figura 2.2. Evolução da Mortalidade dos RN com PN <1500g e <1000g entre 1996 e 2012

Fonte: Mimoso, G., & Almeida, A. (2018)

Paralelamente à evolução da taxa de mortalidade e da taxa de viabilidade assiste-se, em Portugal, a uma tendência de declínio nas últimas décadas do número de nascimentos e da consequente descida na taxa bruta de natalidade. Em 1960 a taxa bruta de natalidade era de 24,1%, em 1990 o valor era de 11,7% e em 2019 decresceu atingindo os 8,4%. Carrilho e Peixoto (1993) destacam, como principais fatores explicativos da queda da natalidade entre 1981 e 1992: o retardar do casamento e da idade ao nascimento do primeiro filho, a difusão dos métodos contracetivos, a dificuldade dos jovens no acesso à habitação e ao emprego e o maior grau de instrução e atividade profissional da mulher. Os fatores anteriores reforçam a ideia de que a natalidade deve ser enquadrada e pensada no contexto socioeconómico.

Mais recentemente, segundo dados do Eurostat, Portugal teve em 2019 a 5ª menor taxa de natalidade (8,4 nascimentos por 1000 habitantes) da União Europeia.

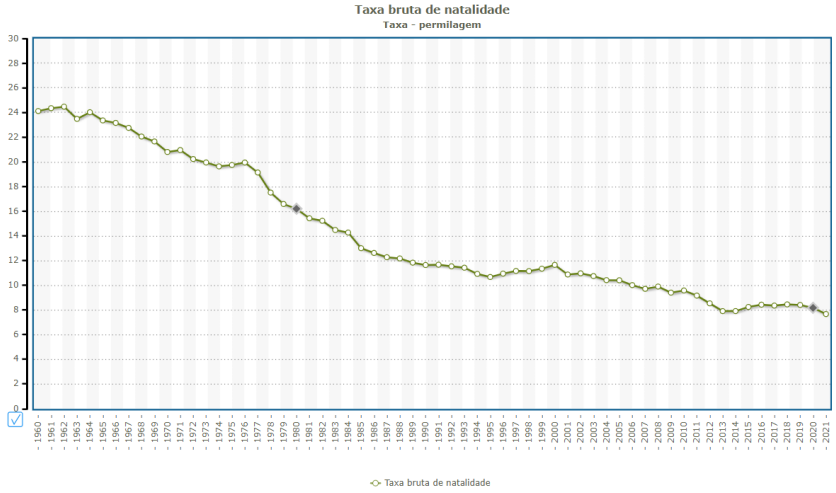


Figura 2.3. Evolução da taxa bruta de natalidade em Portugal, nos anos 1960 a 2021

Fonte: INE/PORDATA (2022)

2.1.2. Evolução da prematuridade na Europa

Na Europa nascem anualmente, em média, 5 milhões de nados vivos, dos quais 5% a 10% correspondem a nados vivos prematuros (Murphy, 2015), sendo que 2/3 das mortes neonatais ocorrem, no contexto europeu, em nados vivos nascidos antes das 37 semanas de gestação.

Na Tabela 2.1 representam-se as percentagens de nascimentos prematuros em 19 países da Europa medidas nos anos 1996,2000,2004 e 2008. As percentagens descritas são heterogéneas entre países e variam consoante a definição de prematuridade adotada, em alguns casos o limiar situa-se nas 24 semanas e noutros em 28 semanas de gestação. Dos 19 países considerados, 15 aumentaram a taxa de prematuridade no intervalo considerado sendo Portugal e a Áustria os países em que a taxa mais subiu entre 1996 e 2008, registando um aumento de 2%, contrariamente a casos como o da Finlândia ou a Suécia em que houve uma diminuição da taxa de prematuridade. Em 2008 as taxas de prematuridade para o total de nascimentos variavam entre os 5,5% (Finlândia) e os 11,1% (Áustria), para os nascimentos singulares o intervalo era de]4,3%,8,7% [e para nascimentos gemelares/múltiplos o intervalo era de]42,1%, 77,8%[(Murphy, 2015).

De referir que apesar da variação da taxa de prematuridade em alguns países ser pouco significativa, o impacto é substancial quando esse valor se traduz em nascimentos ocorridos. Se os restantes países tivessem tido variações semelhantes à da Finlândia/Países Baixos (-0,6% ao ano), teriam nascido menos 24 000 nados vivos prematuros em 2008, valor esse que corresponderia a 1,2% dos 2 milhões de nascimentos na totalidade dos países considerados.

Tabela 2.1. Percentagens de nascimentos prematuros em 19 países europeus entre 1996 e 2008

Country: region/area	All live births				
	n (2008)	1996 %	2000 %	2004 %	2008 %
Austria	77 720	9,1	10,0	11,4	11,1
Belgium: Flanders	69 187	7,0	7,8	8,1	8,0
Czech Republic	119 455		5,4	7,7	8,3
Estonia	16 031	5,5	5,9	5,9	6,2
Finland	59 486	5,8	6,1	5,6	5,5
France*	14 696	5,4	6,2	6,3	6,6
Germany: 3 Länder	215 634		8,8	9,2	9,0
Ireland	75 246		5,4	5,5	5,9
Lithuania	31 287	5,3	5,3	5,3	5,9
Malta**	4 152		6,0	7,2	6,7
the Netherlands	175 160	7,8	7,7	7,4	7,4
Norway	60 744	6,4	6,8	7,1	6,7
Poland	414 480	6,8	6,3	6,8	6,6
Portugal	103 597	7,0	5,9	6,8	9,0
Slovakia	53 624	5,1	5,4	6,3	6,8
Slovenia	21 816	6,0	6,8	7,0	7,4
Spain	417 094	7,1	7,7	8,0	8,2
Sweden**	108 865	6,1	6,4	6,3	5,9
UK: Scotland	58 275	7,0	7,4	7,6	7,7

Fonte: Murphy, M. M., & McLoughlin, G. (2015)

2.1.3. Importância da modelação e previsão no estudo dos Nascimentos Prematuros/Baixo Peso

Toda a informação descrita anteriormente depende, primeiramente, da contabilização do nº de nascimentos prematuros, baixo peso e muito baixo peso. Só então essas observações, mais à frente designadas séries temporais, podem ser modeladas. A modelação permite identificar padrões, tendências e aferir o impacto da implementação de políticas públicas e estratégias de prevenção.

A análise de séries temporais é, com frequência, aplicada no estudo do impacto da aplicação de uma determinada política em saúde pública. Em análise deste tipo, denominadas análises de séries interrompidas, considera-se que o impacto de uma política é significativo se se verificar uma diferença considerável entre observações medidas pré-intervenção e pós-intervenção (McCleary, 1980).

Em Portugal, as políticas de prevenção no âmbito dos nascimentos prematuros e de baixo peso foram descritas pela Comissão Nacional de Saúde Materna e Infantil (1989) e passam por: um melhor acompanhamento pré-natal, a garantia de qualidade e segurança no momento do parto, identificação e acompanhamento precoce dos fatores de risco associados e pela concentração dos partos de grande risco em hospitais Centrais que disponham dos meios para fazer face às necessidades acrescidas da mãe e do recém-nascido. Nas últimas décadas houve também uma evolução tecnológica contínua na área da Neonatologia. As previsões associadas aos nascimentos dos subgrupos descritos são também cruciais para que possam ser tomadas decisões ao nível do sistema de saúde de apoio e dos meios humanos e materiais necessários em cada hospital.

No entanto, para que tal seja possível é imprescindível, numa primeira fase, conhecer os mecanismos teóricos por detrás da modelação e previsão das séries, mecanismos esses que se apresentam seguidamente na *Parte II* do Enquadramento teórico.

2.2. *Parte II* - Séries Temporais

2.2.1. Conceito, objetivos e componentes

Uma **sucessão cronológica** define-se como um conjunto de observações associadas a um determinado fenómeno aleatório feitas em pontos específicos ou períodos sucessivos de tempo, geralmente igualmente espaçados (e.g., dias, meses, trimestres, semestres, anos, etc.) (Murteira et al., 1993) e designa-se por:

$$\{X_t,\ t = 1,2, \dots, n\} \qquad \text{Equação 2.1}$$

O interesse no estudo das sucessões cronológicas surgiu da preocupação com o aspeto dinâmico dos fenómenos, a necessidade de se fazerem previsões e é transversal a vários domínios. Em Economia as sucessões cronológicas utilizam-se em especial no plano macroeconómico, como por exemplo, na análise das flutuações das taxas de juro ou no estudo da variação do produto interno bruto, como forma de a avaliar o crescimento económico de um determinado país ou região; Em meteorologia no registo das temperaturas atmosféricas ou no registo das precipitação atmosférica; Em medicina na interpretação de eletroencefalogramas ou na contabilização de casos e registo da evolução da propagação de uma doença; Ou por fim , como será desenvolvido mais à frente, em demografia, em particular no estudo da evolução demográfica.

Quanto à natureza da variável medida, as séries temporais, também chamadas sucessões cronológicas/sucessões temporais ou em inglês *time series*, podem ser classificadas como séries temporais **contínuas** ou **discretas**. As séries temporais podem ainda ser **univariadas** se são constituídas por uma única observação em cada ponto, ou **multivariadas** se são obtidas observações simultâneas de dois ou mais fenómenos. Neste estudo, como será explicitado mais à frente, serão utilizadas apenas séries temporais univariadas.

Chatfield (2004) defende que o estudo de uma série cronológica deve ter 4 **objetivos**:

- **Descrição**: Inicia-se com a representação gráfica e construção do cronograma para que se tenha uma ideia preliminar dos dados e se identifiquem visualmente padrões de tendência, de sazonalidade, pontos de inflexão ou *outliers*. Faz-se em seguida uma análise da estatística descritiva através dos indicadores apropriados de modo a compreender o mecanismo gerador da série;

- **Explicação**: Após a observação e descrição de uma sucessão cronológica é possível construir um modelo que permita explicar o comportamento da série no período observado. Estimam-se os parâmetros e avalia-se a qualidade do diagnóstico com base na qualidade estatística e do ajustamento do modelo proposto**;**

- **Previsão:** Os métodos de previsão permitem estimar valores futuros, com maior ou menor precisão, com base em valores passados de uma série temporal. As técnicas de previsão do comportamento futuro da série são de especial importância na construção, execução e controlo de planos de médio/longo prazo;

- **Controlo:** É necessária a monotorização da série para que possam ser detetadas alterações nos valores e características da série. Estas alterações podem indiciar que o modelo proposto deixou de ser válido e que é necessário retomar à fase de modelação e repetir o processo.

Uma série temporal pode ainda ser decomposta em diversas **componentes:**

- **Tendência (T_t):** compreende os movimentos suaves e consistentes de longo prazo podendo estes serem lineares ou não lineares, crescentes ou decrescentes;

- **Sazonalidade (E_t):** São flutuações nos valores da variável com duração inferior a um ano que pode ter causas sociais ou naturais e se repetem anualmente, ou num período inferior. Os testes mais usados para estimar a sazonalidade são o teste de Kruskal-Wallis (para amostras independentes), o teste de FRiedman (para amostras dependentes) e o Teste *F*;

- **Movimentos oscilatórios ou cíclicos (C_t):** são flutuações nos valores da variável com duração superior a um ano e que não apresentam periodicidade fixa, ou seja, o seu comprimento varia de ciclo para ciclo;

- **Ruído ou Componente aleatória (e_t):** são flutuações aleatórias fruto do acaso ou resultantes de acontecimentos inesperados.

O modelo tradicional de análise de séries temporais pressupõe que o conjunto de observações que constitui a série é o resultado da interação das componentes anteriores, sendo esta abordagem chamada modelação de componentes ou análise de decomposição. Os modelos mais utilizados que relacionam as diversas componentes são:

1) Modelo aditivo:

$$X_t = T_t + E_t + C_t + e_t \qquad \text{Equação 2.2}$$

2) Modelo multiplicativo:

$$X_t = T_t \times E_t \times C_t \times e_t \qquad \text{Equação 2.3}$$

3) Modelo Misto:

$$X_t = (T_t . + C_t) \times E_t + e_t \qquad \text{Equação 2.4}$$

Nem sempre estão presentes todas as componentes citadas sendo necessário decompor a série temporal para identificar quais as componentes presentes.

2.2.2. Processos Estacionários

Chatfield (2004) considera que a maioria das séries temporais são estocásticas, em que resultados futuros são parcialmente determinados por valores passados, sendo o modelo para estas séries definido como um processo estocástico. Um **processo estocástico** é uma coleção ordenada de variáveis aleatórias em que existe um estado de equilíbrio em torno de um nível médio fixo e as propriedades probabilísticas são estáveis e invariantes ao longo do tempo (Murteira et al., 1994). Uma sequência de variáveis aleatórias definidas em intervalos de tempo fixos designa-se por processo estocástico discreto ou simplesmente sucessão cronológica/série temporal. O objetivo da análise de sucessões cronológicas é fazer inferências sobre um processo estocástico desconhecido tendo como informação disponível uma única realização observada.

Os processos estocásticos subdividem-se em estacionários e não estacionários. Considera-se $\{X_t: t \in \mathbb{R}\}$ um **processo estritamente estacionário**, se a distribuição conjunta de $(X_{t1}, \cdots X_{tn})$ for igual à distribuição conjunta de $(X_{t1+k}, \cdots, X_{tn+k})$ para todo n ∈ ℕ e k ∈ ℝ. Isto é,

$$F_{(X_{t1},\cdots X_{tn})}(x_1, \dots, x_n) = F_{(X_{t1+k},\cdots,X_{tn+k})}(x_1, \dots, x_n) \qquad \text{Equação 2.5}$$

Uma das características principais de uma série temporal é a sua estacionaridade porque, na maioria dos casos, para que uma série temporal possa ser modelada é necessário torná-la estacionária de forma que as suas propriedades não se alterem quaisquer que sejam os instantes de tempo considerados. Os processos estacionários traduzem-se no equilíbrio em torno de um nível médio fixo, ou seja, um processo estacionário possui propriedades que são invariantes e estáveis no tempo.

2.2.2.1. Função de autocorrelação e autocovariância

Seja $\{X_t\}$ um processo estacionário com média e variância constantes, a função de autocovariância do processo define-se por,

$$\gamma_k = E\{(X_t - \mu)(X_{t+k} - \mu)\},\ k \in \mathbb{Z} \qquad \text{Equação 2.6}$$

Em que cada *k* da função γ_k mede a intensidade em que covariam os pares do processo estacionário separados por um intervalo (*lag*) com amplitude k.

Em particular, na modelação de séries temporais é importante apresentar a estimação dos parâmetros que caracterizam o processo, por isso indicam-se também os respetivos estimadores.

A função de autocovariância é estimada por:

$$\hat{\gamma}_k = \frac{1}{N}\sum_{t=1}^{N-k}((X_t - \bar{X})(X_{t+k} - \bar{X})),\quad 0 \leq k \leq \text{N-1} \qquad \text{Equação 2.7}$$

A **função de autocorrelação (FAC)** do processo, define-se por:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \qquad \text{Equação 2.8}$$

Intuitivamente interpreta-se ρ_k como uma medida de de semelhança entre cada realização e essa mesma realização deslocada *k* unidades no tempo (Murteira et al.,1993). A representação gráfica de ρ_k em função de k designa-se por correlograma teórico sendo que para *k* a função p_k mede a correlação entre pares de valores separados por um intervalo de tempo k, ou seja quão forte o valor observado hoje está correlacionado com valores observados no passado. Como será aprofundado mais à frente, o correlograma é um auxiliar importante na identificação do modelo subjacente à análise pois permite caracterizar o desenvolvimento de X_t ao longo do tempo.

A função de autocorrelação é estimada por:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{N-k}((X_t - \bar{X})(X_{t+k} - \bar{X}))}{\sum_{t=1}^{N-k}((X_t - \bar{X}))^2} \qquad \text{Equação 2.9}$$

A função de autocorrelação amostral é estimada, para k=0,...N, por:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{1/N \sum_{t=1}^{N-k}((X_t - \bar{X})(X_{t+k} - \bar{X}))}{1/N \sum_{t=1}^{N-k}((X_t - \bar{X}))^2} \qquad \text{Equação 2.10}$$

Por outro lado, as bandas de confiança de 95% para ρ_k são dadas por:

$$\pm \frac{1{,}96}{\sqrt{N}} \approx \frac{2}{\sqrt{N}} \qquad \text{Equação 2.11}$$

As principais propriedades das funções autocovariância e autocorrelação são:

1) $\gamma_0 = \sigma^2$; $\rho_0 = 1$;
2) $|\gamma_k| \leq \gamma_0$; $|\rho_0| \leq 1$;
3) $\gamma_k = \gamma_{-k}$; $\rho_k = \rho_{-k}$
4) $|k| \to \infty \Rightarrow \gamma_k \to 0 \; e \; \rho_k \to 0$

A **função de autocorrelação parcial (FACP),** contrariamente à FAC, elimina o efeito das variáveis intermédias. Para obter a expressão da FACP ajusta-se uma regressão linear múltipla de X_{t+k} sobre $X_{t+k-1}, X_{t+k-2}, \ldots, X_{t+1}, X_t$,isto é:

$$X_{t+k} = \phi_{k1}X_{t-k-1} + \phi_{k2}X_{t-k-2} + \cdots + \phi_{kk}X_t + \varepsilon_{t+k} \qquad \text{Equação 2.12}$$

Em que ϕ_{kj} , j =1,2…k, são os coeficientes de regressão. Multiplicam-se ambos os membros da expressão anterior por X_{t+k-j}, consideram-se os valores esperados e divide-se por γ_0 obtendo-se:

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{j-k}$$

Que se resolve em ordem aos coeficientes ϕ_{kj} pela regra de Cramer, obtendo a expressão da FACP:

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \dots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \dots & \rho_{k-3} & \rho_2 \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{k-1} & \rho_{k-2} & \dots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{k-1} \\ \rho_1 & 1 & \dots & \rho_1 & \dots & \rho_{k-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \dots & 1 \end{vmatrix}} \qquad \text{Equação 2.13}$$

Ruído branco

Um processo $\{X_t, t = 0, \pm 1, \pm 2, \dots\}$ diz-se ruído branco ou puramente aleatório quando é formado por uma sucessão de variáveis aleatórias não correlacionadas e identicamente distribuídas de média e variância constantes, e designa-se por $\{\varepsilon_t\}$ se:

1) $Cov[(X_t, X_s)] = 0 \; \forall t \neq s$ Equação 2.14

2) $E(X_t) = \mu$ Equação 2.15

3) $Var\,(X_t) = \sigma_x^2$ Equação 2.16

Se para além das condições acima, as variáveis aleatórias seguirem uma distribuição Normal (i.e. $\varepsilon_t \sim N(\mu, \sigma^2)$, o processo designa-se por ruído branco gaussiano.

As séries de ruído branco são raramente identificadas em situações reais, mas desempenham um papel central na construção de modelos de previsão. Assim, um bom modelo deve ser aquele que produz erros de previsão com um comportamento semelhante ao de um ruído branco.

Na Figura 2.4 seguinte representa-se a simulação de um ruído branco de média nula e variância unitária e as respetivas FAC e FACP empíricas.

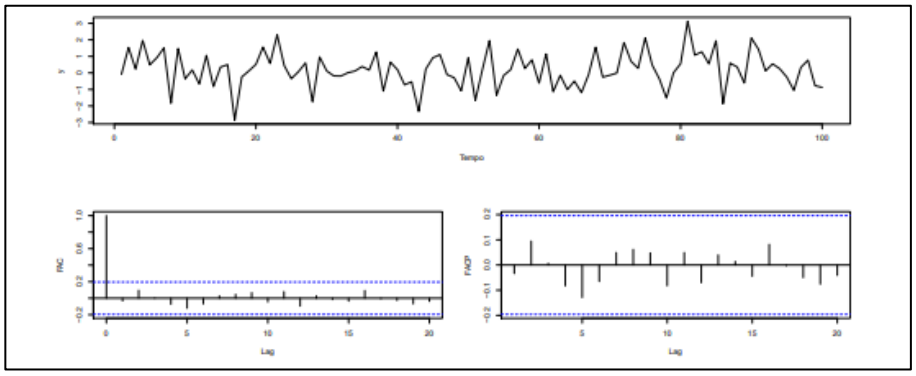


Figura 2.4. Representação de um ruído branco e respetivas FAC e FACP empírica

2.2.2.2. Modelos Autoregressivos (AR)

Em 1927 surgem, devido a Yule, os modelos autoregressivos (**AR**), um processo X_t diz-se um processo autoregressivo de 1ªordem [**AR (1)**] se satisfaz a equação:

$$X_t = \phi X_{t-1} + \varepsilon_t \qquad \text{Equação 2.17}$$

Equivalente a,

$$(1 - \phi B)X_t = \varepsilon_t \qquad \text{Equação 2.18}$$

Ou

$$\phi_1(B)X_t = \varepsilon_t \qquad \text{Equação 2.19}$$

Em que ϕ é um número real, ε_t o ruído branco e B o operador de atraso. O polinómio autoregressico de 1ªordem corresponde a

$$\phi_1(B) = 1 - \phi B \qquad \text{Equação 2.20}$$

O processo X_t diz-se um processo autoregressivo de 2ªordem [**AR (2)**] se satisfaz a equação:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \varepsilon_t \qquad \text{Equação 2.21}$$

Ou

$$\phi_2(B)X_t = \varepsilon_t \qquad \text{Equação 2.22}$$

Onde

$$\phi_2(B) = 1 - \phi_1 B - \phi_2 B^2 \qquad \text{Equação 2.23}$$

Onde $\phi_2(B)$ é um polinómio autoregressivo de 2ªordem.

Os processos autoregressivos de 1ª e 2ªordem podem ser generalizados. O processo X_t diz-se **AR(p)** quando satisfaz a equação:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \qquad \text{Equação 2.24}$$

Com t ϵ $\mathbb{Z}$ e $\phi_1, \ldots, \phi_p$ constantes reais. Pode ainda ser reescrito na forma,

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = \varepsilon_t \qquad \text{Equação 2.25}$$

Ou

$$\phi_p(B)X_t = \varepsilon_t \qquad \text{Equação 2.26}$$

Em que

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \qquad \text{Equação 2.27}$$

É o polinómio autoregressivo de ordem *p*.

2.2.2.3. Modelos de médias móveis (MA)

Os processos médias móveis de ordem q [**MA(q)**] satisfazem a equação:

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \qquad \text{Equação 2.28}$$

Com $\theta_1, \ldots, \theta_q$ constantes reais. Ou podem ser reescritos por,

$$X_t = \theta_q(B)\varepsilon_t \qquad \text{Equação 2.29}$$

Onde

$$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \qquad \text{Equação 2.30}$$

É o polinómio médias móveis de ordem q, ε_t um processo aleatório (ruído branco) e $(\theta_1 \ldots, \theta_q)$ são constantes reais.

2.2.2.4. Modelos Autoregressivos e de Médias Móveis (ARMA)

Yule introduz em 1926 os processos Autoregressivos e, mais tarde, em 1938 é introduzida por Wold a primeira versão dos processos Autoregressivos e de Médias Móveis, ARMA, e a sua posterior utilização na modelação de séries estacionárias.

O processo X_t diz-se um processo Misto Autoregressivo e Médias móveis [**ARMA (p, q)**] e escreve-se $\{X_t, t \in \mathbb{Z}\} \sim ARMA(p,q)$ se satisfaz a equação,

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ \ldots \ + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ \ldots + \theta_q \varepsilon_{t-q} \qquad \text{Equação 2.31}$$

No processo descrito acima, o parâmetro *p* diz respeito ao nº de parâmetros autoregressivos do modelo e *q* o nº de parâmetros de médias móveis.

A equação anterior pode ainda ser reescrita como:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad \text{Equação 2.32}$$

Ou

$$\phi_p(B) X_t = \theta_q(B) \varepsilon_t \quad \text{Equação 2.33}$$

Com

$$\phi_p(B) = 1 + \phi_1 B + \phi_2 B^2 + \ldots + \phi_p B^p \quad \text{Equação 2.34}$$

E

$$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \quad \text{Equação 2.35}$$

Na Figura 2.5 exemplificam-se os vários processos descritos anteriormente. A figura *a)* representa um processo estável autoregressivo de ordem 1, AR(1), com coeficiente positivo; A figura *b)* um processo autoregressivo de ordem 2,AR(2), em que é visível uma oscilação de sinal; A figura *c)* representa um processo de médias móveis de ordem 2, MA(2), em que se destacam dois valores significativos e há uma convergência rápida para 0 e, por fim, a figura *d)* onde se representa um processo autoregressivos e de médias móveis de ordem 2, ARMA(2,2), em que a oscilação de sinal, comparativamente a *b)*, é menos visível.

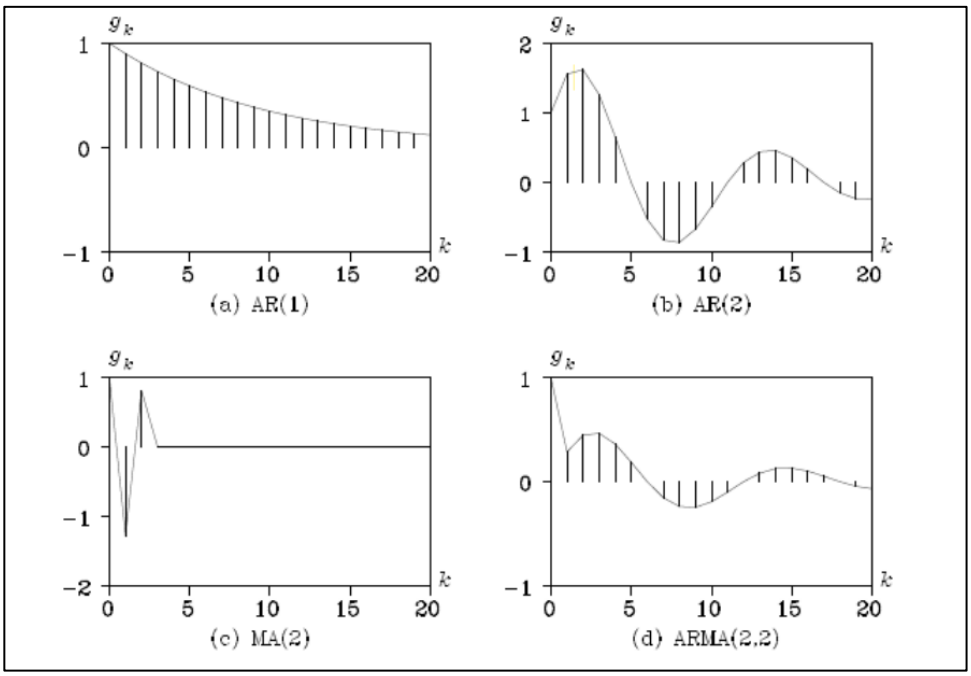


Figura 2.5. Exemplificação de processos AR, MA e ARMA

Fonte: Kitagawa, G. (2010)

A Tabela 2.2 complementa a informação anterior e resume as características principais dos modelos AR, MA e ARMA.

Tabela 2.2. Resumo descritivo das principais características dos modelos AR, MA e ARMA

	AR(p)	**MA(q)**	**ARMA(p, q)**
Modelo em termos dos valores anteriores de X_t	$\phi_p(B)X_t = \varepsilon_t$ Séries finita em X_t	$[\theta_q(B)]^{-1}X_t = \varepsilon_t$ Série infinita em X_t	$[\theta_q(B)]^{-1}\phi_p(B)X_t = \varepsilon_t$ Série infinita em X_t
Modelo em termos dos valores anteriores de ε_t	$X_t = [\phi_p(B)]^{-1}\varepsilon_t$ Séries infinita em ε_t	$X_t = \theta_q(B)\varepsilon_t$ Série finita em ε_t	$[[\phi_p(B)\ ]^{-1}\theta_q(B)\varepsilon_t$ Série infinita em ε_t
Condições de estacionaridade	Raízes de $\phi_p(B) = 0$ fora do círculo unitário	Sempre estacionários	Raízes de $\phi_p(B) = 0$ fora do círculo unitário
FAC	Decaimento exponencial e/ou sinusoidal para zero	Decaimento brusco para zero a partir de *k=q+1*	Decaimento exponencial e/ou sinusoidal para zero
FACP	Decaimento brusco para zero a partir de *k=p+1*	Decaimento exponencial e/ou sinusoidal para zero	Decaimento exponencial e/ou sinusoidal para zero

Fonte: Murteira et al. (1993).

Refere-se, ainda que não seja utilizado na componente empírica, o modelo ARMA sazonal. Os processos mistos autoregressivos e médias móveis estritamente sazonais [**ARMA (P, Q)** s] são definidos por:

$$\phi_p(B^S)X_t = \Theta_Q(B^s)\varepsilon_t \qquad \text{Equação 2.36}$$

Em que,

$$\phi_p(B^S) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^{pS} \qquad \text{Equação 2.37}$$

E

$$\Theta_Q(B^s) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^{qS} \qquad \text{Equação 2.38}$$

Representam os polinómios autoregressivos e médias móveis sazonais.

2.2.3. Processos Não Estacionários

Um dos aspetos a ser validado é a estacionaridade das séries temporais utilizadas, isto é, verificar se as séries estudadas seguem um processo estocástico com média e variância constantes no tempo. Uma série temporal pode ser não estacionária em média quando o nível da série não é estável no tempo, apresentando uma tendência crescente ou decrescente ou não estacionária na variância.

Para converter uma sucessão cronológica não estacionária numa sucessão estacionária recorre-se a transformações, descritas em seguida, que estabilizam a média e/ou a variância. Assim, torna-se possível a inferência estatística, o que não aconteceria que a série fosse não estacionária, sendo que a análise é simplificada dada a estabilidade dos parâmetros estimados.

Inclui-se, em seguida, a representação de três séries temporais ilustrativas de 3 tipos de processos não estacionários:

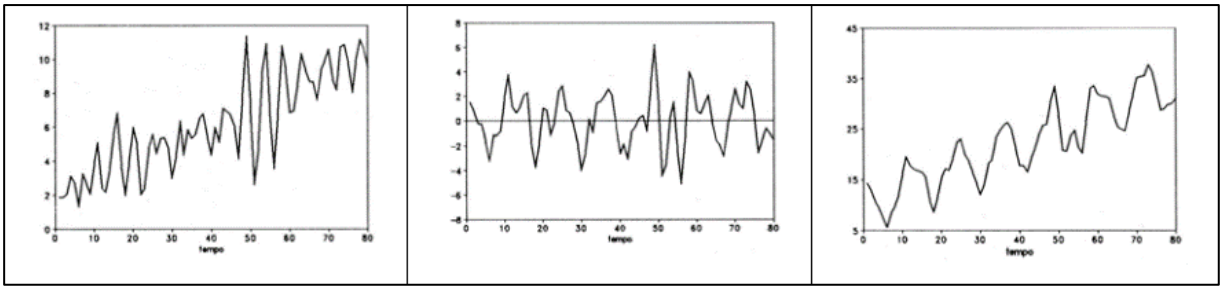


Figura 2.5. Exemplificação de uma série temporal com tendência linear (a), com componente sazonal (b) e com tendência linear e componente sazonal (c)
Fonte: Kitagawa, G. (2010)

2.2.3.1. Tornar a variância estacionária

Box e Cox (1964) descrevem uma transformação paramétrica que permite estabilizar a variância:

$$T_\lambda\ \left(X_t\ \right) = X_t^\lambda = \begin{cases} \left((X_t^\lambda - 1)\right)/\lambda\,, & \text{se } \lambda \neq 0 \\ \ln X_t\,, & \text{se } \lambda = 0 \end{cases} \qquad \text{Equação 2.39}$$

Com λ no intervalo [-1,1].

2.2.3.2. Tornar a média estacionária

Operador Diferença – Seja ∇ o operador diferença,

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t \qquad \text{Equação 2.40}$$

Sendo d a ordem das diferenças e para qualquer $d \geq 0$,

$$\nabla^d X_t = (1-B)^d X_t \qquad \text{Equação 2.41}$$

Os processos não estacionários em média que se transformam em processos estacionários por meio de diferenciação designam-se processos não estacionários homogéneos. Como descrito acima, uma forma de estabilizar a média consiste na utilização de processos de diferenciação que resultam da aplicação do operador diferença à série original não estacionária. Na maioria das situações, a série original X_t não estacionária pode ser transformada numa série estacionária quando aplicada uma diferenciação de 1ª ordem. Contudo, para remover a tendência de uma série poderá ser necessário aplicar o operador diferença mais do que uma vez, ou seja, obterem-se diferenças de 2ªordem, com a reserva de que este processo deve ser usado o número de vezes estritamente necessário para estabilizar a média da série. Se a série temporal for inicialmente não estacionária e esta característica não for tida em consideração, poderão surgir problemas na inferência e resultar em previsões incorretas.

O número de diferenças (*lags*) necessários para que uma série se torne estacionária é denominado ordem de integração. Assim, a série original é integrável de ordem 1 e representa-se por *I(1)* quando a série original é diferenciada uma vez e a série diferenciada é estacionária. No caso geral, a série é integrável de ordem *d* e representa-se por *I(d),* se a série for diferenciada *d* vezes. Quando *d=0,* o processo *I(0)* é um processo estacionário.

2.2.3.3. Modelo ARIMA

Os modelos **ARIMA (p, d,q)** em inglês *Auto-Regressive Integrated Moving Average Model* são utilizados para processos não estacionários que podem ser convertidos em estacionários e representam-se por:

$$\phi_p(B)(1-B)^d X_t = \theta_0 + \theta_q(B)\varepsilon_t \qquad \text{Equação 2.42}$$

Com,

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \qquad \text{Equação 2.43}$$

E

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \qquad \text{Equação 2.44}$$

Os polinómios autoregressivos estacionários e médias móveis invertíveis;

Em síntese, um processo ARIMA compreende três tipos de processos específicos: um processo autoregressivo AR em que ***p*** constitui o número de termos autoregressivos; um processo de integração I em que ***d*** corresponde ao número de diferenciações necessárias para que a série se torne estacionária e um processo de médias móveis MA em que ***q*** corresponde ao número de termos de médias móveis.

Na prática a maioria das séries temporais encontradas são não estacionárias sendo necessário remover dos dados as componentes que a tornam não estacionária (e.g., tendência, sazonalidade). O processo ARIMA é inicialmente um processo não estacionário que se transforma num processo estacionário e

invertível ARMA (p, q) depois de ser diferenciado d vezes. Os modelos vistos anteriormente (AR, MA e ARMA) podem ser descritos sucintamente usando apenas a nomenclatura ARIMA:

a. ARIMA(p,0,0) = AR(p) Equação 2.45

b. ARIMA (0,0, q) = MA(q) Equação 2.46

c. ARIMA(p,0, q) = ARMA(p,q). Equação 2.47

De referir que os modelos ARIMA traduzem uma representação razoável das séries temporais, mas que a qualidade da escolha do modelo correcto depende em boa parte de uma análise intuitiva e subjetiva do analista. Esta análise deve ter presente um conjunto de objetivos e obstáculos, como o objetivo do estudo, o nº de observações disponíveis ou os custos inerentes à utilização de determinado modelo.

- **Modelos ARIMA na previsão de nascimentos**

Saboia (1977) sugere a primeira tentativa de aplicação dos modelos ARIMA na análise de séries temporais relativas ao número de nascimentos num determinado período de tempo. Utilizou dados relativos ao número de nascimentos de nados vivos femininos na Noruega no período entre 1919-1974 e construiu um modelo de previsão para valores futuros. O modelo foi posteriormente testado e utilizada a metodologia *Box Jenkins*, descrita no próximo capítulo, na projeção do número total de nascimentos na Noruega para o ano de 1975. Usando os modelos estimados ARIMA (4,1,1) e ARIMA (3,1,2), Saboia (1977) depreendeu que a grande vantagem na utilização de modelos ARIMA é a possibilidade de utilização de dados apenas a médio prazo, neste caso 50 anos, para que as projeções fossem credíveis e que as previsões feitas com modelos deste tipo se ajustavam facilmente a mudanças inesperadas como, por exemplo, uma quebra no número de nascimentos devido a mudanças nos padrões de fertilidade numa determinada população. Apesar da metodologia *Box Jenkins* (1970) ter sido aplicada anteriormente por outros autores em estudos demográficos e populacionais (Lee 1974; McCleary et al ,1980) esta foi a primeira aplicação específica de modelos ARIMA no estudo da evolução da natalidade.

Posteriormente, no estudo *Time Series Analysis for Social Science* (McCleary et al., 1980) é feita uma análise de mais 50 séries temporais de áreas relacionadas com as ciências sociais e biomédicas, nas quais são ilustradas as principais propriedades das séries temporais e os procedimentos analíticos de modelação com o modelo ARIMA.

No contexto português, o INE coordena, anualmente, desde 2013 o estudo estatístico *Previsões mensais de Nados vivos e Óbitos* no qual, com recurso à utilização de métodos quantitativos de previsão de séries temporais, se antecipam os valores para o nº mensal de nados vivos para um período de 12 meses, por sexo e região NUTS III, valores esses posteriormente utilizados nas Estatísticas mensais da população portuguesa residente. Os métodos de previsão usados baseiam-se na extrapolação das características registadas das observações passadas, na utilização de modelos ARIMA com componente sazonal e na aplicação da metodologia proposta por *Box & Jenkins*.

De destacar uma segunda aplicação metodológica semelhante no estudo *Nados Vivos: Análise e Estimação* pela investigadora Teresa Bago Uva (INE) no qual o principal objetivo foi a obtenção de estimativas para o número de nados vivos ocorridos por mês em Portugal. A amostra utilizada constituiu

a série mensal dos nascimentos em Portugal entre Janeiro e 1980 e Junho de 1998 aos quais foram aplicados métodos de previsão usando modelos dinâmicos univariados (em particular modelos ARIMA e modelos estruturais).

De referir também uma outra aplicação dos modelos ARIMA pelo INE no estudo da natalidade que surgiu da projeção das taxas de fecundidade por idade num horizonte temporal de longo prazo: 2001-2050. Entende-se fecundidade como a relação existente entre o número de nados vivos e o número de mulheres em idade fértil e o modelo escolhido utiliza uma curva ajustada às distribuições da fecundidade portuguesa nos anos considerados. Para as projeções foram utilizados modelos ARIMA nos quais se selecionou um modelo, ajustaram-se os dados e se utilizou o ajustamento para produzir pontos e intervalos de previsão. Do estudo anterior concluiu-se que as projeções feitas com os modelos ARIMA através de observações históricas da série produziram bons resultados a curto prazo para as taxas de fertilidade. Contudo, devido à incerteza associada a um horizonte temporal de 50 anos, os intervalos de confiança aumentam demasiado do ponto de vista demográfico e para atenuar os efeitos da incerteza foi necessário reconsiderar o nível de confiança para 70%.

Capítulo 3. Metodologia

3.1. Metodologia de *Box-Jenkins*

Um dos principais objetivos na análise de sucessões cronológicas é descrever e encontrar um bom modelo para as relações existentes entre as observações. Pankratz (1983) refere *"a good model includes the smallest number of estimated parameters needed to adequately fit patterns in the available data"*. Assim, um bom modelo deve ter em consideração o princípio da parcimónia, isto é conter o menor número de parâmetros possível, deve ser relativamente simples e flexível para que se possa adaptar à incerteza futura e ser capaz de fazer boas previsões.

Box e Jenkins (1970), baseados nos trabalhos de Yule (1926) e Wold (1938), propuseram um processo iterativo com base em 3 etapas:

Etapa 1) Identificação Etapa 2) Estimação Etapa 3) Avaliação do diagnóstico

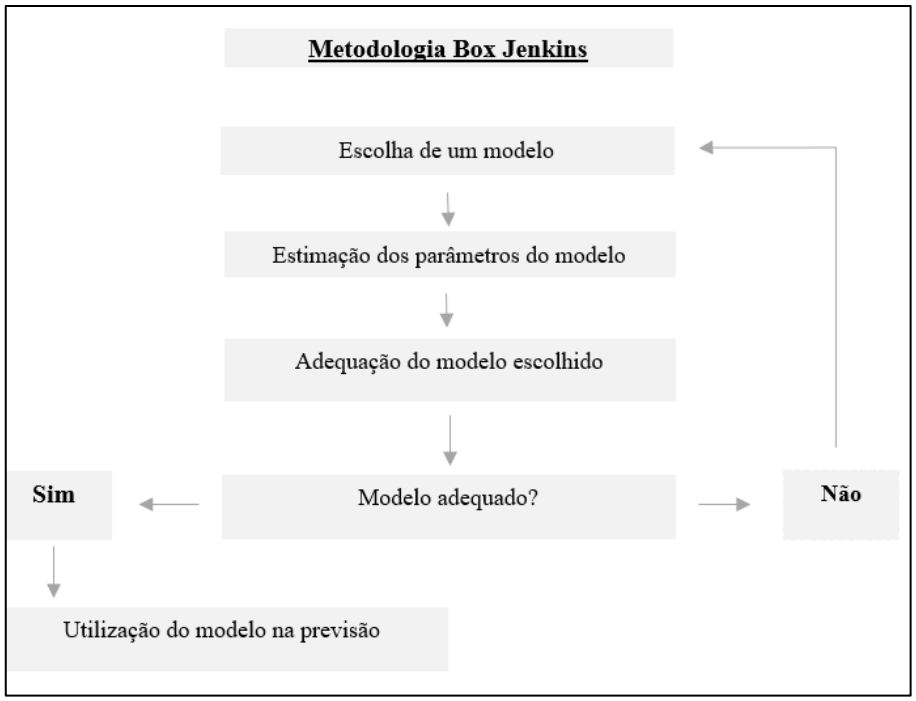


Figura 3.1. Cronograma explicativo da Metodologia *Box Jenkins*

Etapa 1) Identificação

A etapa da identificação tem como objetivo escolher um modelo e subdivide-se em duas etapas:

a) **Representação gráfica e estacionarização da sucessão cronológica;**

b) **Seleção de um modelo**

O estudo de uma série temporal deve começar pela sua **representação gráfica** para que se possa verificar a necessidade ou não de ser estacionarizada. A representação gráfica de uma sucessão cronológica é feita em coordenadas cartesianas com os valores observados no eixo das ordenadas e os instantes observados no eixo das abcissas, sendo que da união dos pontos resulta o cronograma da sucessão cronológica. Uma primeira análise do cronograma da série permite detetar o tipo de tendência, se existe instabilidade da variância/média ou a existência de movimentos periódicos, mas é insuficiente para garantir que a série é estacionária.

Testes de raiz unitária

Com vista a estudar a estacionaridade da série, aplicam-se seguidamente os **testes de raiz unitária**, também chamados testes de estacionaridade. A denominação anterior surge da ideia de que o número de diferenças necessárias para transformar X_t numa série estacionária corresponde ao nº de raízes unitárias presentes no processo gerador de X_t. De um ponto de vista geral, o que os testes de raiz unitária fazem é testar a não estacionaridade de uma série e averiguar se $\rho = 1$, ou seja, se a série a estudar possui uma raiz unitária e, por isso, é não estacionária. Para esse efeito, como descrito em baixo, testa-se a hipótese nula de que a série é não estacionária (e por isso possui uma raiz unitária) versus a hipótese de que a série é estacionária (e não possui raiz unitária).

TESTE DE RAÍZ UNITÁRIA		
$H_0: \rho = 1$	vs	$H_1: \rho \neq 1$
(série não estacionária)		(série estacionária)

Descrevem-se em seguida os dois testes de raiz unitária mais utilizados: o teste Dickey-Fuller (**DF**), e a sua generalização Dickey-Fuller Aumentado (**ADF**), e o teste Kwiatkowski-Phillips-Schmidt-Shin (**KPSS**):

Teste *Augmented Dickey Fuller* (ADF)

O teste DF parte da suposição de que os erros são independentes e identicamente distribuídos (*iid)* e considera a expressão geral,

$$X_t = \rho X_{t-1} + \beta_0 + \beta_1 t + \varepsilon_t \qquad \text{Equação 3.1}$$

À qual se substitui X_{t-1} dos dois lados da equação,

$$X_t - X_{t-1} = (\rho - 1)X_{t-1} + \beta_0 + \beta_1 t + \varepsilon_t \qquad \text{Equação 3.2}$$

O teste de raiz unitária DF obtém-se considerando,

$H_0: \rho = 1$ ($a\ série$ não é $estacionária$)

$H_1: \rho < 1$ ($a\ série$ é $estacionária$)

No caso em que ε_t não é um ruído branco, utiliza-se uma correcção e introduz-se o teste ADF, no qual

se aumenta a regressão, adicionando termos suficientes em ΔX_{t-1} aumentando o desfasamento na equação como forma de eliminar a autocorrelação,

$$\Delta X_t = (\rho - 1)X_{t-1} + \sum_{i=1}^{k} \beta_i \Delta X_{t-i} + \beta_0 + \beta_1 t\varepsilon_t \qquad \text{Equação 3.3}$$

Assim, a principal vantagem do teste ADF em relação ao teste DF é garantir que os resíduos não apresentam autocorrelação. Em alternativa ao teste anterior, e para as mesmas hipóteses a serem testadas, aplica-se o teste Phillips-Perron (**PP**), no qual se permite que os erros sejam correlacionados.

Teste KPSS

O teste *KPSS*, criado por Kwiatkowski et al. (1992) inverte a hipótese nula do teste ADF, ou seja, no caso de a hipótese nula ser rejeitada a série temporal é não estacionária. Assim,

H_0: *A série é estacionária* vs.

H_1: *A série é não estacionária*

A estatística de teste é dada por:

$$LM = \sum_{i=1}^{t} \frac{{S_t}^2}{n^2 \hat{\sigma}^2} \qquad \text{Equação 3.4}$$

Sendo

$$S_t = \sum_{i=1}^{t} \varepsilon_i \text{ e } \hat{\sigma}^2 \qquad \text{Equação 3.5}$$

Um estimador para a variância dos erros.

O critério de rejeição da hipótese nula: Rejeitar H_0 se LM_{KPSS} > valores críticos.

Resumidamente, os testes descritos acima são de dois tipos:

No primeiro caso surgem os testes ADF e PP cuja hipótese nula é testar a presença de uma raiz unitária (a série é não estacionária), e quando esta hipótese não é rejeitada fornecem-se informações sobre o número de diferenciações necessárias para que a série seja estacionária;

No segundo caso surge o teste KPSS cuja hipótese nula é a da estacionaridade da série. Contudo, de referir que deve ser utilizado mais do que um teste em simultâneo para avaliar a estacionaridade da série.

No caso em que a série temporal for identificada como não estacionária aplica-se o operador diferenciação simples com vista a obter uma série diferenciada da série original e testam-se novamente os testes de raiz unitária. O processo de diferenciação e aplicação dos testes é repetido o número de vezes necessário, em geral no máximo duas vezes, até que a série final seja estacionária.

Após a estacionarização da série, seleciona-se um modelo através da comparação das principais FAC e

FACP, para verificar se há necessidade de se aplicarem mais transformações, sendo este um momento crucial na modelação da série. Numa primeira fase deve observar-se o comportamento geral e não prender a atenção em eventuais pormenores, uma vez que o modelo pode ser melhorado numa fase posterior de avaliação do diagnóstico.

A Tabela 3.1 resume as características principais das FAC e FACP das três grandes classes de processos estacionários: AR, MA e ARMA

Tabela 3.1. Principais características das FAC e FACP teóricas dos processos estacionários não sazonais

Processo	FAC	FACP
AR	Decaimento para zero sob forma exponencial ou sinusoidal amortecida	Decaimento brusco para zero a partir de um certo lag k
MA	Decaimento brusco para zero a partir de um certo *lag k*	Decaimento para zero sob forma exponencial ou sinusoidal amortecida
ARMA	Decaimento para zero sob forma exponencial ou sinusoidal amortecida	Decaimento para zero sob forma exponencial ou sinusoidal amortecida

Fonte: Murteira et al. (1993).

Etapa 2) Estimação

Uma vez identificado o modelo, segue-se a fase de estimação dos seus parâmetros. Esta fase exige uma participação pouco ativa por parte do analista porque existem disponíveis softwares estatísticos adequados, neste caso o **R**, que fornecem uma estimativa adequada dos parâmetros a estimar através da aplicação de cálculos computacionais de alguma complexidade. Indicam-se, a título de exemplos, os três métodos mais utilizados: o método dos momentos, dos estimadores de mínimos quadrados ou dos estimadores de máxima verosimilhança.

Etapa 3) Avaliação do Diagnóstico

Após identificar o modelo e estimar os parâmetros necessários, inicia-se a etapa de avaliação da qualidade estatística do modelo e da qualidade de ajustamento. Se o modelo escolhido não for adequado deve retomar-se à fase de identificação e repetir o processo até que se encontre um modelo satisfatório e, se possível, usá-lo para fazer previsões.

Avaliação da qualidade estatística

Para avaliar a qualidade estatística do modelo ajustado estuda-se a significância estatística dos parâmetros, ou seja, tendo em conta o princípio da parcimónia verifica-se se um determinado parâmetro é significativamente diferente de zero e, caso não aconteça, procede-se à sua eliminação.

Para concluir se um parâmetro é ou não significativo para o modelo, aplica-se o seguinte teste de hipóteses:

$$H_0: \beta_i = 0 \; vs \; H_1: \beta_i \neq 0$$

Em que a estatística de teste é:

$$T = \sqrt{N}\frac{B_i}{\sqrt{var_R}(B)} \sim t_{N-p-q} \qquad \text{Equação 3.6}$$

Rejeita-se H_0 para um nível de rejeição α sempre que $|T| > t_{N-p-q,\alpha/2}$ e a não rejeição de H_0, para um nível de significância α, implica a eliminação do parâmetro.

Avaliação da qualidade do ajustamento

A **análise dos resíduos** permite medir a qualidade de ajustamento de um modelo estimado. Após a fase de modelação de uma série temporal, o cálculo dos resíduos pode ser feito através da diferença entre os valores observados e os valores estimados correspondentes, ou seja,

$$e_t = X_t - \hat{X}_t \qquad \text{Equação 3.7}$$

Segundo Box e Jenkins (1970) se o modelo for adequado para descrever a sucessão cronológica, com o aumento do número de observações os resíduos aproximam-se do ruído branco ε_t,

$$\hat{\varepsilon} = \varepsilon_t + O\left(\frac{1}{\sqrt{N}}\right) \qquad \text{Equação 3.8}$$

Um modelo corretamente ajustado a uma série temporal deve gerar resíduos com o comportamento idêntico ao de um ruído branco e, portanto, devem apresentar média nula e satisfazer o pressuposto da não correlação. Assim. as autocorrelações dos resíduos podem ainda ser avaliadas graficamente através da observação da FAC do modelo que deve apresentar o comportamento semelhante ao da FAC de um ruído branco, ou seja autocorrelações não significativamente diferentes de zero. Por outro lado, se se pretender avaliar teoricamente a qualidade de ajustamento de um modelo e verificar se os resíduos estimados são não correlacionados pode utilizar-se o **teste de Box-Pierce** ou o **teste de Ljung Box**.

Teste de *Box-Pierce*

Considerando $H_0: \rho_1 = \rho_2 = \cdots = \rho_m = 0$

A estatística de Box-Pierce define-se:

$$Q(m) = N \sum_{k=1}^{m} \hat{\rho}_k \qquad \text{Equação 3.9}$$

Onde N é o nº total de observações, $\hat{\rho}_k$ é a autocorrelação dos resíduos no lag *k* e *m* o nº de lags a testar, aproxima-se de uma distribuição Qui-Quadrado com m graus de liberdade para grandes amostras e rejeita-se H_0 para um nível de significância α quando $Q > X^2_{1-\alpha,m}$

Teste de *Ljung-Box*

O teste de Ljung-Box desenvolvido por Ljung-Box (1978) é uma variante da estatística Q de Box-Pierce e cuja estatística de teste Q é definida por:

$$LB(m) = N(N+2) \sum_{k=1}^{m} \frac{\hat{\rho}_h^2}{N-h} \qquad \text{Equação 3.10}$$

Aproximando-se também de uma distribuição Qui-Quadrado com m graus de liberdade. Comparativamente ao teste anterior, a estatística de teste utilizada converge mais rapidamente, sendo por isso preferível quando a série a ser estudada não é muito grande.

Em ambos os testes a hipótese nula será rejeitada se o valor da estatística de teste for superior ao valor da distribuição Qui quadrado para o nível de significância escolhido e graus de liberdade m. A rejeição da hipótese nula implica que as autocorrelações para diferentes valores de m (lags) poderão ser diferentes de zero e os valores não são aleatórios e independentes, logo há necessidade de reformular o modelo inicialmente escolhido.

Na fase de previsão, para construir intervalos de previsão, torna-se importante verificar também se os resíduos têm variância constante e apresentam uma distribuição Normal. (Hyndman & Athanasopoulos,2014). A condição de normalidade pode ser avaliada através de testes estatísticos ou pela representação gráfica (análise do histograma ou do *QQ-plot*). No caso em que existe normalidade, o histograma deve assemelhar-se ao comportamento da função densidade de uma distribuição Normal e a maioria dos pontos representados no *QQ-plot* devem posicionar-se sobre uma recta.

Critérios de seleção de modelos

Da análise anterior é possível resultar mais do que um modelo que descreva de forma suficientemente satisfatória a série temporal em análise e será necessário escolher de entre os modelos propostos qual o que melhor se ajusta. Saliente-se ainda que nem sempre a escolha do melhor modelo é imediata e prática aconselha a que se estudem e comparem modelos alternativos que satisfaçam os critérios de seleção, também designados Critérios de Informação, indicados em seguida.

Critério Akaike

Akaike (1973/74) definiu a grandeza AIC para avaliar a qualidade de ajustamento de um modelo,

$$\mathrm{AI}C(m) = -2\,ln(\mathrm{L}) + 2m \qquad \text{Equação 3.11}$$

Sendo *L* a função de máxima verosimilhança e *m* o nº de parâmetros do modelo a ser estimado. Em linha com o que será desenvolvido mais à frente, deverá selecionar-se como melhor modelo o que tiver menor valor de AIC associado.

Critério BIC

O Critério BIC, Critério de Informação Bayesiano, e é definido por:

$$BIC(m) = -2\,ln(\mathrm{L}) + m\,ln(n) \qquad \text{Equação 3.12}$$

Em que *n* representa o nº total de observações do modelo a ser ajustado. Em ambos os casos, a escolha do melhor modelo deve recair no modelo que conduza ao valor mínimo dos critérios AIC e BIC.

Apresentaram-se até aqui de forma sucinta os conceitos mais importantes subjacentes à metodologia de modelação com modelos ARIMA. Seguidamente, uma vez encontrado um modelo satisfatório à luz dos critérios considerados, pode utiliza-se o mesmo na fase de previsão, esperando que o modelo escolhido conduza a erros de previsão inferiores aos de modelos alternativos.

3.1.1. Previsão com séries temporais

Um método de previsão é definido como um procedimento que permite prever o comportamento esperado dos dados a médio ou longo prazo a partir de valores passados e presentes. Wheelwrigth (1998) afirma que a maioria dos métodos de previsão se baseiam na premissa de que as observações passadas contêm a informação sobre o padrão de comportamento da série, padrão esse recorrente no tempo. Chatfield (2004) classifica genericamente os **métodos de previsão** em 3 tipos:

- Métodos subjetivos: que derivam das capacidades subjetivas, intuitivas e conhecimento prévio do analista;

- Métodos univariados: em que as previsões dependem de valores passados de uma série temporal; Serão os métodos utilizados e baseiam-se em exclusivo na própria série a prever e em modelos construídos com esse pressuposto;

- Métodos multivariados: em que as previsões feitas para uma variável dependem, pelo menos em parte, de uma ou mais variáveis explicativas ou preditoras de uma série adicional.

A escolha do método de previsão adequado, tal como do modelo subjacente, depende de um conjunto de indicadores como: o objetivo da previsão, o modelo subjacente identificado, a existência de sazonalidade/tendência, a dimensão da amostra a estudar, o horizonte de previsão, a experiência do analista ou a disponibilidade ou a capacidade preditiva do *software* utilizado. Também segundo Wheelwright (1998) nem sempre os métodos mais complexos de previsão conduzem a melhores resultados e será necessário avaliar a adequabilidade de cada um antes de se iniciar uma previsão.

Um dos objetivos principais na análise de séries temporais é a previsão estatística, ou seja, a previsão do comportamento futuro de uma série tendo em conta o seu comportamento até ao instante *t*.

Considere-se uma série temporal com observações registadas até ao instante *t*, $\{X_t, X_{t-1}, X_{t-2}, \dots\}$ e que com base nestas observações se pretende prever o valor no momento $t+m$ para $m>0$, X_{t+m}. Designando por $X_t(m)$ o preditor de X_{t+m} então,

$$X_t(m) = f(X_t, X_{t-1}, X_{t-2}, \dots) \qquad \text{Equação 3.13}$$

Em que *t* designa a origem de previsão e *m* o horizonte temporal da previsão.

3.1.1.1. Medidas de desempenho

A função de previsão é escolhida com base no **critério de minimização do erro quadrático médio** entre X_{t+m} e $X_t(m)$, definido por,

$$E\,[(X_{t+m} - X_t(m))^2]$$

Em que o melhor preditor de X_{t+m} com base no erro quadrático médio é a **esperança condicional** definida por,

$$X_t(m) = E\,[X_{t+m} \mid X_t, X_{t-1}, X_{t-2}, \dots] \qquad \text{Equação 3.14}$$

Na prática considera-se uma das seguintes condições:

i. X_{t+m} , X_t , X_{t-1},X_{t-2}, ... têm distribuição conjunta Normal, a esperança condicional torna-se fácil de calcular e o melhor preditor de $X_t(m)$ é a função linear de X_t , X_{t-1},X_{t-2}, ...

ii. Admite-se que o preditor tem a forma de funções lineares de X_t , X_{t-1},X_{t-2}, ...

$$X_t(m) = \alpha_0 + \alpha_1 X_t + \alpha_2 X_{t-1} + \cdots \qquad \text{Equação 3.15}$$

E procuram-se os valores de $\alpha_0, \alpha_1, \alpha_2$ ... que minimizam o erro quadrático médio.

Para medir a precisão e eficácia dos resultados obtidos e dos métodos de previsão utilizados, utilizam-se **medidas de desempenho.** Na Tabela 3.2 sumarizam-se as medidas de desempenho destacadas na bibliografia consultada.

Tabela 3.2. Resumo descritivo das principais medidas de desempenho

Medidas de desempenho	
Erro de previsão	$e_t = X_t - \hat{X}_t$
Erro médio (*ME – Mean Error***)**	$\frac{1}{n}\sum_{t=1}^{n} e_t\ (h)$
Erro absoluto médio *(MAE – Mean Absolute Error)*	$\frac{1}{n}\sum_{t=1}^{n} \lvert e_t\ (h)\rvert$
Erro quadrático médio *(MSE –Mean Squared Error)*	$\frac{1}{n}\sum_{t=1}^{n} e_t^2(h)$
Raíz do Erro quadrático médio *(RMSE - Root Mean Squared Error)*	$\sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2(h)}$
Erro médio absoluto em percentagem (*MAPE – Mean absolute percentage error***)**	$100 \times \frac{1}{m}\sum_{i=1}^{n} \frac{\lvert X_t - \hat{X}_t\rvert}{\lvert X_t\rvert}$

Com o objetivo de avaliar o desempenho global de um modelo de previsão, torna-se essencial usar uma medida de erro, como as descritas acima, capaz de traduzir numericamente a capacidade preditiva do modelo. Servem ainda para medir a precisão e a eficácia dos resultados obtidos através dos diversos métodos de previsão, sendo que quanto menor o valor obtido no cálculo das medidas de desempenho, melhor será a capacidade preditiva associada ao método de previsão.

3.1.1.2. Outros métodos de previsão

Existem ainda outros métodos de previsão a utilizar, cuja opção de utilização depende da natureza dos dados usados. Na tabela seguinte resumem-se os principais métodos de previsão alternativos e as suas principais vantagens e limitações.

Tabela 3.3. Resumo descritivo dos principais métodos de previsão alternativos

Método	Descrição	Cálculo da previsão	Vantagens e Desvantagens
Médias Móveis Simples	▪ Consiste em calcular a média das últimas r observações ▪ $\hat{X}_t = \frac{X_t + X_{t-1} + \cdots + X_{t-r+1}}{r}$, para t=1,…N ▪ N designa-se janela de previsão; ▪ Se a série apresentar pouca variação deve ser utilizada uma janela de previsão menor; ▪ O conceito de "média móvel" consiste na atualização permanente da médias das observações, desprezando observações que se encontrem fora da janela de previsão estabelecida;	▪ $\hat{X}_t(h) = \hat{X}_t, \forall h > 0$	▪ Método flexível e de simples aplicação; ▪ Aplicável para um número reduzidos de observações; ▪ Só pode ser utilizado para séries estacionárias; ▪ As observações têm pesos iguais no cálculo da média; ▪ Dificuldade em determinar r;
Alisamento Exponencial Simples	▪ Consiste na aplicação de uma média ponderada nas diversas observações da ST em estudo; São atribuídos pesos diferentes às observações (as observações mais antigas têm pesos inferiores relativamente a observações mais recentes); ▪ $\hat{X}_t = \alpha X_t + \alpha(1-\alpha)X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + \cdots para$ t=1,…N e $\alpha \epsilon$ [0,1] ▪ Quanto mais próximo de 1 for o valor da constante de suavização, designada por α, maior é o peso das observações mais recentes, e mais sensível a mudanças será; Quando mais próximo de 0, menor o ajuste, maior o peso dado a observações mais antigas e mais estável será a previsão;	▪ $\hat{X}_t(h) = \hat{X}_t, \forall h > 0$	▪ Método flexível, rápido e de simples aplicação; ▪ Dificuldade em determinar α; ▪ Não é indicado para séries que apresentem tendência/ sazonalidade;
Holt-Winters	▪ $\hat{X}_t = \alpha X_t + (1-\alpha)\left(X_{t-1} + \hat{T}_{t-1}\right)$ ▪ $\hat{T}_t = \beta(\hat{X}_t - \hat{X}_{t-1}) + (1-\beta)\,\hat{T}_{t-1}$ $onde\ t = 2, \ldots, N, \alpha \epsilon$ [0,1]$e\ \beta \epsilon$ [0,1] ▪ O método Holt-Winters Sazonal é utilizado em alternativa para séries que apresentem tendência e sazonalidade;	▪ $\hat{X}_t(h) = \hat{X}_t + h\hat{T}_t,\ \forall h > 0$	▪ Semelhante aos métodos anteriores, mas aplicável a séries com tendência e sazonalidade; ▪ Dificuldade em determinar os parâmetros α e β;

Segue-se a descrição dos dados utilizados na análise e a explicitação do software utilizado

3.2. Dados utilizados

A maioria dos estudos demográficos envolvem a recolha, análise e interpretação de dados sobre uma determinada população alvo. Em Portugal, é o INE que centraliza a produção, elaboração e divulgação de séries demográficas, nomeadamente relativas à natalidade. Contudo, a origem dos dados advém das Conservatórias do Registo Civil que enviam semanalmente ao INE os verbetes estatísticos relativos aos nados vivos ocorridos em Portugal no período em análise.

Neste trabalho o objeto de estudo são três séries temporais distintas que contabilizam o número de nascimentos Prematuros, de Baixo Peso e de Muito Baixo Peso[1] ocorridos em Portugal entre 1989 e 2019 e se designam, respetivamente, por NP, NBP e NMBP

O INE publica anualmente o documento *Estatísticas Demográficas* no qual é feita uma análise global da situação demográfica atual em Portugal (Instituto Nacional de Estatística [INE], 2020). Os dados referentes ao Nº de Nados Vivos Prematuros e ao Nº de Nados Vivos de Baixo Peso registados em Portugal nos anos de 2001 a 2019 foram retirados, respetivamente, dos documentos *Estatísticas Demográficas* de 2006 a 2019 (INE, 2020).

O INE divulga ainda anualmente o documento *Indicadores Sociais*, acompanhado de uma base de dados adicional, que permite construir um retrato social da população portuguesa através da análise das principais variáveis e indicadores de carácter social (INE, 2012). Os dados referentes ao Nº de Recém-Nascidos de Muito Baixo Peso foram retirados da base de dados referida, tendo-se selecionado para análise o indicador *Nados Vivos (nº) por local de residência da mãe, sexo, grupo etário da mãe e escalão de peso à nascença*, seguido do período de referência 2001 a 2013 e o escalão de peso à nascença.

As séries temporais construídas a partir dos dados descritos anteriormente e que serviram de ponto de partida para a análise são séries temporais univariadas discretas, cujo período de referência é o ano civil e nas quais foram incluídas observações igualmente espaçadas no tempo que estão disponíveis para consulta no Anexo G. Vale ressalvar que os modelos construídos a partir das séries temporais utilizam apenas informação relativa a estas mesmas séries, não incluindo nenhum fator explicativo extra ou série temporal secundária.

3.3. Software utilizado

O *R* é uma linguagem de programação amplamente utilizada como ferramenta de análise de dados e no desenvolvimento de *software* estatístico. Foi criada por Ross Ihaka e Robert Gentleman no departamento de Estatística da Universidade de Auckland no início dos anos 90. Tem uma versão de instalação base onde para além das funcionalidades originais é possível instalar livrarias, em inglês *libraries*, que enriquecem a análise. Optou-se pela utilização da linguagem R, em detrimento de outras linguagens de programação, pelas seguintes razões: *i)* variabilidade dos pacotes de análise estatística disponíveis, *ii)* flexibilidade e rapidez na execução e *iii)* eficiência na manipulação, análise e visualização da informação.

[1] De relembrar que foram considerados nados vivos prematuros aqueles nascidos antes das 37 semanas de gestação, nados vivos de baixo peso com pesos no intervalo de valores]1500g, 2500g[e nados vivos de muito baixo peso no intervalo de valores]500g,1500g[.

Junta-se no Anexo R uma tabela na qual se sumarizam os principais *packages* e respetivas funções aplicadas na análise dos Resultados, alguns desses comandos são específicos para a análise de séries temporais e outros de âmbito geral. Em particular, no código R desenvolvido para a análise dos Resultados foram utilizados, para além das funcionalidades do programa original, o package *stats*, que contém um conjunto de funções de análise estatística e o package *tseries* indicado especificamente na análise de séries temporais e instalado posteriormente.

Para que os dados referidos na secção possam ser lidos pelo *R*, é necessário que sejam criados, primeiramente, dois *data frames* para cada uma das séries em análise, nos quais o primeiro *data frame* contém os anos em que é feita a análise e o segundo os valores registados para cada ano. Os nomes escolhidos para os *data frames* seguem a lógica do nome de cada uma das três séries a analisar: *tabelaNascimentosP, tabelaNascimentosBP e tabelaNascimentosMBP*. Só após esta fase preparatória os dados estão em condições de serem representados graficamente.

Uma vez apresentados os conceitos teóricos e as etapas metodológicas, são apresentados no capítulo 4 os Resultados obtidos que, para simplificar a sua compreensão, são sintetizados em gráficos e tabelas.

Capítulo 4. Resultados

4.1. Representação gráfica e Análise descritiva

A etapa inicial do estudo empírico passa pela caracterização das séries temporais descritas anteriormente. Os gráficos em seguida constituem o primeiro passo na análise dos dados recolhidos e representam o nº de nados vivos Prematuros, de Baixo Peso e de Muito Baixo Peso nos anos de 1989 a 2019, referidos futuramente por NP, NBP e NMBP.

Entre 1989 e 2019 foram registados 3 225 296 nascimentos em Portugal dos quais 253 786 corresponderam a nados vivos prematuros. Uma primeira análise visual sugere um comportamento pouco regular da série NP com um decrescimento acentuado dos valores nos primeiros 12 anos[2]. Destaca-se a variação entre o valor máximo de NP registado em 1989 (14 928 nados vivos) e o valor mínimo em 2001 (6 346), valores esses que contabilizam 12,6% e 5,6% do nº total de nascimentos no ano considerado. A média dos NP no período considerado foi de 8 187 e a mediana de 7 391 nascimentos. Uma primeira análise visual não sugere qualquer comportamento sistemático em termos da variância ou a existência de padrões sazonais, de tendência ou de movimentos oscilatórios.

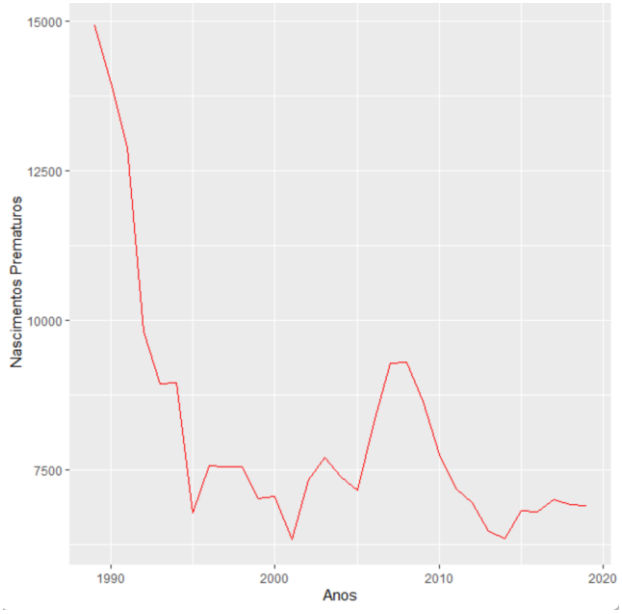


Figura 4.1. Representação gráfica dos Nados vivos Prematuros em Portugal no período de 1989 a 2019

No mesmo período, do nº total de nascimentos em Portugal, 234 762 foram de nados vivos de baixo peso. O valor máximo de NBP registou-se no ano de 1999 (8 568 nados vivos) e o valor mínimo em 1992 (6 165 nados vivos), valores esses que correspondem a 7,4% e 5,4% do nº total de nascimentos no ano considerado. A média dos NBP no período considerado foi de 7 573 e a mediana de 7 667 nascimentos. Uma primeira análise visual não sugere qualquer comportamento sistemático em termos da variância ou a existência de padrões sazonais ou de movimentos oscilatórios.

[2] A variação brusca observada no período entre 1989 e 1993 deve-se à alteração na definição de prematuridade que até aos anos 90 para além da IG incluía também todos os nados vivos nascidos com menos de 2500g

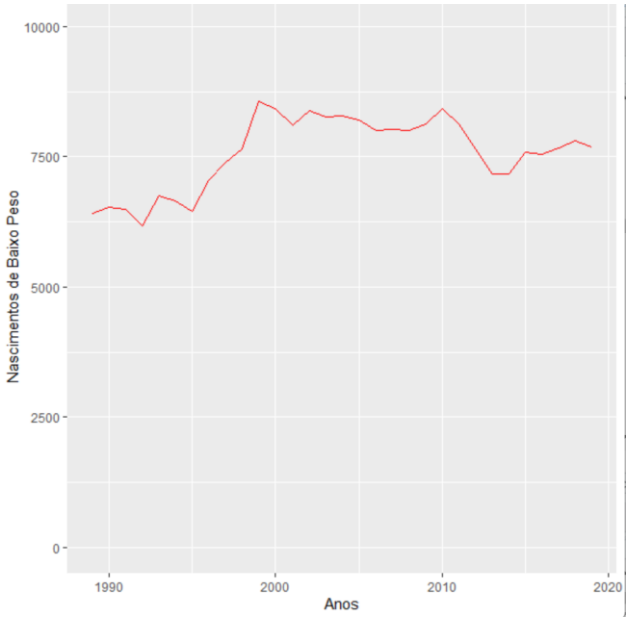


Figura 4.2. Representação gráfica dos Nados vivos de Baixo Peso em Portugal no período de 1989 a 2019

No mesmo período, do Nºtotal de nascimentos em Portugal, 30 300 foram de nados vivos de muito baixo peso. O valor máximo de NMBP registou-se em 2012 (1 148 nados vivos) e o valor mínimo em 1989 (611 nados vivos). A média dos NMBP no período considerado foi de 997 e a mediana de 1 020 nascimentos. Uma primeira análise visual não sugere qualquer comportamento sistemático em termos da variância ou a existência de padrões sazonais ou de movimentos oscilatórios.

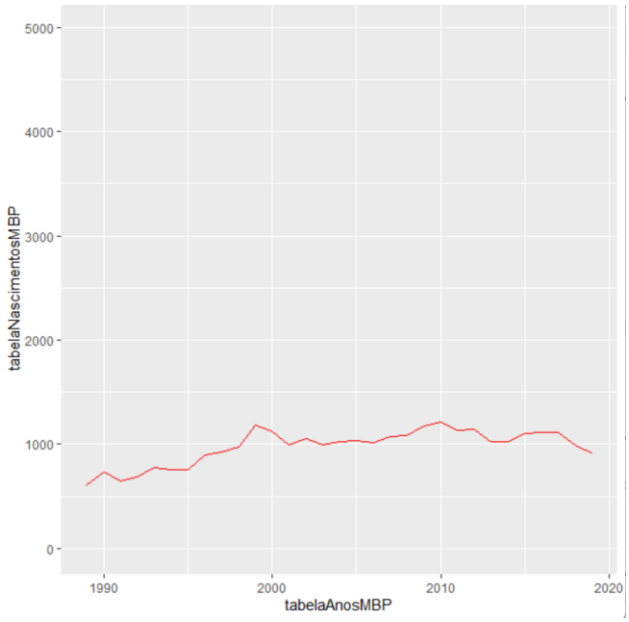


Figura 4.3. Representação gráfica dos Nados vivos de Muito Baixo Peso em Portugal no período de 1989 a 2019

Tabela 4.1. Estatística Descritiva das séries NT, NP, NBP e NMBP para o período de 1989 a 2019

série	*N*	*Min*	*1Q*	*Mediana*	*Média*	*3Q*	*Max*
Nasc.Totais	3 225 296	82 367	93 349	109 287	104 042	113 770	119 455
NP	253 786	6 346	6 942	7 391	8 187	8 797	14 928
NBP	234 762	6 165	7 107	7 667	7 573	8 130	8 568
NMBP	30 300	611	902	1 020	977	1 111	1 215

Os histogramas apresentados em baixo são a representação gráfica da distribuição de frequências de nascimentos nas três séries estudadas, em que no eixo horizontal os dados foram agrupados em classes que variam consoante a série a analisar e o eixo vertical mede a frequência de nascimentos.

Nenhum dos histogramas apresentados na Figura 4.4 indicia que as séries consideradas sejam normalmente distribuídas. O histograma da série NP sugere um enviesamento à direita, ou seja, uma assimetria positiva dos dados, em que a classe]6000,8000[é a que apresenta maior frequência relativa. Os histogramas das séries NBP e NMBP sugerem, pelo contrário, um enviesamento à esquerda, ou seja uma assimetria negativa dos dados, sendo que do histograma da série NBP se destaca a classe]8000,8500[e do histograma da série NMBP as classes]1000,1100[e]1100,1200[.

Pela análise dos *QQplots* e dos resultados do teste *Shapiro-Wilk*, no Anexo I, verifica-se que é rejeitada a hipótese nula da normalidade, pelo que existem evidências estatísticas para admitir a não normalidade da distribuição de valores das séries consideradas.

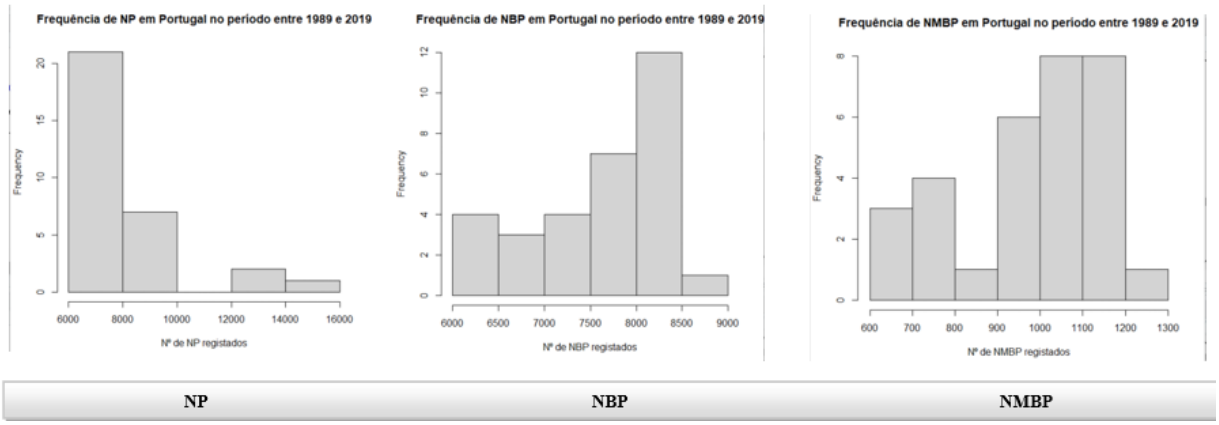


Figura 4.4. Histogramas representativos da frequência de NP, NBP e NMBP em Portugal no período de 1989 a 2019

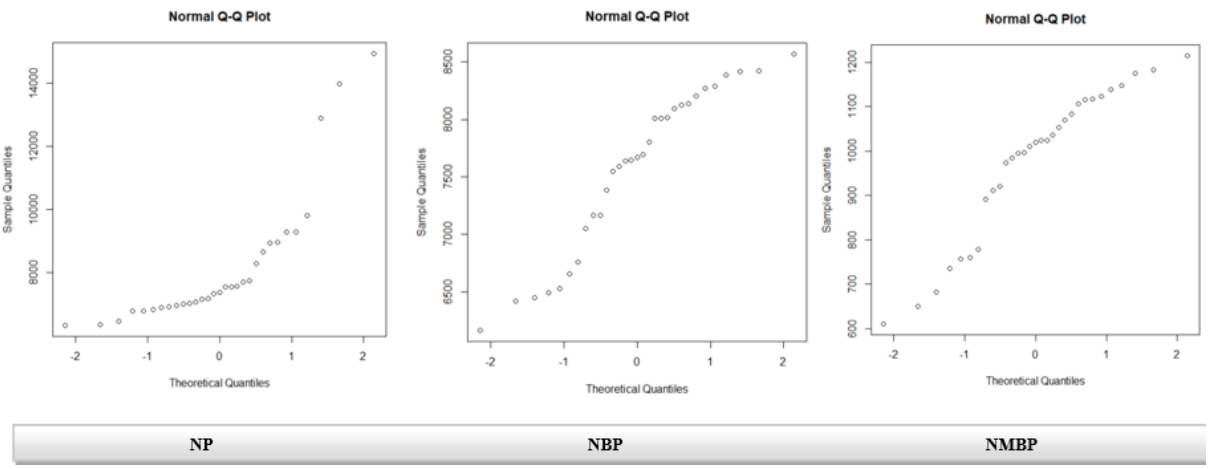


Figura 4.5. Gráficos QQPlot representativos das séries NP, NBP e NMBP

4.1.1. Análise das FAC e FACP das séries originais

Como visto previamente na metodologia, as FAC e FACP (ou, em inglês, *ACF* e *PACF*) são dois dos conceitos mais importantes na fase de identificação, considerando que estas têm um comportamento idêntico às FAC e FACP do modelo teórico que deu origem à série temporal. A par da inspeção visual já efetuada à série e para uma identificação preliminar dos modelos que deram origem às séries consideradas, observa-se em seguida o comportamento das FAC e FACP das observações originais. As bandas horizontais a tracejado representam os limites de confiança a um nível de 95%. De relembrar que as conclusões retiradas pela análise das FAC e FACP são unicamente indicações que serão confirmadas, ou não, nas fases seguintes.

4.1.1.1. Série NP

A FAC da série original NP apresenta decaimento exponencial nos 4 primeiros lags, o que sugere que a média das observações é não estacionária. Os valores empíricos são superiores ao limite a tracejado nos 2 primeiros lags, a partir do lag 5 assumem valores negativos não inferiores a -0,2 e não ultrapassam os limites definidos.

A FACP apresenta um padrão irregular, um valor muito superior ao limite definido a tracejado para o *lag=1* e um decaimento brusco para zero a partir desse mesmo *lag*, o que é indicativo de um processo autoregressivo de ordem 1. Nenhum dos valores dos lags registados sugere a existência de sazonalidade nem de nenhuma tendência aparente.

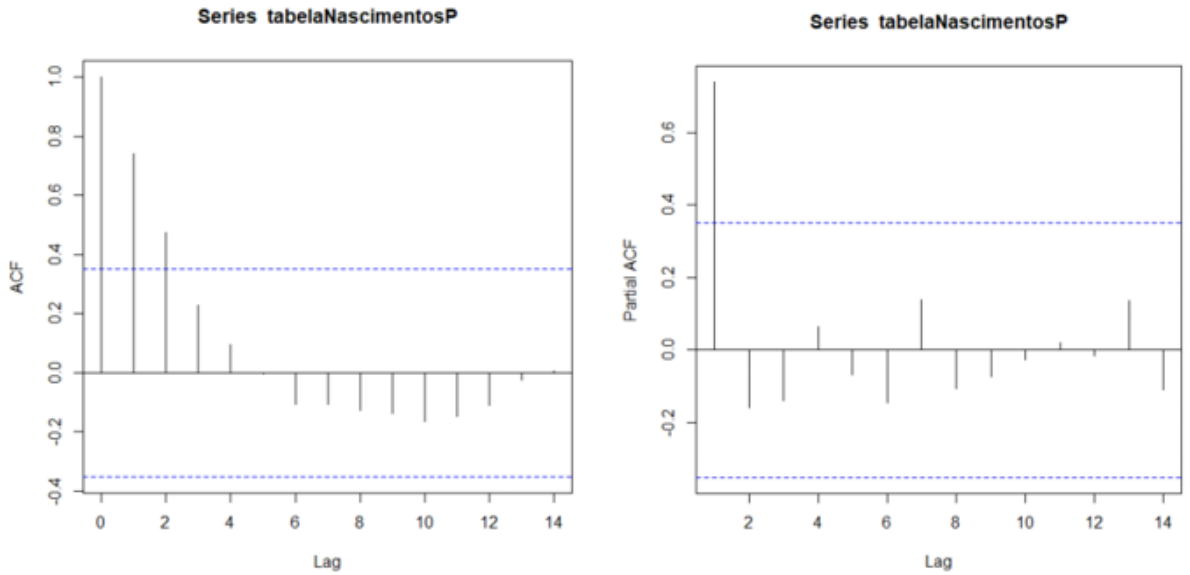


Figura 4.6. Representação das FAC e FACP da série original NP

Tabela 4.2. Valores empíricos das FAC e FACP da série original NP, representando-se com um * os valores que ultrapassam os limites a tracejado

	Série NP														
lag	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
FAC	1.000*	0.740*	0.475*	0.228	0.095	-0.003	-0.107	-0.108	-0.126	-0.136	-0.163	-0.149	-0.111	-0.024	0.006
FACP	-	0.740*	-0.159	-0.141	0.062	-0.068	-0.144	0.139	-0.108	-0.075	-0.026	0.019	-0.017	0.136	-0.109

4.1.1.2. Série NBP

A FAC da série original NBP apresenta decaimento exponencial lento nos 6 primeiros lags para zero, o que sugere que a média das observações é não estacionária. Os valores empíricos são superiores ao limite a tracejado nos 3 primeiros lags, a partir do lag 7 assumem valores negativos e decaem progressivamente, não ultrapassando para os valores negativos o limite definido a tracejado.

A FACP apresenta um padrão irregular e um valor muito superior ao limite definido a tracejado para o *lag=1*, o que é indicativo de um processo autoregressivo de ordem 1. Regista-se para o lag = 4, um valor negativo superior ao limite defindo, não sendo este indicativo de sazonalidade.

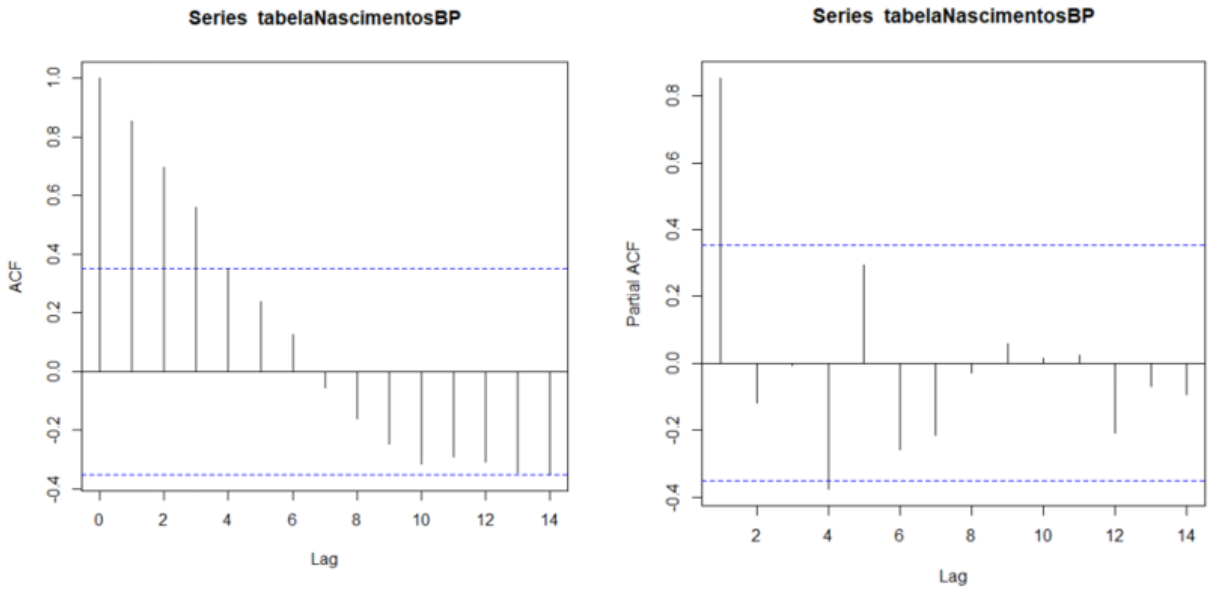


Figura 4.7. Representação das FAC e FACP da série original NBP

Tabela 4.3. Valores empíricos das FAC e FACP da série original NBP, representando-se com um * os valores que ultrapassam os limites a tracejado

Série NBP															
lag	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
FAC	1.000*	0.853*	0.695*	0.560*	0.349	0.238	0.125	-0.056	-0.162	-0.246	-0.314	-0.292	-0.308	-0.347	-0.348
FACP	-	0.853*	-0.117	-0.008	-0.375*	0.293	-0.257	-0.214	-0.027	0.060	0.014	0.026	-0.207	-0.069	-0.093

4.1.1.3. Série NMBP

A FAC da série original NMBP apresenta decaimento exponencial lento e progressivo nos 7 primeiros lags, o que sugere que a média das observações é não estacionária. Os valores empíricos são superiores ao limite a tracejado nos 4 primeiros lags, a partir do lag 8 assumem valores negativos não inferiores a -0,2 e não ultrapassando os limites definidos.

A FACP apresenta um padrão irregular e um valor muito superior ao limite definido a tracejado para o *lag=1*, o que é indicativo de um processo autoregressivo de ordem 1. Nenhum dos valores dos lags registados sugere a existência de sazonalidade nem de tendência.

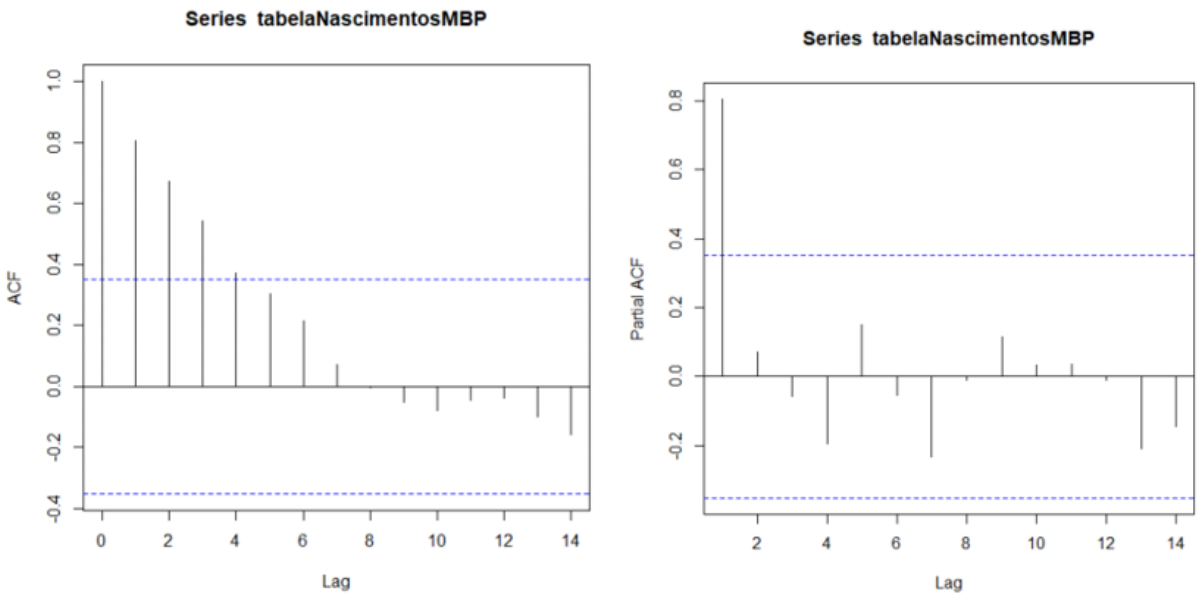


Figura 4.8. Representação das FAC e FACP da série original NMBP

Tabela 4.4. Valores empíricos das FAC e FACP da série original NMBP, representando-se com um * os valores que ultrapassam os limites a tracejado

	Série NMBP														
lag	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
FAC	1.000	0.806*	0.674*	0.542*	0.371*	0.302	0.213	0.072	-0.004	-0.052	-0.079	-0.047	-0.037	-0.101	-0.157
FACP	-	0.806*	0.070	-0.056	-0.194	0.150	-0.054	-0.234	-0.010	0.114	0.035	0.035	-0.010	-0.210	-0.144

4.2. Estacionaridade e diferenciação

4.2.1. Testes de raíz unitária

Segue-se o resumo dos resultados dos testes de raíz unitária efetuados a cada uma das séries temporais. Aplicou-se o operador da diferenciação quando necessário, ou seja, quando se verifica que a série é não estacionária, estando os resultados da diferenciação disponíveis para consulta no Anexo H.

Verifica-se que para a série original NP, a hipótese nula de que existe uma raíz unitária é rejeitada para os testes ADF e PP a um nível de significância de 1%. Assim, conclui-se que a série NP é estacionária não havendo necessidade de ser diferenciada.

Tabela 4.5. Aplicação dos testes de raiz unitária à série de Nascimentos Prematuros (NP)

Série NP	X_t
ADF	-4,2598***
PP	-4,1099***
KPSS	0,53035

* rejeita-se H_0 para o nível de significância de 10%
** rejeita-se H_0 para o nível de significância de 5%
*** rejeita-se H_0 para o nível de significância de 1%

No caso da série original de NBP, os valores da estatística de teste são sempre superiores para qualquer nível de significância considerado logo a série original é não estacionária. Segue-se a primeira diferenciação e a aplicação dos testes de raiz unitária à série diferenciada. A hipótese nula de que existe raiz unitária é rejeitada para os testes ADF e PP a um nível de significância de 1%. Conclui-se que a série de NBP é estacionária para a primeira diferença.

Tabela 4.6. Aplicação dos testes de raiz unitária à série de Nascimentos de Baixo Peso (NBP)

Série NBP	X_t	ΔX_t
ADF	-1,8596	-3,9296 ***
PP	-1,8391	-4,8026 ***
KPSS	0,2384	0,0717

* rejeita-se Ho para o nível de significância de 10%
** rejeita-se H0 para o nível de significância de 5%
*** rejeita-se H0 para o nível de significância de 1%

Por fim, para a terceira série original de NMBP, os valores da estatística de teste são sempre superiores para qualquer nível de significância considerado logo a série original de NMBP é não estacionária. Segue-se a aplicação da primeira diferenciação e dos testes de raiz unitária. A hipótese nula de que existe uma raiz unitária é rejeitada para os testes ADF e PP a um nível de significância de 1%. Conclui-se que a série NMBP é estacionária para a primeira diferença.

Tabela 4.7. Aplicação dos testes de raiz unitária à série de Nascimentos de Muito Baixo Peso (NMBP)

Série NMBP	X_t	ΔX_t
ADF	-1,8109	-3,8843 ***
PP	-2,3381	-5,4291 ***
KPSS	0,2264	0,0517

* rejeita-se Ho para o nível de significância de 10%
** rejeita-se H0 para o nível de significância de 5%
*** rejeita-se H0 para o nível de significância de 1%

A série NP, e as séries transformadas NBP e NMBP são agora estacionárias, como indicam os testes de raíz unitária em que existe forte evidência estatística para rejeitar a hipótese nula de não estacionaridade, estando assim satisfeitas as condições para que a análise prossiga.

Após a diferenciação:

A Figura 4.9 é a representação gráfica da aplicação anterior do operador diferença às séries NBP e NMBP. Em ambos os casos, as primeiras diferenças das séries parecem ter eliminado a não estacionaridade em média.

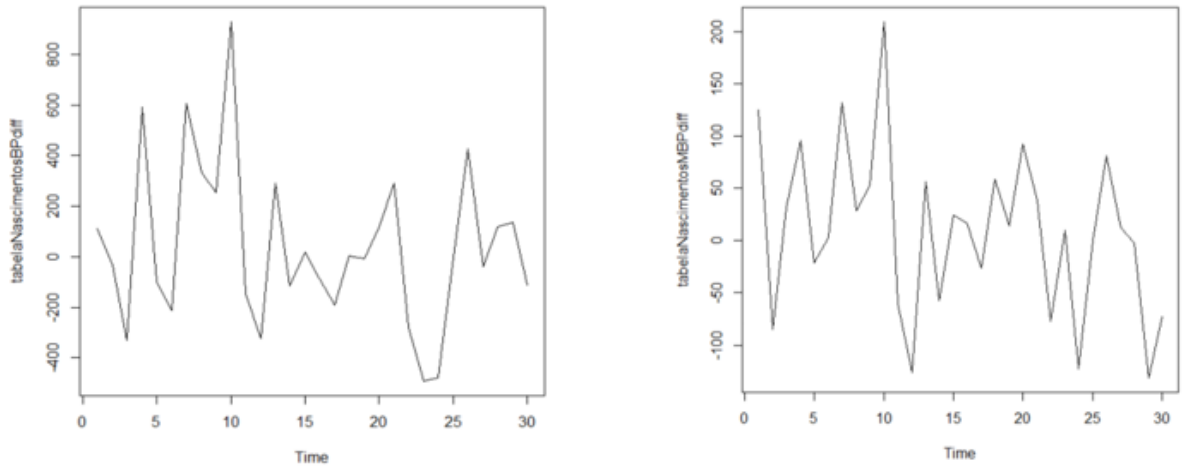
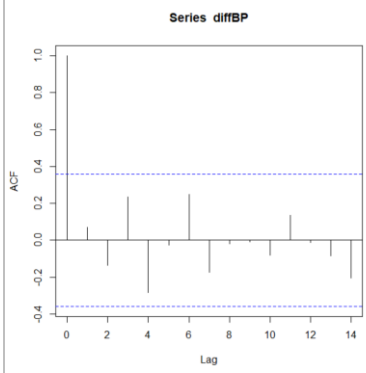
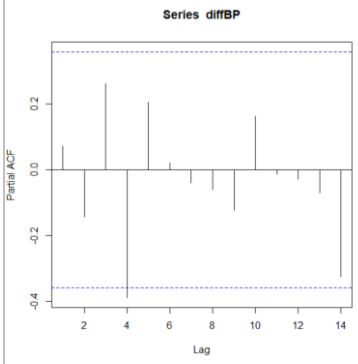
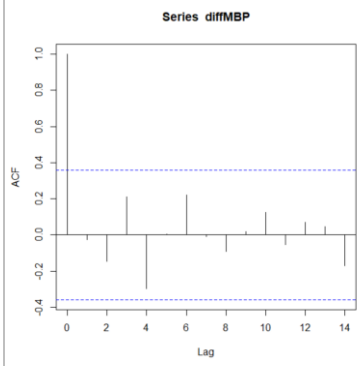
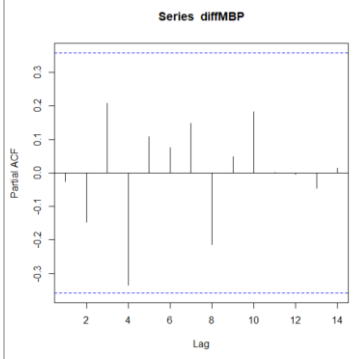


Figura 4.9. Representação gráfica do resultado da aplicação do operador de diferenciação às séries NBP e NMBP

Tabela 4.8. Resultados das FAC e FACP das séries NP,NBP e NMBP após a diferenciação

Após a diferenciação	FAC	FACP	Descrição
Série NP	-	-	Não houve necessidade de diferenciar a série NP
Série NBP			A FAC da série NBP diferenciada uma vez não apresenta decaimento exponencial, como se verificava antes do processo de diferenciação e não se evidencia um padrão nas FAC ou FACP, logo não é necessária uma segunda diferenciação. Nenhum dos valores empíricos, exceto o *lag 0*, ultrapassam os limites definidos. A FACP apresenta um padrão irregular, o *lag 1* regista um valor inferior ao registado anteriormente e o valor empírico para o lag 4 mantém-se fora dos limites definidos.
Série NMBP			A FAC da série NMBP diferenciada uma vez não apresenta decaimento exponencial, como se verificava antes de ser diferenciada, não se evidencia um padrão nas FAC ou FACP, logo não é necessária uma segunda diferenciação. Nenhum dos valores empíricos, exceto o *lag 0*, ultrapassam os limites definidos. A FACP apresenta um padrão irregular, o *lag 1* regista um valor inferior ao registado anteriormente e o valor empírico para o lag 4 é significativo, mas mantém-se dentro dos limites definidos.

4.3. Identificar e estimar o modelo

Na estimação de um modelo é necessário considerar toda a informação disponível, havendo neste processo o cruzamento entre a teoria subjacente aos modelos e os dados empíricos do modelo a estimar. O estudo teórico dos modelos de séries temporais apresentado no capítulo do Enquadramento teórico vai ser agora aplicado na estimação do modelo subjacente a cada uma das três séries temporais.

Pretende-se identificar o modelo correto dentro da classe geral dos Modelos Mistos Autoregressivos e de Médias Móveis (ARMA) ou o Modelo Misto Integrado Autoregressivo e de Médias Móveis (ARIMA). É de notar que, segundo Saboia (1977), quando se trabalha com dados demográficos raramente os parâmetros *p* e *q* são superiores a dois. Tendo isso em consideração, começa-se por aplicar os critérios de informação a diferentes combinações de modelos ARIMA, cujos parâmetros *p* e *q* variam entre 0 e 2 e o parâmetro *d* varia consoante a necessidade da série ser ou não diferenciada. Seguidamente escolhe-se o melhor modelo e estimam-se os parâmetros do modelo selecionado e a respetiva expressão geral.

4.3.1. Série NP

A comparação dos valores dos Critérios de Informação calculados para diferentes modelos ARIMA resultou na escolha do modelo ARIMA (2,0,0) como o que melhor se ajusta à série original NP. O valor AIC do modelo ARIMA (2,0,0) corresponde ao menor valor de entre os modelos possíveis (AIC= 516,65) ao qual se associa um valor BIC= 522,3886.

Tabela 4.9. Critérios de informação (AIC e BIC) aplicados a diferentes modelos ARIMA (p,d,q) para a série NP

Série NP		
Modelo ARIMA (p,d,q)	**AIC**	**BIC**
ARIMA (0,0,0)	566,23	569,096
ARIMA (1,0,0)	517,84	522,1401
ARIMA (0,0,1)	543,48	547,7851
ARIMA (1,0,1)	517,53	523,2664
ARIMA (2,0,0)	516,65	522,3886
ARIMA (0,0,2)	533,78	539,517
ARIMA (2,0,2)	518,44	539,5177

Foram estimados os parâmetros do modelo escolhido ARIMA (2,0,0), cuja estimação resultou na expressão geral:

$$NP_t = 1{,}2721\ NP_{t-1} - 0{,}3279\ NP_{t-2} + \varepsilon_t \qquad \text{Equação 4.1}$$

Com,

$$\hat{\phi}_1 = 1{,}2721$$

$$\hat{\phi}_2 = -0{,}3279$$

Os parâmetros estimados, sendo que são significativamente diferentes de zero ao nível de significância considerado. Na expressão do modelo estimado NP_t representa a série original e ε_t os erros associados quando é feito o ajuste do modelo aos dados.

```
Call:
arima(x = tabelaNascimentosP, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      1.2721  -0.3279   9428.855
s.e.  0.1700   0.1815   2265.051

sigma^2 estimated as 716668:  log likelihood = -254.33,  aic = 516.65
```

Figura 4.10. *Output* descritivo do melhor modelo ARIMA (2,0,0) para a série NP

4.3.2. Série NBP

A comparação dos valores dos Critérios de Informação calculados para diferentes modelos ARIMA resultou na escolha do modelo ARIMA (2,1,2) como o que melhor se ajusta à série diferenciada NBP. O valor AIC do modelo ARIMA (2,1,2) corresponde ao menor valor de entre os modelos possíveis (AIC= 428,92) ao qual se associa um valor BIC= 435,9235.

Tabela 4.10. Critérios de informação (AIC e BIC) aplicados a diferentes modelos ARIMA (p,d,q) para a série NBP

Série NBP		
Modelo ARIMA (p,d,q)	**AIC**	**BIC**
ARIMA (0,1,0)	433,26	434,6653
ARIMA (1,1,0)	435,04	437,8408
ARIMA (0,1,1)	434,94	437,7379
ARIMA (1,1,1)	432,16	436,3644
ARIMA (2,1,0)	436,58	440,7821
ARIMA (0,1,2)	433,30	437,5043
ARIMA (2,1,2)	428,92	435,9235

Foram estimados os parâmetros do modelo escolhido ARIMA (2,1,2), cuja estimação resultou na expressão geral:

$$NBP_t = -1{,}2124\, NBP_{t-1} - 0{,}8063\, NBP_{t-2} + \varepsilon_t + 1{,}6872\, \varepsilon_{t-1} + 0{,}9999\, \varepsilon_{t-2} \qquad \text{Equação 4.2}$$

Com,

$$\hat{\phi}_1 = -1{,}2124$$

$$\hat{\phi}_2 = -0{,}8063$$

$$\hat{\theta}_1 = 1{,}6872$$

$$\hat{\theta}_2 = 0{,}9999$$

Os parâmetros estimados, sendo que são significativamente diferentes de zero ao nível de significância considerado. Na expressão do modelo estimado NBP_t representa a série original diferenciada uma vez e ε_t os erros associados quando é feito o ajuste do modelo aos dados.

```
Call:
arima(x = tabelaNascimentosBP, order = c(2, 1, 2))

Coefficients:
          ar1      ar2     ma1     ma2
      -1.2124  -0.8063  1.6872  0.9999
s.e.   0.1070   0.1183  0.1580  0.1650

sigma^2 estimated as 57339:  log likelihood = -209.46,  aic = 428.92
```

Figura 4.11. Output descritivo do melhor modelo ARIMA (2,1,2) para a série NBP

4.3.3. Série NMBP

A comparação dos valores dos Critérios de Informação calculados para diferentes modelos ARIMA resultou na escolha do modelo ARIMA (2,1,2) como o que melhor se ajusta à série diferenciada NMBP. O valor AIC do modelo ARIMA (2,1,2) corresponde ao menor valor de entre os modelos possíveis (AIC= 346,60) ao qual se associa um valor BIC= 353,6036.

Tabela 4.11. Critérios de informação (AIC e BIC) aplicados a diferentes modelos ARIMA (p,d,q) para a série NMBP

Série NMBP		
Modelo ARIMA (p,d,q)	**AIC**	**BIC**
ARIMA (0,1,0)	349,69	351,0901
ARIMA (1,1,0)	351,68	353,4869
ARIMA (0,1,1)	351,68	354,4854
ARIMA (1,1,1)	353,65	357,8540
ARIMA (2,1,0)	353,17	357,3780
ARIMA (0,1,2)	349,69	353,8941
ARIMA (2,1,2)	346,60	353,6036

Foram estimados os parâmetros do modelo escolhido ARIMA(2,1,2), cuja estimação resultou na expressão geral:

$$NMBP_t = -1{,}3587\ NMBP_{t-1} - 0{,}7842\ NMBP_{t-2} + \varepsilon_t + 1{,}8516\ \varepsilon_{t-1} + 1{,}0000\ \varepsilon_{t-2} \qquad \text{Equação 4.3}$$

Com,

$$\hat{\phi}_1 = -1{,}3587$$

$$\hat{\phi}_2 = -0{,}7842$$

$$\hat{\theta}_1 = 1{,}8516$$

$$\hat{\theta}_2 = 1{,}0000$$

E os parâmetros estimados, sendo que todos os parâmetros ajustados são significativamente diferentes de zero ao nível de significância considerado. Na expressão do modelo estimado $NMBP_t$ representa a série original diferenciada uma vez e ε_t os erros associados quando é feito o ajuste do modelo aos dados.

```
Call:
arima(x = tabelaNascimentosMBP, order = c(2, 1, 2))

Coefficients:
          ar1      ar2     ma1     ma2
      -1.3587  -0.7842  1.8516  1.0000
s.e.   0.1212   0.1210  0.1804  0.1872

sigma^2 estimated as 3632:  log likelihood = -168.3,  aic = 346.6
```

Figura 4.12. *Output* descritivo do melhor modelo ARIMA (2,1,2) para a série NMB

4.4. Diagnóstico

A fase de diagnóstico compreende duas etapas: a análise da qualidade estatística e da consequente significância estatística dos parâmetros e a análise da qualidade do ajustamento do modelo, através da inspeção dos resíduos.

4.4.1. Avaliação da qualidade estatística

Os Resultados do estudo da significância estatística dos parâmetros estimados para cada um dos modelos estão disponíveis para consulta em Anexo K.

Os parâmetros estimados anteriormente para as séries NBP e NMBP são todos significativamente diferentes de zero e, consequentemente, mantêm-se no modelo. Contudo, o parâmetro AR2, contrariamente ao que indica o critério de informação AIC, não é significativamente diferente de zero e deve ser eliminado do modelo. Como previsto, em certos casos, pela Metodologia *Box Jenkins* o modelo calculado surge como menos adequado do que inicialmente se esperava, havendo necessidade de retomar o ciclo. Segue-se a escolha de um novo modelo ARIMA para a série NP, escolha essa que resultou no modelo ARIMA (1,0,0). O processo anterior encontra-se explicitado no Anexo L.

4.4.2. Análise dos Resíduos

A **análise dos resíduos** permite medir a qualidade de ajustamento de um modelo estimado e, para que tal seja possível, analisam-se em seguida 3 gráficos essenciais na análise da qualidade de ajustamento de um modelo: o gráfico dos resíduos *standartizados*, o gráfico das FAC dos resíduos e o gráfico dos *pvalues* associados à estatística do Teste *Ljung-Box*. Para um modelo corretamente especificado, os resíduos deverão ter um comportamento semelhante ao de um ruído branco. Os resultados dos testes *Ljung-Box* e *Box-Pierce* que sustentam as conclusões seguintes encontram-se no Anexo N. As análises dos resíduos apresentam resultados semelhantes nas 3 séries, como se confirma nas tabelas seguintes.

Tabela 4.12. Resultados da Análise dos Resíduos do modelo ARIMA (1,0,0) para a série NP

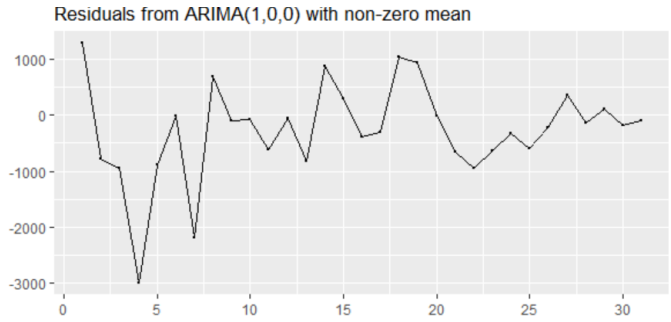
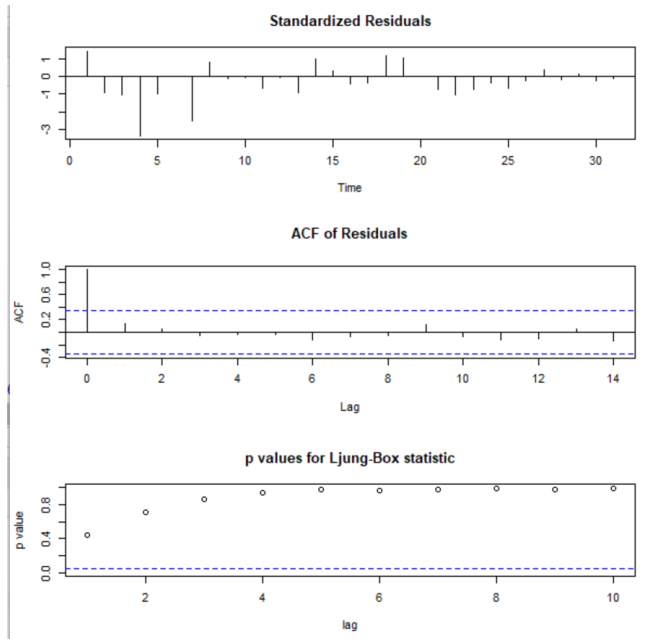
Série NP	Gráficos Análise Resíduos
▪ O gráfico dos resíduos padronizados sugere uma distribuição aleatória em torno do zero; ▪ Da análise dos FAC dos Resíduos resultantes do ajustamento da série NP conclui-se que pelo menos 95% das correlações estão dentro dos limites de confiança indicados pelas retas a tracejado e que não há necessidade de proceder a um novo ajustamento; ▪ As FAC dos resíduos têm um comportamento semelhante ao comportamento das FAC de um processo puramente aleatório ou de um ruído branco e os resíduos são não correlacionados com valores das correlações muito baixos. Pelas razões anteriores, não há necessidade de proceder a um novo ajustamento; ▪ O gráfico dos *pvalues* calculado para a estatística de teste do Teste *Ljung-Box* revela valores do *pvalue* superiores a 0,5 pelo que não se rejeita a hipótese nula e, consequentemente, não se rejeita a normalidade de os resíduos constituírem um ruído branco; Os resultados da aplicação direta do Teste *Ljung-Box* (Anexo N) ao modelo NP para os 10 primeiros *lags* corrobora as conclusões anteriores;	 **Figura 4.13.** Gráfico dos resíduos ARIMA (1,0,0) – série NP  **Figura 4.14.** Resultados Análise dos Resíduos ARIMA (1,0,0) - série NP

Tabela 4.13. Resultados da Análise dos Resíduos do modelo ARIMA (2,1,2) para a série NBP

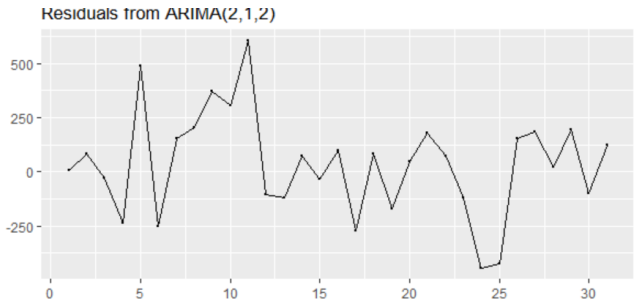
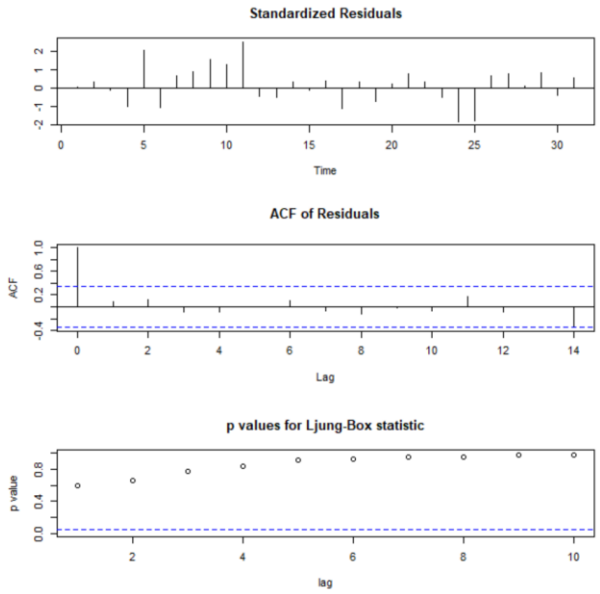
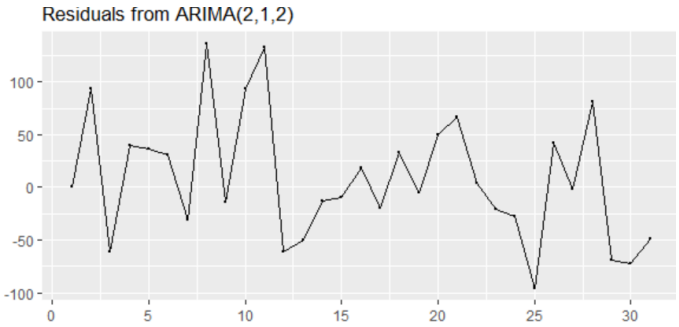
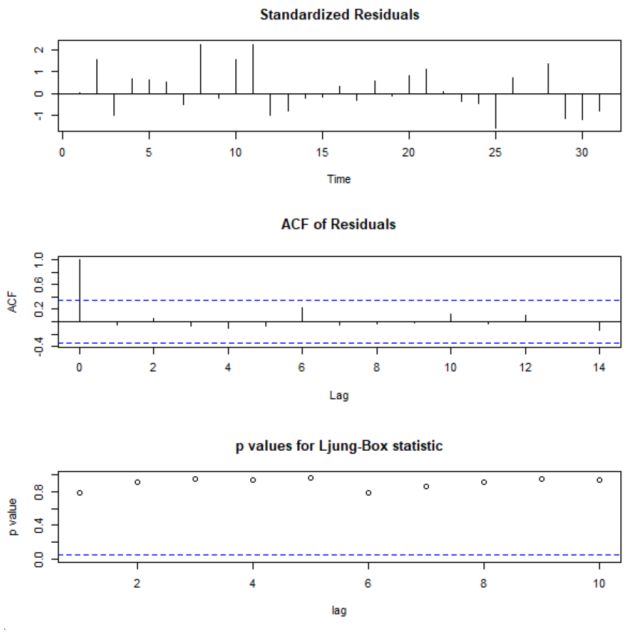
Série NBP	Gráficos Análise Resíduos
▪ O gráfico dos resíduos padronizados sugere uma distribuição aleatória em torno de zero; ▪ Da análise dos FAC dos Resíduos resultantes do ajustamento da série NBP conclui-se que pelo menos 95% das correlações estão dentro dos limites críticos indicados pelas retas a tracejado; ▪ As FAC dos resíduos têm um comportamento semelhante ao comportamento das FAC de um processo puramente aleatório ou de um ruído branco e os resíduos são não correlacionados com valores das correlações muito baixos. Pelas razões anteriores, não há necessidade de proceder a um novo ajustamento. De referir que nos *lags* 11, 24 e 25 existe uma correlação muito próxima da banda de confiança, não a ultrapassando; ▪ O gráfico dos *pvalues* calculado para a estatística de teste do Teste *Ljung-Box* revela valores do *pvalue* superiores a 0,5 pelo que não se rejeita a hipótese nula e, consequentemente, não se rejeita a normalidade de os resíduos constituírem um ruído branco; Os resultados da aplicação direta do Teste *Ljung-Box (*Anexo N) ao modelo NBP para os 10 primeiros *lags* corrobora as conclusões anteriores;	 **Figura 4.15.** Gráfico dos resíduos ARIMA (2,1,2) – série NBP  **Figura 4.16**. Resultados Análise dos Resíduos ARIMA (2,1,2) - série NBP

Tabela 4.14. Resultados da Análise dos Resíduos do modelo ARIMA(2,1,2) para a série NMBP

Série NMBP	Gráficos Análise Resíduos
▪ O gráfico dos resíduos padronizados sugere uma distribuição aleatória em torno do zero; ▪ Da análise dos FAC dos Resíduos resultantes do ajustamento da série NMBP conclui-se que pelo menos 95% das correlações estão dentro dos limites críticos, indicados pelas retas a tracejado; ▪ As FAC dos resíduos têm um comportamento semelhante ao comportamento das FAC de um processo puramente aleatório ou de um ruído branco e os resíduos são não correlacionados com valores das correlações muito baixos. Pelas razões anteriores, não há necessidade de proceder a um novo ajustamento; ▪ Também para a série NMBP, o gráfico dos *pvalues* calculado para a estatística de teste do Teste *Ljung-Box* revela valores do *pvalue* superiores a 0,5, neste caso superiores a 0,8, pelo que não se rejeita a hipótese nula e, consequentemente, não se rejeita a normalidade de os resíduos constituírem um ruído branco; Os resultados da aplicação direta do Teste *Ljung-Box* (Anexo N) ao modelo NMBP para os 10 primeiros lags corrobora as conclusões anteriores;	 **Figura 4.17.** Gráfico dos resíduos ARIMA (2,1,2) – série NMBP  **Figura 4.18**. Resultados Análise dos Resíduos ARIMA (2,1,2) - série NMBP

Uma vez encontrado o modelo adequado, que cumpre os requisitos propostos para cada uma das três séries, estão reunidas as condições para que se inicie a etapa final das previsões dos valores futuros para cada uma das séries.

Previsão

A última etapa passa pela previsão das 3 séries temporais em estudo, sendo que os resultados retirados desta secção permitirão responder a um dos objetivos propostos, que se traduz em encontrar a melhor previsão possível para cada uma das 3 séries NP, NBP e NMBP.

A previsão feita para cada uma das séries será executada através de 3 modelos: ARIMA, Alisamento Exponencial e Médias Móveis, sendo que será posteriormente comparado o poder preditivo de cada um deles. Do ponto de vista prático, para comparar os 3 modelos é feita uma análise gráfica que é complementada pela comparação das medidas de desempenho descritas na componente teórica, nomeadamente as medidas ME, RMSE e MAE.

Previsão com modelos ARIMA

Os modelos ARIMA usados na fase de previsão foram ARIMA (1,0,0), ARIMA (2,1,2) e ARIMA (2,1,2), respetivamente para as séries NP, NBP e NMBP. A previsão de cada série será composta por uma previsão *in sample*, que representa uma previsão para a janela temporal da própria série, e por uma previsão *out of sample*, onde são previstos valores para fora do intervalo temporal utilizado. Na previsão *in sample* é utilizada a janela temporal de 2016 a 2019 e na previsão *out of sample* consideram-se os 4 anos seguintes à última observação da série (2020, 2021, 2022 e 2023), sendo que, destes últimos, estão apenas disponíveis para comparação os valores efetivamente registados dos anos 2020 e 2021.

- Previsão *in sample*

Os resultados posteriores sumarizam o desempenho da previsão *in sample* para as séries NP, NBP e NMBP entre 2016 e 2019, na qual é feita a comparação com os valores efetivamente registados.

Série NP

Verifica-se pela observação da Figura 4.19 e pelos resultados obtidos na tabela que os valores previstos para os NP são, em módulo, muito próximos dos valores registados tendo em consideração a amplitude dos valores históricos. A previsão obtida através do modelo ARIMA está dentro do IC considerado, tem um forte poder preditivo e adequa-se significativamente bem à série NP.

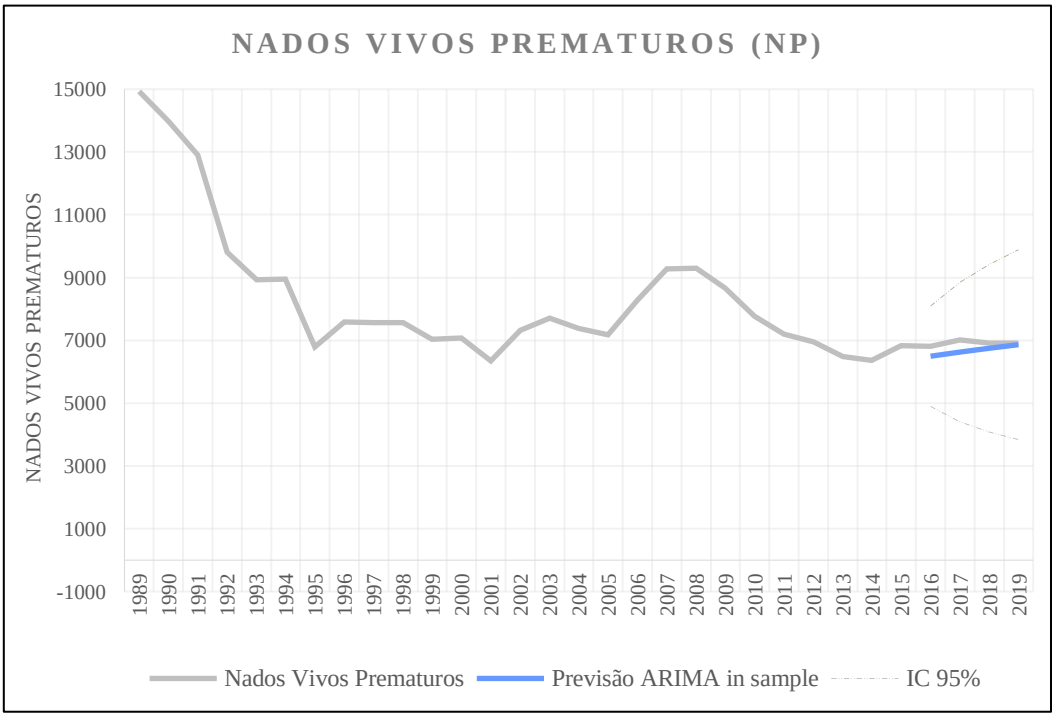


Figura 4.19. Previsão da série NP usando o modelo ARIMA (1,0,0) para h=4

Tabela 4.15. Previsão da série NP para os anos 2016,2017,2018 e 2019 pelo modelo ARIMA (1,0,0) para um nível de confiança de 90% e 95%

Ano	**Previsão NP**	**IC de 90%**		**IC de 95%**		**Valor registado**	**|erro $_{prev}$|**
		Lim inf	*Lim sup*	*Lim inf.*	*Lim sup.*		
2016	6495,389	4897,941	8092,837	4591,912	8398,865	6801	306
2017	6622,835	4405,472	8840,197	3980,684	9264,985	7011	389
2018	6745,523	4079,388	9411,657	3568,627	9922,418	6922	177
2019	6863,639	3840,509	9886,752	3261,359	10465,902	6913	50

Série NBP

Pela observação da Figura 4.20 e pelos resultados obtidos na Tabela 4.16, verifica-se que os valores previstos estão dentro do IC considerado e que o modelo tem um relativo poder preditivo. Os valores previstos foram, para os 4 anos considerados, inferiores aos valores registados.

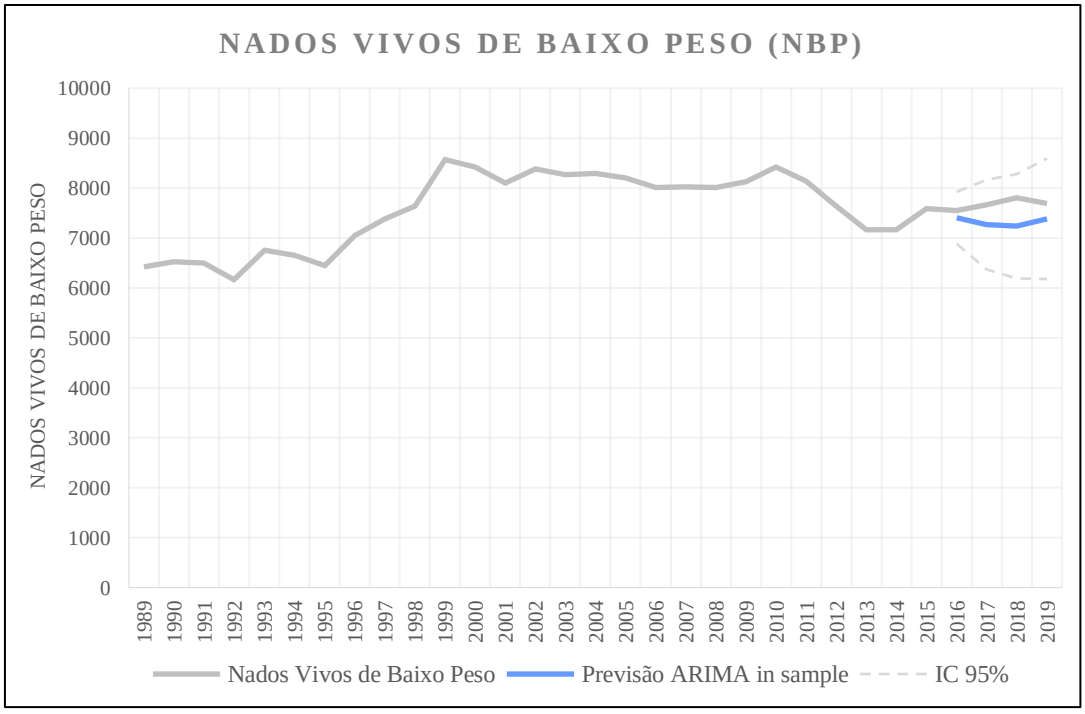


Figura 4.20. Previsão da série NBP usando o modelo ARIMA (2,1,2) para h=4

Tabela 4.16. Previsão da série NBP para os anos 2016,2017,2018 e 2019 pelo modelo ARIMA (2,1,2) para um nível de confiança de 90% e 95%

Ano	**Previsão NBP**	**IC de 90%**		**IC de 95%**		**Valor registado**	**|erro $_{prev}$|**
		Lim inf.	*Lim sup.*	*Lim inf.*	*Lim sup.*		
2016	7404,876	6969,663	7840,089	6886,287	7923,464	7550	146
2017	7271,771	6519,709	8023,833	6375,634	8167,908	7667	396
2018	7238,143	6359,489	8116,797	6191,162	8285,124	7804	566
2019	7385,430	6372,893	8397,968	6178,917	8591,943	7694	309

Série NMBP

Verifica-se pela observação da Figura 4.20 e pelos resultados obtidos na Tabela 4.17 que a previsão obtida através do modelo ARIMA foi em 3 dos 4 anos superior aos valores registados, os valores estão dentro do IC considerado e que o modelo tem um relativo poder preditivo.

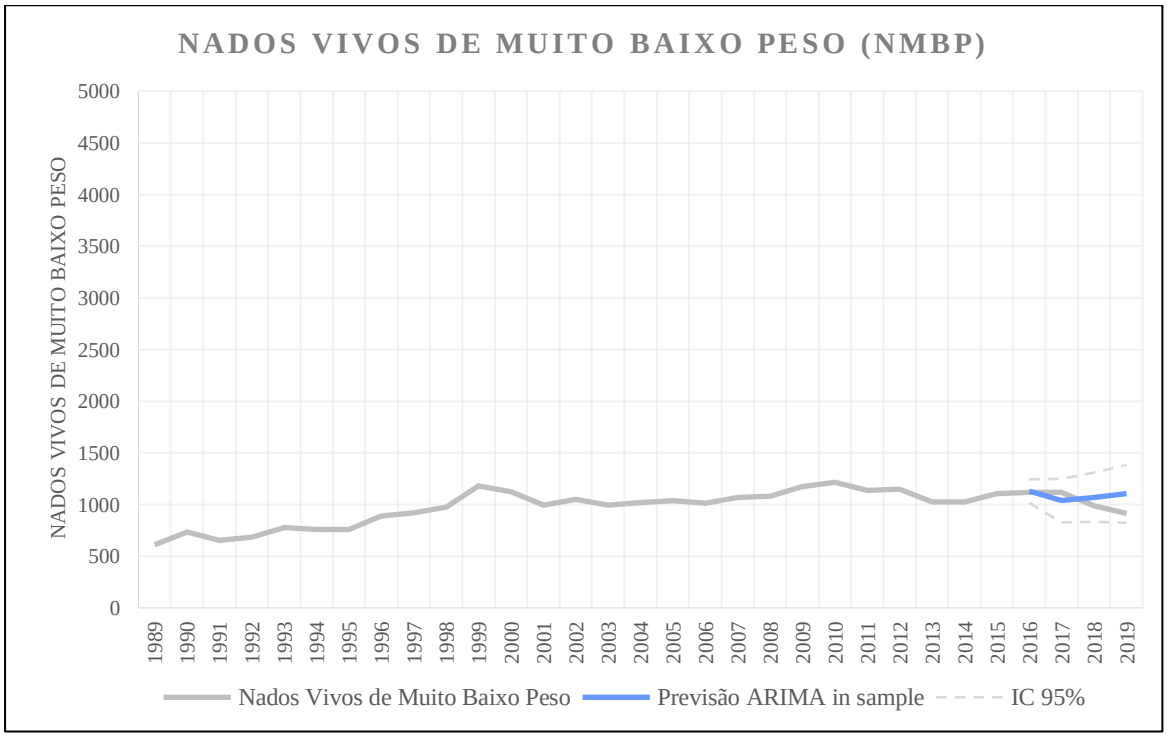


Figura 4.21. Previsão da série NMBP usando o modelo ARIMA (2,1,2) para h=4

Tabela 4.17. Previsão da série NMBP para os anos 2016,2017,2018 e 2019 pelo modelo ARIMA (2,1,2) para um nível de confiança de 90% e 95%

Ano	Previsão NMBP	IC de 90%		IC de 95%		Valor registado	|erro prev|
		Lim inf.	*Lim sup.*	*Lim inf.*	*Lim sup.*		
2016	1130,347	1033,906	1226,789	1015,430	1245,264	1118	12
2017	1039,233	862,711	1215,755	828,894	1249,572	1116	77
2018	1071,412	872,018	1270,806	833,820	1309,005	985	86
2019	1103,603	869,752	1337,453	824,952	1382,253	912	191

Tabela 4.18. Medidas de desempenho da previsão in sample das séries NP, NBP e NMBP

Medidas de desempenho	NP	NBP	NMBP
ME	-327,115	27,764	13,958
RMSE	971,179	250,391	55,244
MAE	721,029	197,088	44,411

- Previsão *out of sample*

As tabelas em seguida sumarizam o desempenho da previsão *out of sample* para as séries NP, NBP e NMBP entre 2020 e 2023, na qual é feita a comparação com os valores efetivamente registados. As representações gráficas complementares aos resultados obtidos encontram-se no Anexo O.

Série NP

Tabela 4.19. Previsão da série NP para os anos 2020,2021,2022 e 2023 pelo modelo ARIMA (1,0,0) para um nível de confiança de 90/95%

Ano	Previsão NP	IC de 90%		IC de 95%		Valor registado	|erro $_{prev}$|
		Lim inf.	*Lim sup.*	*Lim inf.*	*Lim sup.*		
2020	7015,185	5549,191	8481,178	5268,345	8764,024	5765	1250
2021	7113,948	5075,134	9152,762	9543,344	9543,344	5997	1116
2022	7209,404	4753,352	9665,456	4282,838	10135,970	-	-
2023	7301,664	3977,168	10626,160	3977,168	10626,160	-	-

Série NBP

Tabela 4.20. Previsão da série NBP para os anos 2020,2021,2022 e 2023 pelo modelo ARIMA (2,1,2) para um nível de confiança de 90% e 95%

Ano	Previsão NP	IC de 90%		IC de 95%		Valor registado	|erro $_{prev}$|
		Lim inf.	*Lim sup.*	*Lim inf.*	*Lim sup.*		
2020	7832,679	7426,464	8238,893	7348,644	8316,713	6663	1169
2021	7874,613	7168,282	8580,945	7032,967	8716,259	6708	1166
2022	7711,957	6888,131	8535,783	6730,307	8693,607	-	-
2023	7875,345	6925,782	8824,908	6743,871	9006,819	-	-

Série NMBP

Tabela 4.21. Previsão da série NMBP para os anos 2020,2021,2022 e 2023 pelo modelo ARIMA (2,1,2) para um nível de confiança de 95%

Ano	Previsão NMBP	IC de 90%		IC de 95%		Valor registado	|erro $_{prev}$|
		Lim inf.	*Lim sup.*	*Lim inf.*	*Lim sup.*		
2020	952,1443	849,9094	1054,379	830,3239	1073,965	730	222
2021	907,0977	728,1751	1086,020	693,8983	1120,297	789	118
2022	936,8225	730,8276	1142,817	691,3644	1182,281	-	-
2023	931,7598	689,2634	1174,256	642,8075	1220,712	-	-

Método de Alisamento Exponencial (AE)

O método do Alisamento Exponencial, como exposto na componente teórica, é utilizado para previsões de séries temporais de curto prazo, com a vantagem de ser prático uma vez que os coeficientes são atualizados a cada momento, não permanecendo fixos com o avançar do processo. O método de Alisamento Exponencial subdivide-se em Alisamento Exponencial Simples e Holt Winters (para séries com tendência e sazonalidade) sendo que no caso das séries em análise é apenas aplicado o método de Alisamento Exponencial Simples.

A previsão feita através do método de Alisamento Exponencial contabiliza a previsão *in sample* e *out of sample*. Em particular, a previsão *in sample* para as três séries NP, NBP e NMBP incide no intervalo temporal 2016-2019.

- Previsão *in sample*

  - Série NP

Os valores previstos para os NP pelo método do AE estão dentro dos intervalos de confiança considerados, mas comparativamente com o modelo ARIMA registam valores mais afastados dos registados, sendo que o poder preditivo do modelo é relativo.

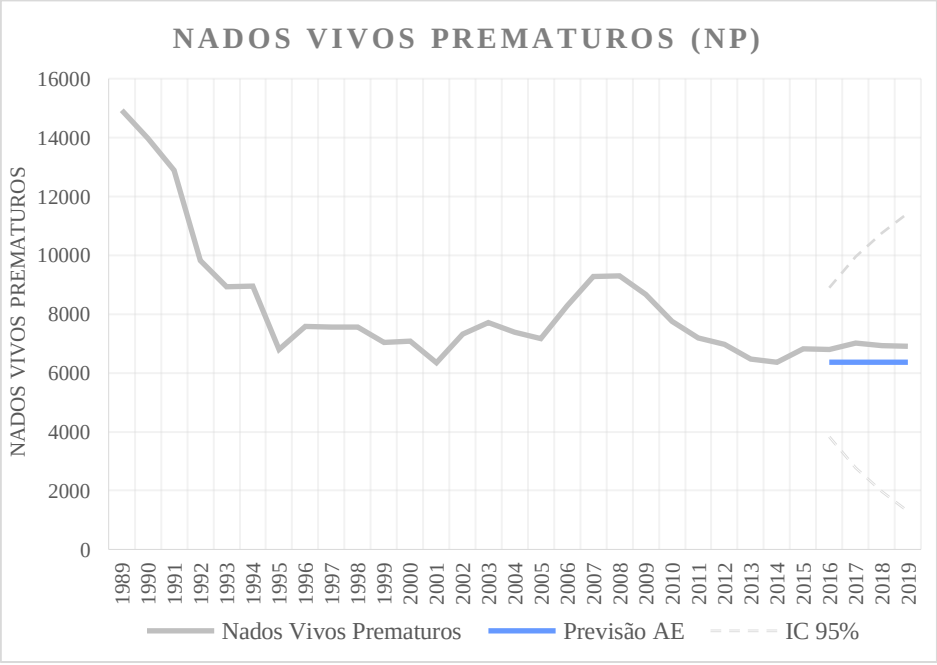


Figura 4.22. Representação gráfica da previsão da série NP pelo Método de AE para *h=4*

Tabela 4.22. Aplicação do Método de Alisamento Exponencial à série NP para os anos 2016, 2017, 2018, 2019 para um nível de confiança de 90% e 95%

Ano	Previsão NP	IC de 90%		IC de 95%		Valor registado	|erro prev|
		Lim inf	*Lim sup*	*Lim inf.*	*Lim sup.*		
2016	6363,011	4236,672	8489,349	3829,324	8896,699	6801	438
2017	6363,011	3356,066	9369,957	2780,015	9946,008	7011	648
2018	6363,011	2680,332	10045,691	1974,828	10751,195	6922	559
2019	6363,011	2110,655	10615,368	1296,016	11430,007	6913	550

- Série NBP

Os valores previstos para os NBP pelo método AE estão dentro do intervalo de confiança considerado de 95%, o modelo tem um poder de previsão relativo e os erros de previsão são relativamente superiores aos registados com o modelo ARIMA.

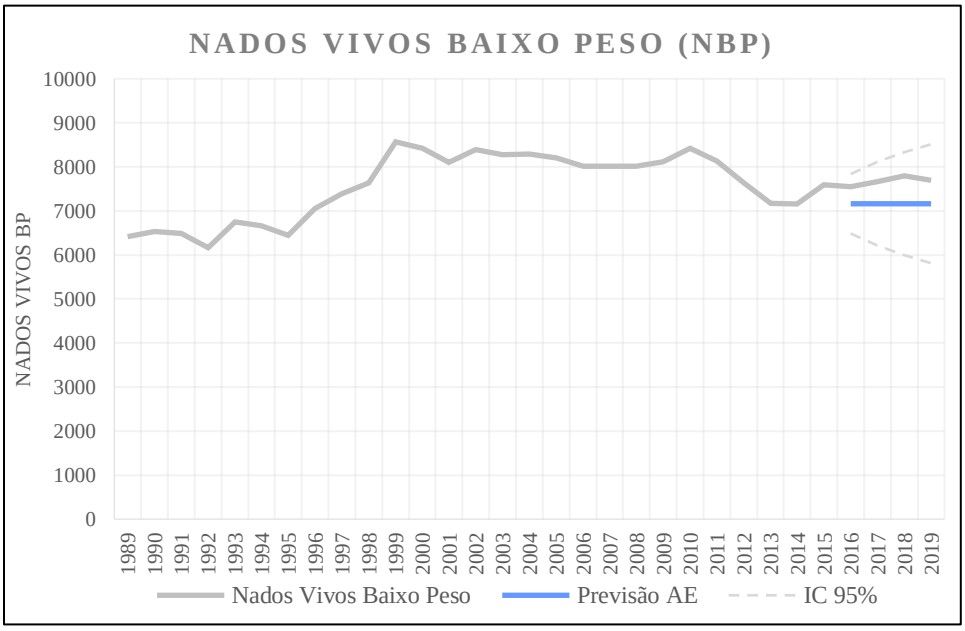


Figura 4.23. Representação gráfica da previsão da série NBP pelo Método de AE para *h=4*

Tabela 4.23. Aplicação do Método de Alisamento Exponencial à série NBP para os anos 2016, 2017, 2018, 2019 para um nível de confiança de 90% e 95%

Ano	Previsão NP	IC de 90%		IC de 95%		Valor registado	|erro prev|
		Lim inf	*Lim sup*	*Lim inf.*	*Lim sup.*		
2016	7163	6596,464	7729,536	6487,931	7838,069	7550	387
2017	7163	6361,838	7964,163	6208,356	8117,644	7667	504
2018	7163	6181,797	8144,204	5993,824	8332,176	7804	641
2019	7163	6030,014	8295,987	5812,963	8513,037	7694	531

- Série NMBP

Os valores previstos para os NMBP pelo método de AE encontram-se dentro do intervalo de confiança considerado de 95% e registam-se erros de previsão pouco significativos. De referir que em 2 dos anos previstos os valores estão acima dos registados.

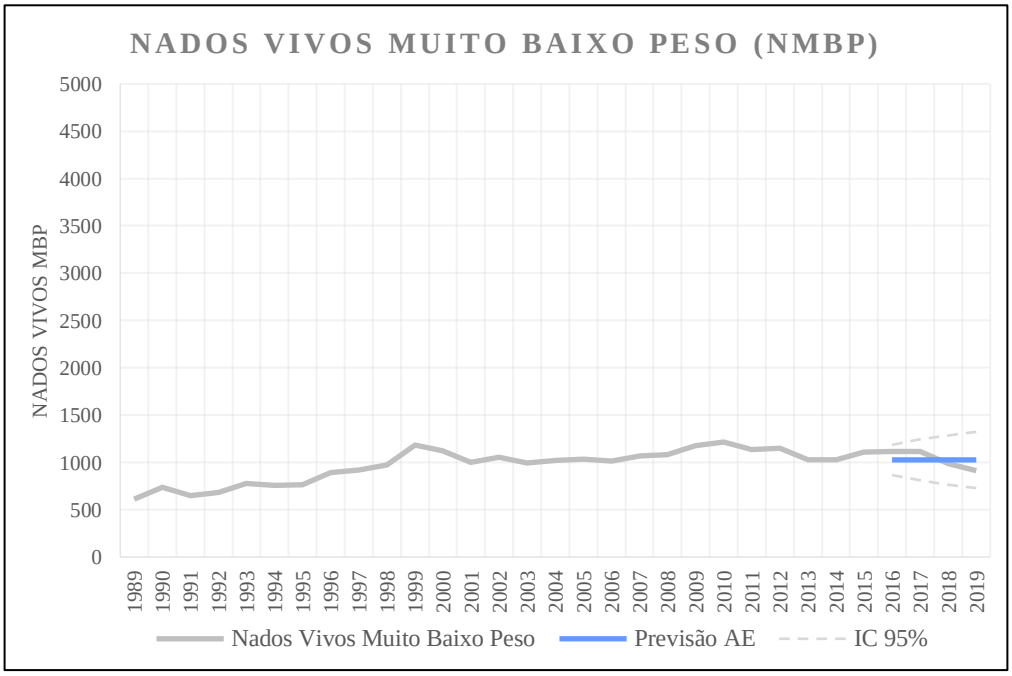


Figura 4.24. Representação gráfica da previsão da série NMBP pelo Método de AE para *h=4*

Tabela 4.24. Aplicação do Método de Alisamento Exponencial à série NMBP para os anos 2016, 2017, 2018, 2019 para um nível de confiança de 90% e 95%

Ano	Previsão NP	IC de 90%		IC de 95%		Valor registado	|erro prev|
		Lim inf	*Lim sup*	*Lim inf.*	*Lim sup.*		
2016	1026	892,113	1160,037	866,449	1185,701	1118	92
2017	1026	845,275	1206,876	810,639	1241,512	1116	90

2018	1026	808,287	1243,803	766,565	1285,586	985	41
2019	1026	776,727	1275,413	728,959	1323,192	912	114

- Previsão *out of sample*

As tabelas em seguida sumarizam o desempenho da previsão *out of sample* para as séries NP, NBP e NMBP entre 2020 e 2023 através do Método de Alisamento Exponencial, na qual é feita a comparação com os valores efetivamente registados. As representações gráficas complementares aos resultados obtidos encontram-se no Anexo.

Série NP

Tabela 4.25. Previsão da série NP através do método de AE para um IC de 90-95% e *h=4*

Ano	**Previsão NP**	**IC de 90%**		**IC de 95%**		**Valor registado**	**|erro $_{prev}$|**
		Lim inf.	*Lim sup.*	*Lim inf*	*Lim sup*		
2020	6913,001	4972,133	8853,869	4600,314	9225,687	5765	1148
2021	6913,001	4168,337	9657,665	3642,532	10183,470	5997	916
2022	6913,001	3551,543	10274,458	2907,577	10918,475	-	-
2023	6913,001	3031,557	10794,445	2287,975	11538,027	-	-

Série NBP

Tabela 4.26. Previsão da série NBP através do método de AE para um IC de 90-95% e h=4

Ano	**Previsão NBP**	**IC de 90%**		**IC de 95%**		**Valor registado**	**|erro $_{prev}$|**
		Lim inf.	*Lim sup.*	*Lim inf*	*Lim sup*		
2020	7694,011	7158,400	8229,622	7055,791	8332,231	6663	1031
2021	7694,011	6936,581	8451,441	6791,477	8596,545	6708	986
2022	7694,011	6766,367	8621,655	6588,655	8799,367	-	-
2023	7694,011	6622,869	8765,153	6417,667	8970,355	-	-

Série NMBP

Tabela 4.27. Previsão da série NMBP através do método de AE para um IC de 90-95% e h=4

Ano	**Previsão NMBP**	**IC de 90%**		**IC de 95%**		**Valor registado**	**|erro $_{prev}$|**
		Lim inf.	*Lim sup.*	*Lim inf*	*Lim sup*		
2020	913,316	780,313	1046,320	754,833	1071,800	730	183
2021	913,316	726,858	1099,775	691,137	1135,495	789	124
2022	913,316	685,625	1141,008	642,005	1184,628	-	-
2023	913,316	650,789	1175,843	600,496	1226,136	-	-

Tabela 4.28. Medidas de desempenho da aplicação do método AE para as 3 séries

Medidas de desempenho	NP	NBP	NMBP
ME	-174,122	27,745	17,146
RMSE	1242,007	330,917	78,248
MAE	819,439	244,871	59,433

Modelo das Médias Móveis Simples

Na aplicação do Modelo das médias móveis simples escolheu-se um valor *N* (sendo que *N* se refere ao nº de observações incluídas em cada média, denominado período da média móvel). O objetivo consistiu em "alisar" a série através do operador média, filtrar flutuações aleatórias e identificar um padrão nos dados. O valor *N* ajusta-se consoante se quer o alisamento da série, tendo em consideração que quanto maior for mais informação relevante será perdida relativamente ao padrão da série original.

Para a previsão de Médias Móveis foi aplicada a previsão *in sample*, em que as séries NP,NBP e NMBP são divididas em intervalos de *N* valores.Tendo em consideração que cada uma das séries originais tem apenas 30 observações, utilizaram-se os valores *N=2 e N=5* e apresentam-se os resultados para os 4 últimos anos (de 2016 a 2019) do modelo com menor RMSE [3], explicitando-se o modelo alternativo no Anexo Q. Sumariza-se também no Anexo Q o desempenho *out of sample* das séries NP, NBP e NMBP.

Série NP

A Tabela 4.29 representa o desempenho da previsão *in sample* da série NP através do Modelo de médias móveis para *N=2 e h=4*. Os erros associados a esta previsão são, em média, baixos o que se traduz numa previsão com um desempenho significativamente bom. Observa-se também que a medida de desempenho *RMSE* para *N=2* (*RMSE= 57,387*) é um valor baixo e inferior ao valor calculado para *N=5 (RMSE= 184,396*), sustentando as conclusões retiradas da exclusiva observação dos erros.

[3] Contrariamente aos outros modelos, a medida de desempenho RMSE foi calculada com o auxílio do Excel

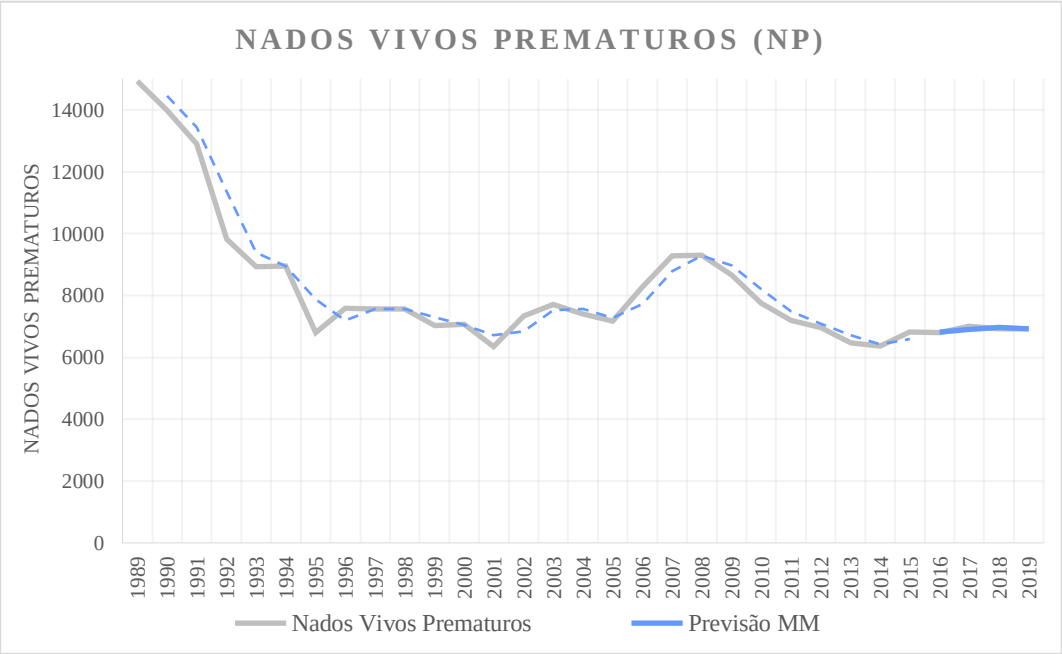


Figura 4.25. Aplicação do Modelo das Médias Móveis à série NP para N=2 e h=4

Tabela 4.29. Previsão *in sample* da série NP através do Modelo de Médias Móveis Simples para N=2

Ano	Previsão NP	Valor registado	|erro $_{prev}$|	RMSE
2016	6815,0	6801	14	455,4
2017	6906,0	7011	105	
2018	6966,5	6922	44	
2019	6917,5	6913	4	

Série NBP

A Tabela 4.30 representa o desempenho da previsão *in sample* da série NBP através do Modelo de médias móveis para *N=2 e h=4*. Os erros associados a esta previsão são, em média, baixos o que se traduza numa previsão com um desempenho significativamente bom. Observa-se também que a medida de desempenho *RMSE* para *N=2* (*RMSE= 54,007*) é um valor baixo e inferior ao valor calculado para *N=5 (RMSE= 185,454* sustentando assim as conclusões retiradas da exclusiva observação dos erros.

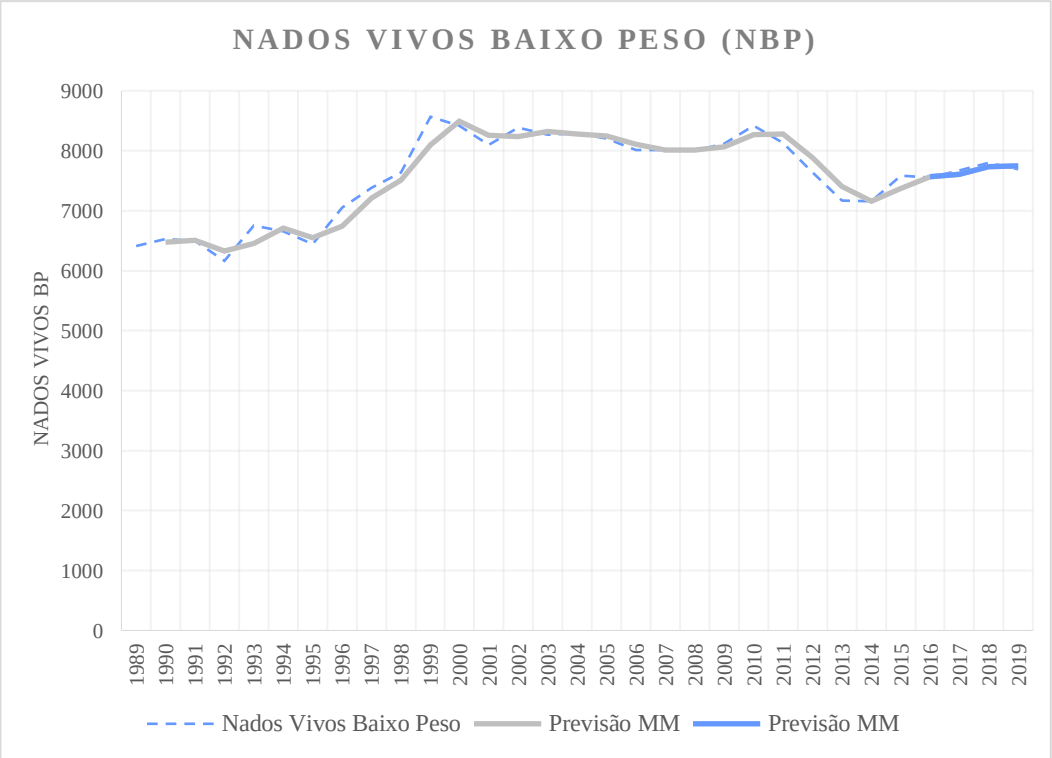


Figura 4.26. Aplicação do Modelo das Médias Móveis à série NBP para N=2 e h=4

Tabela 4.30. Previsão in sample da série NBP através do Modelo de Médias Móveis Simples para N=2

Ano	Previsão NBP	Valor registado	|erro $_{prev}$|	RMSE
2016	7570,0	7550	20	162,9
2017	7608,5	7667	59	
2018	7735,5	7804	69	
2019	7749,0	7694	55	

Série NMBP

A Tabela 4.31 representa o desempenho da previsão *in sample* da série NMBP através do Modelo de médias móveis para *N=2 e h=4.* Os erros associados a esta previsão são, em média, baixos o que se traduza numa previsão com um desempenho significativamente bom. Observa-se também que a medida de desempenho *RMSE* para *N=2* (*RMSE= 37,276*) é um valor baixo e inferior ao valor calculado para *N=5 (RMSE= 83,741)*

sustentando as conclusões retiradas da exclusiva observação dos erros.

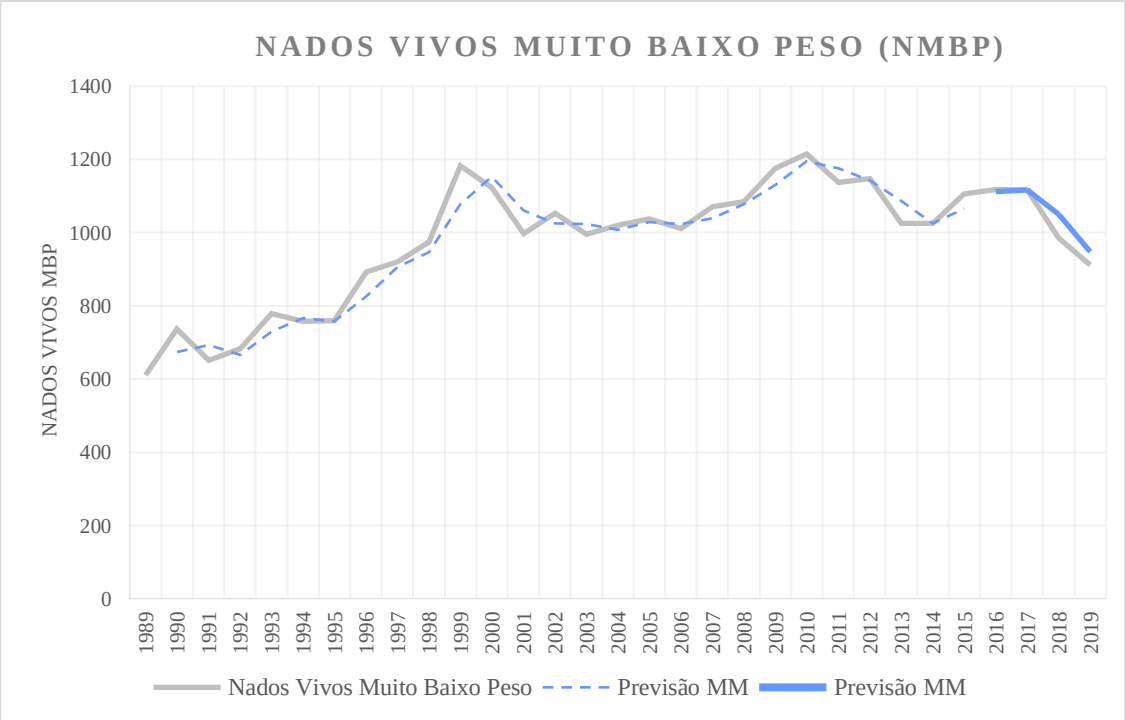


Figura 4.27. Aplicação do Modelo das Médias Móveis à série NMBP para N=2 e h=4

Tabela 4.31. Previsão in sample da série NMBP através do Modelo de Médias Móveis Simples para N=2

Ano	Previsão NMBP	Valor registado	$\|erro_{prev}\|$	RMSE
2016	1112,0	1118	6	40,4
2017	1117,0	1116	1	
2018	1050,5	985	65	
2019	948,5	912	36	

Como será detalhado na secção seguinte, os Resultados obtidos para o modelo MM, comparativamente aos modelos ARIMA e AE, mostram-se para as 3 séries, mais próximos dos valores efetivamente registados pelo INE para os anos de 2016 a 2020.

Análise comparativa dos métodos de previsão

Na tabela 4.32 registam-se para cada uma das 3 séries as medidas de desempenho (ME, RMSE, MAE, MAPE) calculadas para os modelos preditivos ARIMA in sample, Modelo de Alisamento Exponencial e Método das Médias Móveis.

Tabela 4.32. Comparação das medidas de desempenho na aplicação dos diferentes modelos de previsão às séries NP, NBP e NMBP

	NP				**NBP**				**NMBP**			
	ME	*RMSE*	*MAE*	*MAPE*	*ME*	*RMSE*	*MAE*	*MAPE*	*ME*	*RMSE*	*MAE*	*MAPE*
ARIMA	-327,1	971,2	721,0	8,6	27,8	250,4	197,1	2,6	13,9	55,2	44,4	4,7
AE	-174,1	1242,0	819,4	9,1	27,7	330,9	244,9	3,3	17,1	78,2	59,4	6,2
MM	-138,1	455,4	311,4	3,7	22,4	162,9	123,7	1,6	5,5	40,4	32,2	3,3

A **análise comparativa** dos modelos de previsão apresenta conclusões semelhantes para as 3 séries NP, NBP e NMBP. Apesar da qualidade preditiva dos 3 modelos (ARIMA, AE e MM) ser relativamente boa e não diferir muito entre si (com vantagem para o modelo MM), é de assinalar, primeiramente, uma vantagem e uma desvantagem na aplicação do modelo ARIMA, o modelo ao qual se despendeu mais tempo de análise esse tipo de modelo exigiu cuidados em termos de análise inicial e pré-processamento dos dados (em particular no estudo da estacionaridade), o que não aconteceu nos outros 2 modelos que exigiram menos tempo implícito na execução e obtenção de resultados.

No que diz respeito às medidas de desempenho, apesar das 3 séries terem horizontes temporais iguais, têm intervalos de observações muito diferentes e, como refere Chatfield (1988), as medidas de desempenho não são independentes da escala utilizada na contabilização. Neste sentido, as medidas de desempenho ME/RMSE/MAE/MAPE são apenas interpretadas no contexto de cada série e não inter-relacionáveis. De entre as medidas de desempenho, a mais facilmente interpretável no contexto é a MAPE que mede o erro absoluto médio percentual, ou seja, avalia a dimensão do erro em termos percentuais. Importa, contudo, ressalvar que foi possível utilizar a medida MAPE porque nenhuma série continha valores nulos, o que não seria possível se tal se verificasse. De entre os 3 modelos, o modelo MM é aquele ao qual estão associados valores menores da medida MAPE, valores esses de 3,7%; 1,6% e 3,3%. Assim, as estimativas anuais para os NP erram, no máximo, 3,7%; para os NBP 1,6% e, por último, para os NMBP 3,3%, erros esses muito pouco significativos e inferiores quando comparados com os modelos ARIMA e AE. De igual modo, a raíz do erro quadrático médio (RMSE) é, para as 3 séries, inferior no modelo das médias móveis, seguindo-se o modelo ARIMA e o modelo de AE.

Conclui-se que as medidas de desempenho registam, nos 3 casos, valores significativamente superiores para os modelos de previsão ARIMA e Alisamento Exponencial e, consequentemente, um pior desempenho preditivo. Assim, para as 3 séries NP/NBP e NMBP, o Modelo de Médias Móveis (MM, para $N=2$) é aquele com melhor capacidade preditiva do número de nascimentos respetivos e o Método de Alisamento Exponencial (AE) o que evidencia um pior poder preditivo.

Interpretação no contexto da prematuridade

O conceito de **prevalência** define-se, segundo a OMS, como a proporção de indivíduos com uma determinada doença/condição de saúde associada numa população num dado instante de tempo. No contexto deste estudo interessou, como complemento da análise das séries temporais, estudar a prevalência de NP/NBP e NMBP na totalidade dos nascimentos ocorridos e, cujos resultados se encontram no Anexo G. Se se retirar da análise os anos de 1989 a 1993 para os NP (período em que houve uma alteração na definição de prematuridade), nas três séries NP/NBP e NMBP, registou-se um aumento significativo na prevalência associada:

- Na série NP de 7,8% (1993) para 8,0% (2019); Na série NBP de 5,4% (1989) para 8,9% (2019) e, por último, na série NMBP de 0,5% (1989) para 1,1% (2019);

Os resultados anteriores corroboram as conclusões de Peixoto et al. (2002) na qual identificava, contabilizando apenas os anos de 1996 a 2000, uma tendência crescente na prevalência de nascimentos prematuros, em particular NMBP. Como referido no capítulo do Enquadramento teórico, apesar das causas da prematuridade serem ainda desconhecidas, os valores anteriores surgem acompanhados de um agravamento dos fatores de risco associados. Veja-se que, em particular, a proporção de nados vivos de partos gemelares atingiu os 3% em 2019 (Anexo B) e houve um aumento do nº de nados vivos nascidos no grupo etário das mães entre os 25 a 39 anos e >40 anos (Anexo F).

Com o intuito de fazer previsões para fora do horizonte temporal considerado, e verificar se os resultados anteriores se mantinham para as 3 séries, aplicaram-se os modelos ARIMA out of sample, AE e MM na previsão dos nascimentos entre 2020 e 2023. (Tabelas 4.19/4.20/4.21, Tabelas 4.25/4.26/4.27 e Anexo Q). Os valores previstos, nas 3 séries, apesar de estarem contidos nos intervalos de confiança considerados caracterizaram-se por estarem associados a erros de previsão significativos. Apesar do elevado poder preditivo do modelo ARIMA, comprovado na previsão in sample, a previsão out of sample revelou-se pouco eficaz, em parte porque a contabilização do nº de nascimentos totais foi afetada pelas circunstâncias da Covid19.

Capítulo 5. Conclusões

No decorrer do presente trabalho caracterizou-se e modelou-se o fenómeno demográfico dos nascimentos Prematuros/Baixo Peso e Muito Baixo Peso, em Portugal, no horizonte temporal de 1989 a 2019. Para que tal fosse possível, foi aplicada a *Metodologia Box-Jenkins* nas suas 3 principais etapas: identificação, estimação e avaliação do diagnóstico do modelo.

Na primeira fase de identificação, representaram-se graficamente as 3 séries, fez-se uma análise preliminar das FAC e FACP das séries originais e estudou-se a estacionaridade de cada uma delas. Seguidamente, nos casos das séries NBP e NMBP, em que a hipótese de estacionaridade foi rejeitada, procedeu-se à aplicação do operador diferenciação de ordem 1 e, após a diferenciação, foram novamente analisadas as FAC e FACP das séries diferenciadas. Na segunda fase da metodologia, foi identificado e estimado o modelo ARIMA (2,0,0), que na fase subsequente se substituiu por ARIMA (1,0,0), o modelo ARIMA (2,1,2) e ARIMA (2,1,2) para as séries NP, NBP e NMBP e, por último, estimados os respetivos parâmetros. Na fase final da metodologia, a fase de diagnóstico, avaliou-se a qualidade estatística e de ajustamento do modelo, através da inspeção dos resíduos, dos 3 modelos propostos.

Da análise anterior retira-se que os modelos previamente estimados cumpriam os requisitos propostos e estavam reunidas as condições para uma posterior previsão. Seguidamente, previu-se o nº de nascimentos das categorias anteriores no horizonte temporal de curto prazo (4 anos), através da aplicação de 3 modelos de previsão distintos (ARIMA, Alisamento Exponencial e Médias Móveis) e comparou-se posteriormente a capacidade preditiva e cada um deles. Concluiu-se que o modelo de Médias Móveis (MM para N=2) era aquele com melhor capacidade preditiva para as 3 séries NP,NBP e NMBP. Os modelos escolhidos foram, tanto quanto possível, descritivos ajudando à compreensão dos fenómenos a que dizem respeito, situação essa de especial importância no domínio da natalidade. Conclui-se, portanto, que os objetivos delineados na introdução foram devidamente alcançados.

5.1. Limitações

Uma das limitações surge do nº de observações consideradas, nas 3 séries utilizam-se 30 observações referentes a 30 anos, sendo que, idealmente, o nº de observações deveria ser superior. Chatfield (2004) refere que as previsões anuais baseadas em registos mensais, quando possíveis de obter, geram melhores resultados. De facto, a importância estratégica dos modelos ARIMA reside na capacidade de modelar estruturas complexas e interdependentes no tempo, estruturas essas que ocorrem com maior frequência em intervalos de tempo mensais, precisamente pela ocorrência de flutuações sazonais. Por essa razão deveriam ser utilizados dados mais precisos (observações mensais, trimestrais, semestrais) contudo o estudo de uma população está condicionado pela disponibilidade das fontes estatísticas que, neste caso, forneciam apenas informação anual.

Um outro inconveniente resulta dos modelos de Médias Móveis, identificados na aplicação às 3 séries NP/ NBP e NMBP como aqueles com melhor capacidade preditiva, darem demasiada importância aos últimos

anos (mais próximos da data de previsão), resultando numa tendência prevista similar à verificada nos anos anteriores. Assim, contrariamente aos modelos ARIMA, apesar desta abordagem resultar em previsões tendencialmente corretas, não atende à possibilidade de uma inversão de tendência.

De referir que os modelos de previsão constituem uma área de contínua investigação com o objetivo de melhorar a modelação e o rigor das previsões. Nos modelos ARIMA, na fase de identificação do modelo, surgem diferentes modelos com ajustes semelhantes, como foi o caso dos modelos sugeridos para a série NP, que, mais tarde na fase de diagnóstico se revelam inapropriados exigindo tempo do analista e um recomeço do processo de estimação.

5.2. Sugestões para investigações futuras

Salienta-se que os modelos de previsão de séries temporais constituem uma área de contínua investigação com o objetivo de melhorar a modelação e o rigor das previsões, pelo que se apresenta em seguida uma sugestão que abre portas a investigações futuras.

Uma sugestão para investigações futuras passa pela modelação e aplicação de um modelo econométrico que estude a influência de múltiplos fatores, entre os quais fatores demográficos e socioeconómicos no fenómeno da natalidade, em particular nos NP/ NBP e NMBP, e que permita a melhoria das previsões comparativamente aos modelos univariados aplicados neste estudo. Como referido, os modelos univariados mostraram ser capazes de fornecer boas previsões, mas dependem unicamente de valores associados a observações passadas. Contudo, sabe-se que existem outros fatores como o sexo, a idade da mãe, a conjuntura económica ou os cuidados de saúde prestados que influem significativamente no nº de nascimentos ocorridos desta natureza, pelo que seriam fatores qualitativos e quantitativos a incluir em trabalhos futuros.

Capítulo 6. Referências Bibliográficas

Almeida, A. N. D., André, I. M., & Lalanda, P. (2002). Novos padrões e outros cenários para a fecundidade em Portugal. *Análise Social*, *XXXVII*(163), 371-409. http://hdl.handle.net/10451/36782

Alpuim, T. (1998). *Séries Temporais*, Associação de Estudantes da Faculdade de Ciências de Lisboa.

Beck, S., Wojdyla, D., Say, L., Betran, A. P., Merialdi, M., Requejo, J. H., Rubens, C., Menon, R., & Look, P. F. V. (2009). *The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity*. Bulletin of the World Health Organization.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1993). *Time Series Analysis: Forecasting and Control* (3.ª ed.). John Wiley & Sons.

Carrilho, M. J., & Peixoto, J. (1993). *A evolução demográfica em Portugal entre 1981 e 1992* (Vol. 31). Instituto Nacional de Estatística.

Chatfield, C. (2004). *The analysis of time series: an introduction*. CRC Press.

Chatfield, C., & Xing, H. (2019). *The Analysis of Time Series: An Introduction with R* (7.ª ed.). Chapman and Hall/CRC.

Chawanpaiboon, S., Vogel, J. P., Moller, A. B., Lumbiganon, P., Petzold, M., Hogan, D., Landoulsi, S., Jampathong, N., Kongwattanakul, K., Laopaiboon, M., Lewis, C., Rattanakanokchai, S., Teng, D. N., Thinkhamrop, J., Watananirun, K., Zhang, J., Zhou, W., & Gülmezoglu, A. M. (2019). Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health*, *7*(1), e37-e46. https://doi.org/10.1016/s2214-109x(18)30451-0

Comissão Nacional de Saúde Materna e Infantil. (1989). *Cuidados de Saúde Materna e Neo-Natal*.

d'Uva, T. B. (1999). Nados Vivos: Estimação e Análise. *Revista de Estatística*, *12*(3ºQuad).

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2018). *Forecast: Forecasting functions for time series and linear models*. R package.

Instituto Nacional de Estatística. (2005). *Documento Metodológico Nados Vivos*.

Instituto Nacional de Estatística. (2012). *Indicadores Sociais: 2011*. https://www.ine.pt/xurl/pub/149279938

Instituto Nacional de Estatística. (2020). *Estatísticas Demográficas: 2019*. https://www.ine.pt/xurl/pub/71882686

Kitagawa, G. (2010). *Introduction to time series modeling*. CRC Press.

Lee, R. D. (1974). Forecasting births in post-transition population: stochastic renewal with serially correlated fertility. *Journal of the American Statistical Association*, *69*(347), 607-617. https://doi.org/10.1080/01621459.1974.10480177

Makridakis, S., Wheelwright, S., & Hyndman, R. (1998). *Forecasting: Methods and Applications* (3.ª ed.). John Wiley & Sons.

McCleary, R., & Hay, R. A. (1980). *Applied Time Series Analysis for the Social Science*. Sage.

Mimoso, G., & Almeida, A. (2018). *Registo MBP de 1996 a 2018*. Unidade de Formação Contínua em Neonatologia CMIN/ICBAS

Murphy, D. J. (2007). Epidemiology and environmental factors in preterm labour. *Best Practice & Research Clinical Obstetrics and Gynaecology*, *21*(5), 773-789. https://doi.org/10.1016/j.bpobgyn.2007.03.001

Murphy, M. M., & McLoughlin, G. (2015). *Born too soon: preterm birth in Europe trends, causes and prevention*. Entre Nous, *81*, 10-12.

Murteira, B. J. F., Muller, D., & Turkman, K. F. (1993). *Análise de Sucessões Cronológicas*. McGraw-Hill.

Pankratz, A. (1983). *Forecasting with Univariate Box-Jenkins Models*. Wiley & Sons, Inc.

Peixoto, J. C., Guimarães, H., Machado, M.-C., V, M., Mimoso, G., Neto, T., Tomé, T., Virella, D., & Peso, R. (2002). *Nascer Prematuro em Portugal. Estudo Multicêntrico Nacional 1996 - 2000*. Fundação Bial. https://doi.org/10.13140/2.1.3716.6086

Peixoto, J. C. (1999). Registo Nacional dos Recém-Nascidos de Muito Baixo Peso. Rede de Investigação Neonatal Nacional. *Acta Pediátrica Portuguesa*, 6, 485-91.

Rodrigues, T., & Barros, H. (1998). Factores de Risco para Trabalho de Parto Pré-Termo. *Acta Médica Portuguesa*, *11*, 901-905.

Saboia, J. L. (1977). Autoregressive Integrated Moving Average (ARIMA) Models for Birth Forecasting. *Journal of the American Statistical Association*, *72*(358), 264-270. https://doi.org/10.2307/2286787

World Health Organization. (2012). *Born too soon: the global action report on preterm birth*. WHO.

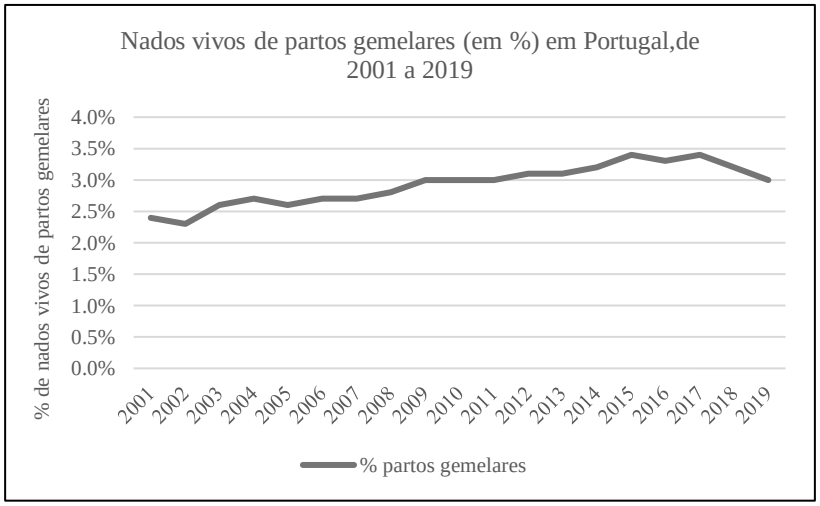
Zeitlin, J., Szamotulska, K., Drewniak, N., Mohangoo, A. D., Chalmers, J., Sakkeus, L., Irgens, L., Gatt, M., Gissler, M., & Blondel, B. (2013). Preterm birth time trends in Europe: a study of 19 countries. *BJOG: An International Journal of Obstetrics & Gynaecology*, *120*(11), 1356-1365. https://doi.org/10.1111/1471-0528.12281

Anexos

Anexo A – Definição dos Principais Conceitos

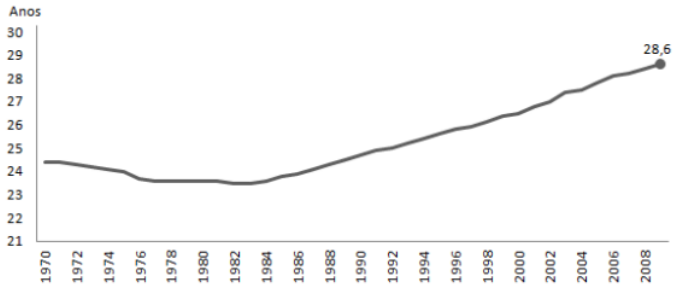
- *Idade gestacional*: Duração da gestação, a qual é expressa em dias ou semanas completas e é calculada a partir do primeiro dia do último período menstrual normal;
- *Nado Vivo*: O produto do nascimento vivo;
- *Nascimento de Baixo Peso (NBP)*: Nascimento de um nado vivo com peso compreendido entre 1500g e 2500g;
- *Nascimento de Muito Baixo Peso (NMBP)*: Nascimento de um nado vivo com peso compreendido entre 1500g e 500g; Também designado Nascimento de extremo baixo peso;
- *Nascimento Prematuro (NP)*: Nascimento que ocorre antes das 37 semanas de gestação completas;
- *Peso à nascença*: Primeira medida de peso (em gramas) do nado vivo obtida após o nascimento. feita de preferência durante a primeira hora de vida, antes que ocorra uma significativa perda de peso pós-natal;
- *Prevalência:* Proporção de indivíduos numa população doentes num dado instante de tempo;
- *Taxa bruta de natalidade:* Número de nados vivos ocorrido durante um determinado período de tempo, normalmente um ano civil, referido à população média desse período (habitualmente expressa em número de nados vivos por 1000 habitantes);
- *Taxa de Prematuridade:* Número de nascimentos prematuros em relação ao número total de nascimentos no ano considerado;

Anexo B – Nados vivos de partos gemelares (em%) em Portugal, de 2001 a 2019

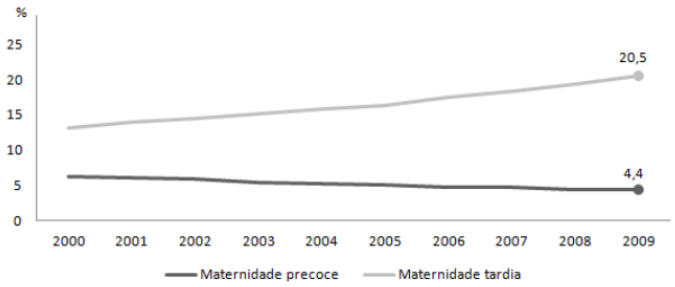


Fonte: INE, I.P, Nados Vivos

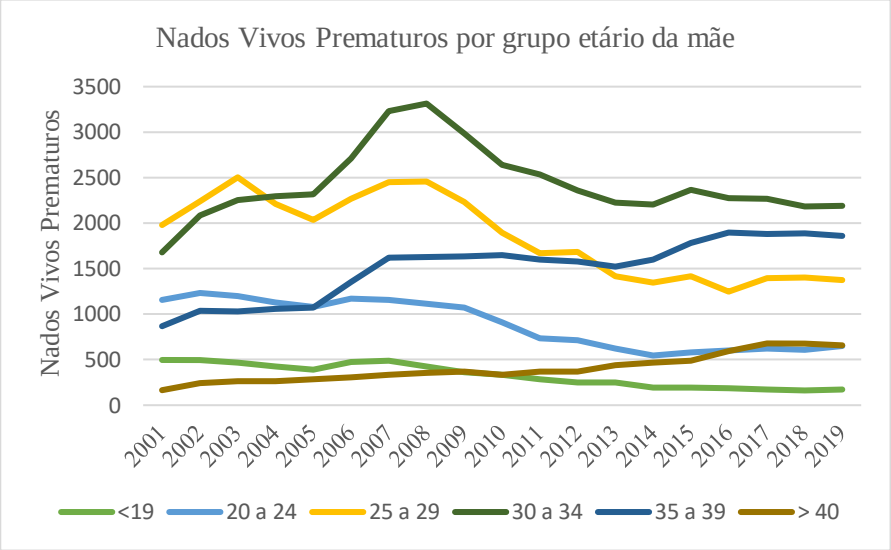
Anexo C – Idade Média da mãe ao nascimento do primeiro filho em Portugal, 1970-2009



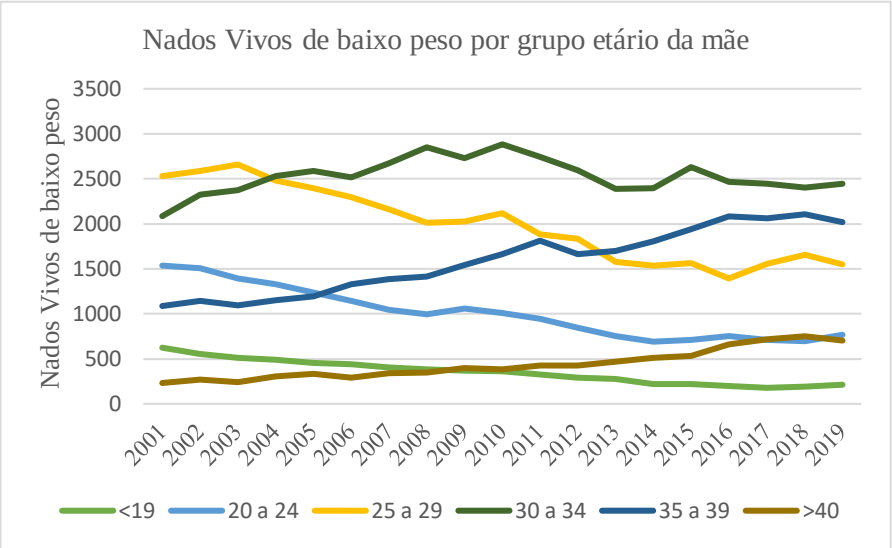
Anexo D – Maternidade Precoce e Maternidade Tardia



Anexo E – Nados Vivos Prematuros por Grupo Etário da Mãe



Anexo F – Nados Vivos de baixo peso por grupo etário da mãe



Anexo G - Nascimentos Totais, Prematuros, de Baixo Peso e de Muito Baixo Peso

Anos	NT	NP		NBP		NMBP	
		n	Prevalência	*n*	Prevalência	*n*	Prevalência
1989	118 560	14928	12,6%	6416	5,4%	611	0,5%
1990	116 383	13976	12,0%	6526	5,6%	736	0,6%
1991	116 415	12896	11,1%	6494	5,6%	651	0,6%
1992	115 018	9813	8,5%	6165	5,4%	682	0,6%
1993	114 030	8937	7,8%	6756	5,9%	778	0,7%
1994	109 287	8956	8,2%	6658	6,1%	757	0,7%
1995	107 184	6794	6,3%	6446	6,0%	760	0,7%
1996	110 363	7582	6,9%	7051	6,4%	892	0,8%
1997	113 047	7556	6,7%	7384	6,5%	921	0,8%
1998	113 510	7560	6,7%	7639	6,7%	974	0,9%
1999	116 038	7026	6,1%	8568	7,4%	1183	1,0%
2000	119 455	7070	5,9%	8421	7,0%	1123	0,9%
2001	112 774	6346	5,6%	8097	7,2%	997	0,9%
2002	114 383	7328	6,4%	8386	7,3%	1053	0,9%
2003	112 515	7716	6,9%	8272	7,4%	996	0,9%
2004	109 298	7391	6,8%	8290	7,6%	1020	0,9%
2005	109 399	7167	6,6%	8204	7,5%	1037	0,9%
2006	105 449	8286	7,9%	8012	7,6%	1011	1,0%
2007	102 492	9280	9,1%	8017	7,8%	1070	1,0%
2008	104 594	9293	8,9%	8008	7,7%	1084	1,0%
2009	99 491	8657	8,7%	8124	8,2%	1176	1,2%
2010	101 381	7759	7,7%	8416	8,3%	1215	1,2%
2011	96 856	7191	7,4%	8135	8,4%	1138	1,2%
2012	89 841	6963	7,8%	7644	8,5%	1148	1,3%
2013	82 787	6476	7,8%	7165	8,7%	1025	1,2%
2014	82 367	6363	7,7%	7163	8,7%	1025	1,2%
2015	85 500	6829	7,8%	7590	8,9%	1106	1,3%
2016	87 126	6801	7,8%	7550	8,7%	1118	1,3%
2017	86 154	7011	8,1%	7667	8,9%	1116	1,3%
2018	87 020	6922	8,0%	7804	9,0%	985	1,1%
2019	86 579	6913	8,0%	7694	8,9%	912	1,1%

Anexo H - Valores dos Nascimentos das séries originais e das séries diferenciadas

NP		NBP			NMBP		
t	X_t	t	X_t	∇X_t	t	X_t	∇X_t
1989	14928	1989	6416		1989	611	
1990	13976	1990	6526	110	1990	736	125
1991	12896	1991	6494	-32	1991	651	-85
1992	9813	1992	6165	-329	1992	682	31
1993	8937	1993	6756	591	1993	778	96
1994	8956	1994	6658	-98	1994	757	-21
1995	6794	1995	6446	-212	1995	760	3
1996	7582	1996	7051	605	1996	892	132
1997	7556	1997	7384	333	1997	921	29
1998	7560	1998	7639	255	1998	974	53
1999	7026	1999	8568	929	1999	1183	209
2000	7070	2000	8421	-147	2000	1123	-60
2001	6346	2001	8097	-324	2001	997	-126
2002	7328	2002	8386	289	2002	1053	56
2003	7716	2003	8272	-115	2003	996	-57
2004	7391	2004	8290	19	2004	1020	24
2005	7167	2005	8204	-86	2005	1037	17
2006	8286	2006	8012	-192	2006	1011	-26
2007	9280	2007	8017	5	2007	1070	59
2008	9293	2008	8008	-9	2008	1084	14
2009	8657	2009	8124	116	2009	1176	92
2010	7759	2010	8416	292	2010	1215	39
2011	7191	2011	8135	-281	2011	1138	-77
2012	6963	2012	7644	-491	2012	1148	10
2013	6476	2013	7165	-479	2013	1025	-123
2014	6363	2014	7163	-2	2014	1025	0
2015	6829	2015	7590	427	2015	1106	81
2016	6801	2016	7550	-40	2016	1118	12
2017	7011	2017	7667	117	2017	1116	-2
2018	6922	2018	7804	137	2018	985	-131
2019	6913	2019	7694	-110	2019	912	-73

Anexo I - Resultados Teste Shapiro-Wilk (Séries originais)

Teste *Shapiro-Wilk*		
Série	**W**	***pvalue***
NP	0,71969	2,3e-06
NBP	0,92187	0,02647
NMBP	0,91655	0,01914

Anexo J - Valores empíricos das FAC e FACP das séries NBP e NMBP diferenciadas uma vez

Série NBP diferenciada															
lag	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
FAC	1.000	0.071	-0.137	0.234	-0.283	-0.027	0.247	-0.173	-0.019	-0.009	-0.080	0.134	-0.012	-0.084	-0.203
FACP	-	0.071	-0.143	0.263	-0.388*	0.206	0.021	-0.039	-0.061	-0.122	0.162	-0.013	-0.029	-0.070	-0.324

Série NMBP diferenciada															
lag	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
FAC	1.000	0.071	-0.137	0.235	-0.283	-0.027	0.247	-0.173	-0.019	-0.009	-0.080	0.134	-0.012	-0.084	-0.203
FACP	-	0.071	-0.143	0.263	-0.388	0.206	0.021	-0.039	-0.061	-0.122	0.162	-0.013	-0.029	-0.070	-0.324

Anexo K - Avaliação da qualidade estatística - Significância estatística dos parâmetros dos modelos

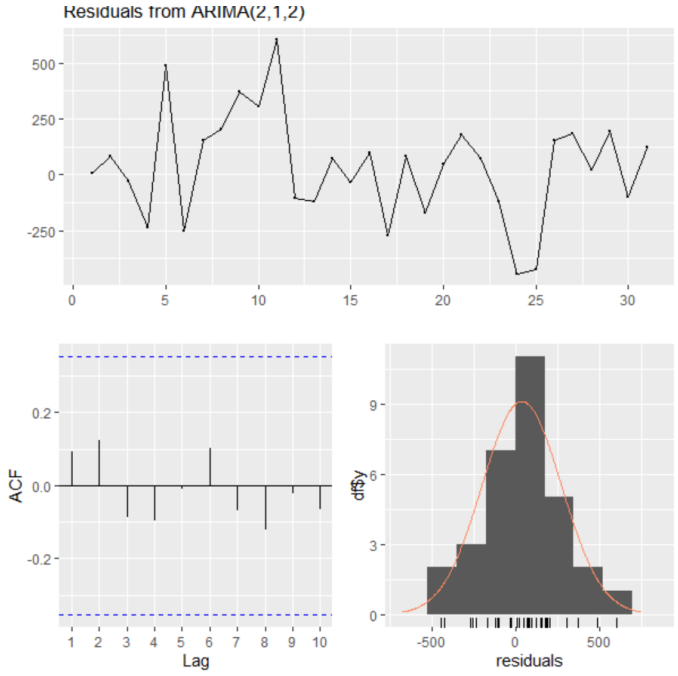
		IC (95%)	
		Limite inf.	*Limite sup.*
NP	AR1	0,9389655	1,605301
	AR2	-0,6836296	2,773666e-02
NBP	AR1	-1,4220537	-1,002686
	AR2	-1,0382549	-0,574338
	MA1	1,3775777	1,996762
	MA2	0,6764848	1,323398
NMBP	AR1	-1,5963379	-1,1210730
	AR2	-1,0213453	-0,5470083
	MA1	1,4980479	2,2051314
	MA2	0,6331014	1,3668603

Anexo L - Escolha de um novo modelo ARIMA para a série NP

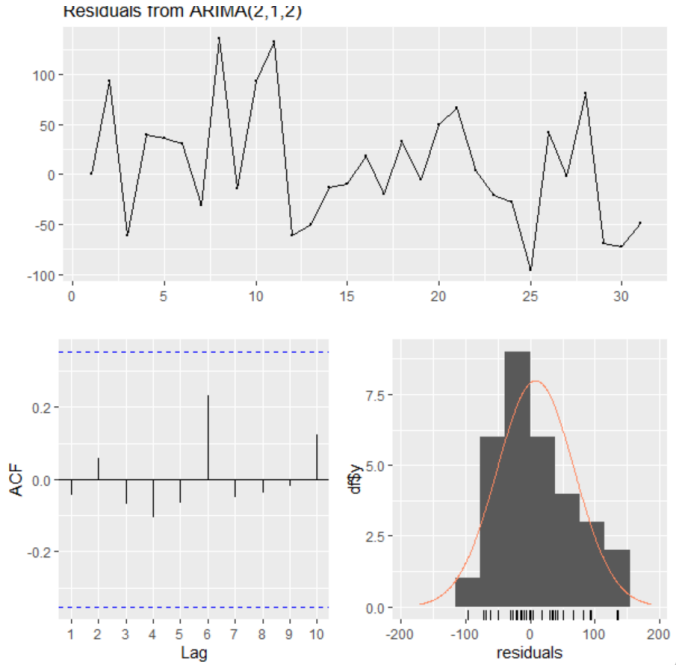
Série NP					
ARIMA (1,0,1)	*AIC*	*BIC*	IC (95%)		
	517,53	523,2664	Parâmetros	*Limite inf.*	*Limite sup.*
			AR1	0,82728193	1,065893e+00
			MA1	-0,07293505	5,873334e-04
ARIMA (1,0,0)	517,84	522,1401	AR1	0,8890357	1,043996

Anexo M - Análise dos Resíduos

- Série NBP



- Série NMBP

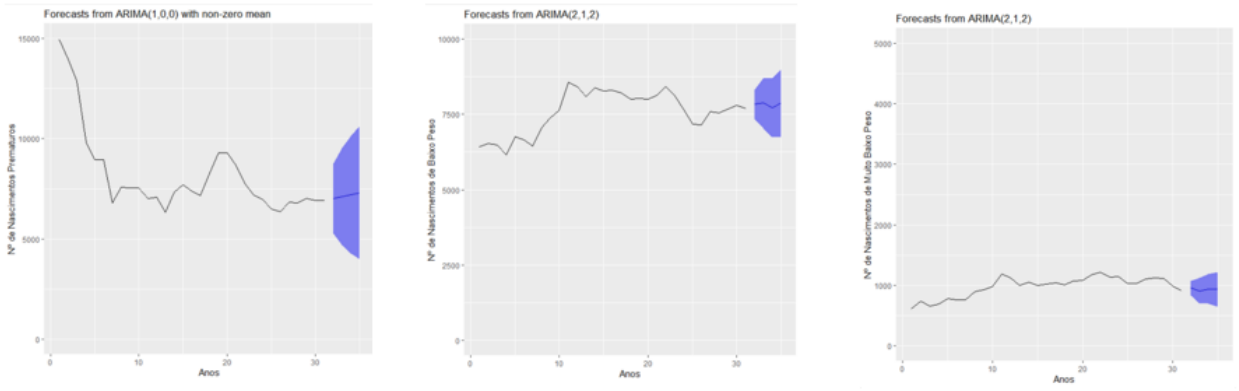


Anexo N - Resultados do teste *Ljung Box*

Teste *Ljung Box*						
	NP		**NBP**		**NMBP**	
lag	*Q*	*pvalue*	*Q*	*pvalue*	*Q*	*pvalue*
1	0,57792	0,4471	0,2907	0,5898	0,0696	0,7920
2	0,67614	0,7131	0,8255	0,6618	0,1850	0,9116
3	0,74747	0,8620	1,1120	0,7742	0,3515	0,9501
4	0,81818	0,9360	1,4535	0,8349	0,7678	0,9427
5	0,86593	0,9726	1,4564	0,9180	0,9383	0,9674
6	1,5040	0,9592	1,8874	0,9298	3,1408	0,7910
7	1,6675	0,9759	2,0903	0,9547	3,2472	0,8612
8	1,7697	0,9827	2,7431	0,9494	3,8106	0,9134
9	2,4298	0,9827	2,7639	0,9729	3,3272	0,9499
10	2,6923	0,9878	2,9794	0,9819	4,0563	0,9448

Anexo O - Previsão Método ARIMA out of sample

- Representação gráfica previsão ARIMA out of sample – NP; NBP e NMBP

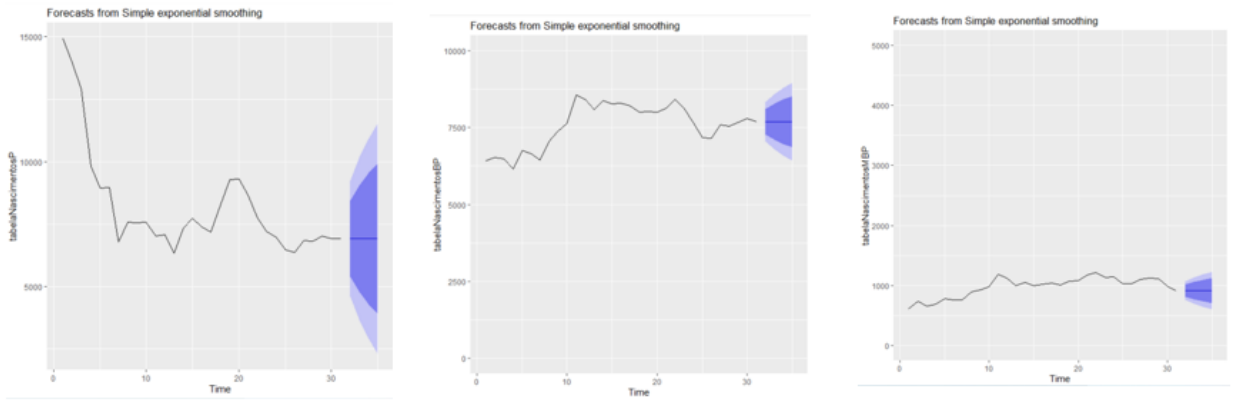


(A) (B) (C)

- Previsão da série NP usando o modelo ARIMA (1,0,0) para um *IC de 95%* e *h=4; (A);*
- Previsão da série NBP usando o modelo ARIMA (2,1,2) para um *IC de 95%* e *h=4; (B);*
- Previsão da série NMBP usando o modelo ARIMA (2,1,2) para um *IC de 95%* e *h=4; (C)*

Anexo P - Previsão Método Alisamento Exponencial

- Representação gráfica previsão AE *out of sample* – NP; NBP e NMBP



(A) (B) (C)

  - Previsão da série NP pelo método AE para um IC de 95% e h=4; (A);
  - Previsão da série NBP pelo método AE para um IC de 95% e h=4; (B);
  - Previsão da série NMBP pelo método AE para um IC de 95% e h=4; (C)

- Output explicativo Método Alisamento Exponencial - NP

```
ETS(A,N,N)

Call:
 ets(y = tabelaNascimentosP, model = "ANN")

  Smoothing parameters:
    alpha = 0.9999

  Initial states:
    l = 10889.7462

  sigma:  1179.964

     AIC     AICc      BIC
548.9270 549.8159 553.2290
```

- Output explicativo Método Alisamento Exponencial - NBP

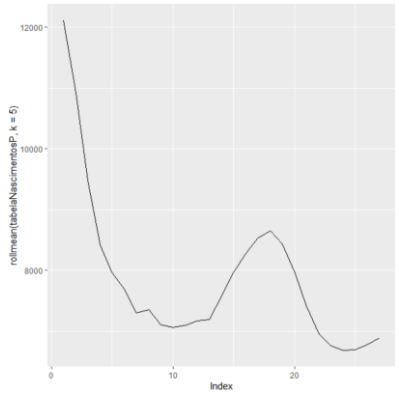
```
ETS(A,N,N)

Call:
 ets(y = tabelaNascimentosBP, model = "ANN")

  Smoothing parameters:
    alpha = 0.9999

  Initial states:
    l = 6417.0895

  sigma:  325.6284

     AIC     AICc      BIC
469.1031 469.9920 473.4051
```

- Output explicativo Método Alisamento Exponencial - NMBP

```
ETS(A,N,N)

Call:
 ets(y = tabelaNascimentosMBP, model = "ANN")

  Smoothing parameters:
    alpha = 0.9824

  Initial states:
    l = 613.2994

  sigma:  80.8604

     AIC     AICc      BIC
382.7351 383.6240 387.0371
```

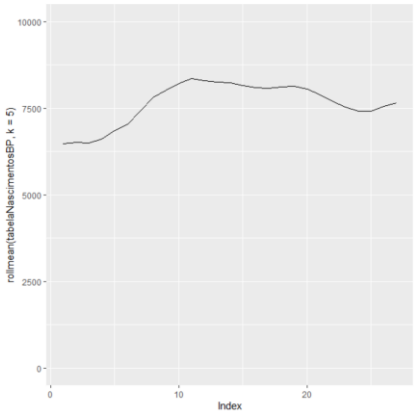
Anexo Q - Previsão Modelo Médias Móveis Simples

- Previsão in sample NP pelo Modelo das Médias Móveis Simples para N=5 e h=4



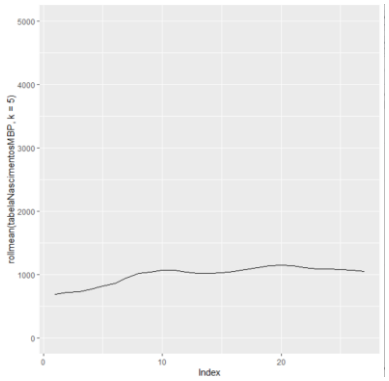
Ano	Previsão NP	Valor registado	|erro $_{prev}$|	RMSE
2016	6686,4	6801	115	184,396
2017	6696,0	7011	315	
2018	6785,2	6922	137	
2019	6895,2	6913	18	

- Previsão in sample NBP pelo Modelo das Médias Móveis Simples para N=5 e h=4



Ano	Previsão NBP	Valor registado	|erro $_{prev}$|	RMSE
2016	7422,4	7550	128	185,454
2017	7427,0	7667	240	
2018	7554,8	7804	250	
2019	7661,0	7694	33	

- Previsão in sample NMBP pelo Modelo das Médias Móveis Simples para N=5 e h=4



Ano	Previsão NMBP	Valor registado	|erro $_{prev}$|	RMSE
2016	1084,4	1118	34	83,741
2017	1078,0	1116	38	
2018	1070,0	985	85	
2019	1047,4	912	135	

- Previsão *out of sample*

Série NP

Ano	Previsão NBP	IC de 90%		IC de 95%		Valor registado	|erro $_{prev}$|
		Lim inf.	*Lim sup.*	*Lim inf*	*Lim sup*		
2020	6878,3	6256,190	7500,444	5926,856	7829,778	5765	1113
2021	6846,9	5564,962	8128,963	4886,312	8807,613	5997	849
2022	6821,9	4830,677	8813,081	3776,599	9867,159	-	-
2023	6801,8	4082,214	9521,411	2642,546	10961,079	-	-

Série NBP

Ano	Previsão NBP	IC de 90%		IC de 95%		Valor registado	|erro $_{prev}$|
		Lim inf.	*Lim sup.*	*Lim inf*	*Lim sup*		
2020	7748,9	7436,361	8061,636	7270,861	8227,137	6663	1085
2021	7748,9	7306,884	8191,113	7072,843	8425,154	6708	1040
2022	7748,9	7207,531	8290,467	6920,895	8577,103	-	-
2023	7748,9	7123,770	8374,227	6792,795	8705,203	-	-

Série NMBP

Ano	Previsão NBP	IC de 90%		IC de 95%		Valor registado	|erro $_{prev}$|
		Lim inf.	*Lim sup.*	*Lim inf*	*Lim sup*		
2020	948,5	876,466	1020,554	838,329	1058,692	730	218
2021	948,5	846,630	1050,390	792,698	1104,322	789	159
2022	948,5	823,735	1073,285	757,603	1139,337	-	-
2023	948,5	804,434	1092,587	728,164	1168,856	-	-

Anexo R - Explicitação dos principais comandos utilizados

Comandos de instalação	
install.packages("stats") *library(stats)*	Instala a *library* "stats" e carrega-a no *workspace*;
install.packages("TSA") *library (TSA)*	Instala a *library* "TSA" e carrega-a no *workspace;*
install.packages("tseries") *library(tseries)*	Instala a *library* "tseries" e carrega-a no *workspace;*
Comandos gerais	
data. frame ( )	Cria um objecto do tipo data frame; Um *data frame* é uma estrutura bidimensional semelhante a uma matriz na qual cada coluna contém as variáveis a estudar e cada linha os respetivos valores associados;
summary ()	Calcula o mínimo, 1º,2º e 3º quartil, a média, mediana e o máximo de uma função;
diff ()	Diferencia uma série temporal;
plot () ou ggplot()	Representa graficamente uma função;
hist ()	Representa o histograma da função;
confint ()	Calcula o intervalo de confiança para o coeficiente estimado para um determinado nível de confiança;

Package	**Função**	**Descrição**
stats	*acf ()*	Representa a função de autocorrelação;
	pacf ()	Representa a função de autocorrelação parcial;
	box.test ()	Aplica o teste Box-Pierce;
	checkresiduals ()	Executa o diagnóstico dos resíduos incluindo os resíduos standartizados, as funções ACF/PACF e o teste Ljung Box;
tseries	*ts ()*	Transforma a função num objecto do tipo *time series*;
	arma ()	Estima um modelo autoregressivo de médias móveis,
	arima ()	Estima um modelo autoregressivo integrado e de médias móveis,
	adf.test () *pp.test ()* *kpss.test ()*	Executa o teste *Augmented Dickey Fuller* (ADF) para a hipótese nula de que existe raiz unitária; Executa o teste de raiz unitária PP para a hipótese nula de que a série é estacionária; Executa o teste de raiz unitária KPSS para a hipótese nula de que a série é estacionária;

- **Código *R* utilizado:**

```
#Análise de Dados

library(tseries)
library(forecast)
library(tidyverse)
library(ggplot2)
library(urca)
library(TSstudio)
library(zoo)



# NASCIMENTOS PREMATUROS

tabelaAnosP <-c(Ano=1989:2019)
tabelaNascimentosP <-c(NascimentosP = c(14928,  13976,  12896,  9813,  8937,  8956,  6794,  7582,
        7556,  7560,  7026,  7070,
6346,7328,7716,7391,7167,8286,9280,9293,8657,7759,7191,6963,6476,6363,6829,6801,7011,6922,6913))

#data frame tabelaDados
tabelaDados <- data.frame(tabelaAnosP,tabelaNascimentosP)
tabelaDados

#Gráfico da ST NP
plot(tabelaDados,type="l",xlab="Anos",ylab="Nascimentos Prematuros",ylim=c(0,15000))
ggplot(tabelaDados,aes(x=tabelaAnosP ,y=tabelaNascimentosP))+geom_line(color="red")+ ylab("Nascimentos
Prematuros") + xlab("Anos")

#Gráfico em %
tabelaAnosPPerc <-c(Ano=1989:2019)
tabelaNascimentosPPerc <-c (NascimentosPPerc =
c(13,12,11,09,08,08,06,07,07,07,06,06,06,06,07,07,07,08,09,09,09,08,07,08,08,08,08,08,08,08,08,08
))
tabelaDadosPerc <- data.frame(tabelaAnosPPerc,tabelaNascimentosPPerc)
tabelaDadosPerc
ggplot(tabelaDadosPerc,aes(x=tabelaAnosPPerc ,y=tabelaNascimentosPPerc))+ geom_line(color="red")+
scale_y_continuous(limits = c(0, 100)) + ylab("Nascimentos Prematuros (%)") + xlab("Anos")

#Estatística descritiva NP
summary(tabelaNascimentosP)
windowsFonts(A = windowsFont("Times New Roman"))
hist(tabelaNascimentosP,main="Frequência de NP em Portugal no período entre 1989 e 2019", xlab="N° de NP
registados",
     col="lightgrey",family="A")

#Testar a normalidade (série original NP)
qqnorm(tabelaNascimentosP)
decompose(tabelaNascimentosP)
shapiro.test(tabelaNascimentosP)
```

```
#Análise estacionaridade
acftab <- acf(tabelaNascimentosP) #autocorrelação
pacftab <- pacf(tabelaNascimentosP) #autocorrelação parcial
acf(tabelaNascimentosP,lag=5,pl=FALSE)

#Análise da estacionaridade (Série original NP)
testdf0 <- ur.df(tabelaNascimentosP, type="drift")
summary(testdf0)
testpp0 <- ur.pp(tabelaNascimentosP, model="constant",type="Z-tau", lags="short")
summary(testpp)
testkpss0 <-ur.kpss(tabelaNascimentosP,type="tau", lags="short")
summary(testkpss0)

#FAC e FACP
acf(tabelaNascimentosP)
pacf(tabelaNascimentosP)

#FAC e FACP - valores por lag
acf_valuesP = acf(tabelaNascimentosP,plot=F)
acf_valuesP
pacf_valuesP = pacf(tabelaNascimentosP,plot=F)
pacf_valuesP


#Ajustar um modelo auto ARIMA

tabelaDadosmodelP2 <- auto.arima(tabelaNascimentosP,ic="aic",trace=TRUE)

#Estimar modelos ARIMA
arima(tabelaNascimentosP,order = c(2, 0,0))
summary(arima(tabelaNascimentosP,order = c(0, 1, 0)))
BIC(arima(tabelaNascimentosP,order = c(0, 1, 0)))

#Melhor modelo: ARIMA (2,0,0)
arima(tabelaNascimentosP,order = c(2, 0, 0))
summary(arima(tabelaNascimentosP,order = c(2, 0, 0)))

cor(arima(tabelaNascimentosP,order = c(2, 0, 0)))

#Análise de resíduos
tsdiag(arima(tabelaNascimentosP,order = c(1,0,0)))
shapiro.test(residuals(arima(tabelaNascimentosP,order = c(1,0,0))))
checkresiduals(arima(tabelaNascimentosP,order = c(1,0,0)))
Box.test(residuals(arima(tabelaNascimentosP,order = c(1,0,0))),lag=14,type=c("Box-Pierce"))

#Significância dos parâmetros
confint(arima(tabelaNascimentosP,order = c(1,0,0)))

#Validar o modelo
Box.test(tabelaDadosmodel$residuals,lag=2,type="Ljung-Box")

#Análise dos resíduos
plot(residuals(tabelaDadosmodel))
acf(residuals(tabelaDadosmodel))
pacf(residuals(tabelaDadosmodel))
```

```
qqnorm(residuals(tabelaDadosmodel))
checkresiduals(arima(tabelaNascimentosP,order = c(1,0,0)))
tsdiag(arima(tabelaNascimentosP,order = c(1,0,0)))

#Forecast - Previsão NP

#Forecast out of sample
autoplot(forecast(arima(tabelaNascimentosP,order = c(1,0,0)),level=c(90),h=4)) +
scale_y_continuous(limits = c(0, 15000))+ labs(y = "N° de Nascimentos Prematuros", x = "Anos")
forecast(arima(tabelaNascimentosP,order = c(1,0,0)),level=c(95),h=4)
accuracy(arima(tabelaNascimentosP,order = c(1,0,0)),level=c(95),h=4)

#Forecast in sample
forP <- window(tabelaNascimentosP,start=1,end=26)
forecast(arima(forP,order=c(1,0,0)),level=c(90),h=4)
f_in_sample <- forecast(arima(forP,order=c(1,0,0)),level=c(95),h=4)
accuracy(forecast(arima(forP,order=c(1,0,0)),level=c(95),h=4))
autoplot(forecast(arima(forP,order=c(1,0,0)),level=c(95),h=4),PI=FALSE)+ labs(y = "N° de Nascimentos
Prematuros", x = "Anos")
accuracy(forecast(arima(forP,order=c(1,0,0)),level=c(95),h=15))

# Forecast NP - Alisamento Exponencial
ses(forP,h=4,level = c(95))
autoplot(ses(tabelaNascimentosP,h=4))
aeP <- ets(forP, "ANN")
aeP
accuracy(ses(forP,h=4))

# Forecast NP - Médias Móveis
rm <- rollmean(tabelaNascimentosP,k=2)
rm
accuracy(rollmean(tabelaNascimentosP,k=5))
autoplot.zoo(rollmean(tabelaNascimentosP,k=5))
mae(rollmean(tabelaNascimentosP,k=5))

# NASCIMENTOS DE BAIXO PESO

tabelaAnosBP <-c(Ano=1989:2019)
```

```
tabelaNascimentosBP <-c(Nascimentos = c(6416,  6526,  6494,  6165,  6756,  6658,  6446,  7051,
        7384,  7639,  8568,  8421,
8097,8386,8271,8290,8204,8012,8017,8008,8124,8416,8135,7644,7165,7163,7590,7550,7667,7804,7694))

#data frame tabelaDados
tabelaDadosBP <- data.frame(tabelaAnosBP,tabelaNascimentosBP)
tabelaDados



#gráfico da ST
plot(tabelaDadosBP,type="l",xlab="Anos",ylab="Nascimentos de Baixo Peso",ylim=c(0,10000))
ggplot(tabelaDadosBP,aes(x=tabelaAnosBP
,y=tabelaNascimentosBP))+geom_line(color="red")+scale_y_continuous(limits=c(0, 10000)) +
ylab("Nascimentos de Baixo Peso") + xlab("Anos")



#Gráfico em %
tabelaAnosBPPerc <-c(Ano=1989:2019)
tabelaNascimentosBPPerc <-c (NascimentosBPPerc =
c(05,06,06,05,06,06,06,06,07,07,07,07,07.2,07.3,07.4,07.6,07.5,07.6,07.8,07.7,08.2,08.3,08.4,08.5,08.7,0
8.7,08.9,08.7,08.9,09.0,08.9))
tabelaDadosBPPerc <- data.frame(tabelaAnosBPPerc,tabelaNascimentosBPPerc)
tabelaDadosBPerc
ggplot(tabelaDadosBPPerc,aes(x=tabelaAnosBPPerc ,y=tabelaNascimentosBPPerc))+ geom_line(color="red")+
scale_y_continuous(limits = c(0, 100)) + ylab("Nascimentos Baixo Peso (%)") + xlab("Anos")



#Estatística descritiva
summary(tabelaNascimentosBP)
decompose(tabelaNascimentosBP)
windowsFonts(A = windowsFont("Times New Roman"))
hist(tabelaNascimentosBP,main="Frequência de NBP em Portugal no período entre 1989 e 2019", xlab="N° de
NBP registados",
     col="lightgrey",family="A")


#Testar a normalidade (série original NBP)
qqnorm(tabelaNascimentosBP)
decompose(tabelaNascimentosBP)
shapiro.test(tabelaNascimentosBP)



#Análise da autocorrelação
acftab <- acf(tabelaNascimentosBP) #autocorrelação
pacftab <- pacf(tabelaNascimentosBP) #autocorrelação parcial
acf(tabelaNascimentosBP,lag=5,pl=FALSE)

#Análise da estacionaridade - Série original NBP
testdfBP <- ur.df(tabelaNascimentosBP, type="drift")
summary(testdfBP)
testppBP <- ur.pp(tabelaNascimentosBP, model="constant",type="Z-tau", lags="short")
```

```
summary(testppBP)
testkpssBP <-ur.kpss(tabelaNascimentosBP,type="tau", lags="short")
summary(testkpssBP)

#FAC e FACP
acf(tabelaNascimentosBP)
pacf(tabelaNascimentosBP)

#FAC e FACP - valores por lag
acf_valuesBP = acf(tabelaNascimentosBP,plot=F)
acf_valuesBP
pacf_valuesBP = pacf(tabelaNascimentosBP,plot=F)
pacf_valuesBP

#Diferenciação - 1 diferenciação NBP
diffBP <- diff(tabelaNascimentosBP,differences =1)
tabelaNascimentosBPdiff = ts(diff(tabelaNascimentosBP,differences =1))
plot(tabelaNascimentosBPdiff)+ geom_line(color="red")

#Análise estacionaridade - 1 diferenciação (2 opçao)
testdfBP <- ur.df(diffBP, type="drift")
summary(testdfBP)
testppBP <- ur.pp(diffBP, model="constant",type="Z-tau", lags="short")
summary(testppBP)
testkpssBP <-ur.kpss(diffBP,type="tau", lags="short")
summary(testkpssBP)

#FAC e FACP (Série diferenciada)
acf(diffBP)
pacf(diffBP)

#FAC e FACP - valores por lag (Série diferenciada)
acf_valuesdiffBP = acf(diffBP,plot=F)
acf_valuesdiffBP
pacf_valuesdiffBP = pacf(diffBP,plot=F)
pacf_valuesdiffBP

#Ajustar um modelo auto ARIMA
tabelaDadosmodeldiffBP <- auto.arima(diffBP,ic="aic",trace=TRUE)


#Estimar modelos ARIMA (d=1)
arima(tabelaNascimentosBP,order = c(2, 1, 2))
summary(arima(tabelaNascimentosBP,order = c(2, 1, 2)))
BIC(arima(tabelaNascimentosBP,order = c(2, 1, 2)))

#Melhor modelo: ARIMA (2,1,2)
arima(tabelaNascimentosBP,order = c(2, 1, 2))
summary(arima(diffBP,order = c(2, 1, 2)))

#Significância dos parâmetros
confint(arima(tabelaNascimentosBP,order = c(2,1,2)))

#Análise de resíduos
tsdiag(arima(tabelaNascimentosBP,order = c(2,1,2)))
```

```
shapiro.test(residuals(arima(tabelaNascimentosBP,order = c(2,1,2))))
Box.test(residuals(arima(tabelaNascimentosBP,order = c(2,1,2))),type=c("Ljung-Box"))
checkresiduals(arima(tabelaNascimentosBP,order = c(2,1,2)))

#Forecast NBP
autoplot(forecast(arima(tabelaNascimentosBP,order = c(2,1,2)),level=c(95),h=4)) +
scale_y_continuous(limits = c(0, 10000)) + labs(y = "N° de Nascimentos de Baixo Peso", x = "Anos")
forecast(arima(tabelaNascimentosBP,order = c(2,1,2)),level=c(90),h=4)
accuracy(forecast(arima(tabelaNascimentosBP,order = c(2,1,2)),level=c(95),h=4))

#Forecast in sample
forBP <- window(tabelaNascimentosBP,start=1,end=26)
forecast(arima(forBP,order=c(2,1,2)),level=c(90),h=4)
fBP_in_sample <- forecast(arima(forBP,order=c(2,1,2)),level=c(95),h=4)
accuracy(forecast(arima(forBP,order=c(2,1,2)),level=c(95),h=4))

# Forecast BP - Alisamento Exponencial
ses(forBP,h=4,level = c(90, 95))
autoplot(ses(forBP,h=4),ylim=c(0,10000))
aeBP <- ets(forBP, "ANN")
aeBP
accuracy(ses(forBP,h=4))

# Forecast NBP - Médias Móveis
rmBP <- rollmean(tabelaNascimentosBP,k=2)
rmBP
accuracy(rollmean(tabelaNascimentosBP,k=2))
autoplot.zoo(rollmean(tabelaNascimentosBP,k=5),ylim=c(0,10000))
```

```
# NASCIMENTOS DE MUITO BAIXO PESO

tabelaAnosMBP <-c(Ano=1989:2019)
tabelaNascimentosMBP <-c(NascimentosMBP = c(611,  736,  651,  682,  778,  757,  760,
  892,  921,  974,  1183,  1123,
997,1053,996,1020,1037,1011,1070,1084,1176,1215,1138,1148,1025,1025,1106,1118,1116,985,912))

#data frame tabelaDados
```

```
tabelaDadosMBP <- data.frame(tabelaAnosMBP,tabelaNascimentosMBP)
tabelaDadosMBP

#gráfico da ST NMBP
plot(tabelaDadosMBP,type="l",xlab="Anos",ylab="Nascimentos de Muito Baixo Peso",ylim=c(0,5000))
ggplot(tabelaDadosMBP,aes(x=tabelaAnosMBP
,y=tabelaNascimentosMBP))+geom_line(color="red")+scale_y_continuous(limits=c(0, 5000))

#Gráfico em %
tabelaAnosMBPPerc <-c(Ano=1989:2019)
tabelaNascimentosMBPPerc <-c (NascimentosMBPPerc =
c(0.5,0.6,0.6,0.6,0.7,0.7,0.7,0.8,0.8,0.9,1.0,0.9,0.9,0.9,0.9,0.9,0.9,1.0,1.0,1.0,1.2,1.2,1.2,1.3,1.2,1.
2,1.3,1.3,1.3))
tabelaDadosMBPPerc <- data.frame(tabelaAnosMBPPerc,tabelaNascimentosMBPPerc)
tabelaDadosMBPPerc
ggplot(tabelaDadosMBPPerc,aes(x=tabelaAnosMBPPerc ,y=tabelaNascimentosMBPPerc))+ geom_line(color="red")+
scale_y_continuous(limits = c(0, 100)) + ylab("Nascimentos de Muito Baixo Peso (%)") + xlab("Anos")

#Estatística descritiva
summary(tabelaNascimentosMBP)windowsFonts(A = windowsFont("Times New Roman"))
hist(tabelaNascimentosMBP,main="Frequência de NMBP em Portugal no período entre 1989 e 2019", xlab="N°
de NMBP registados",
     col="lightgrey",family="A")

#Testar a normalidade (série original)
qqnorm(tabelaNascimentosMBP)
decompose(tabelaNascimentosMBP)
shapiro.test(tabelaNascimentosMBP)

#Análise da autocorrelação
acftab <- acf(tabelaNascimentosMBP) #autocorrelação
pacftab <- pacf(tabelaNascimentosMBP) #autocorrelação parcial
acf(tabelaNascimentosMBP,lag=5,pl=FALSE)

#Análise da estacionaridade (Série original MBP)
testdfMBP <- ur.df(tabelaNascimentosMBP, type="drift")
summary(testdfMBP)
testppMBP <- ur.pp(tabelaNascimentosMBP, model="constant",type="Z-tau", lags="short")
summary(testppMBP)
testkpssMBP <-ur.kpss(tabelaNascimentosMBP,type="tau", lags="short")
summary(testkpssMBP)



#FAC e FACP
acf(tabelaNascimentosMBP)
pacf(tabelaNascimentosMBP)

#FAC e FACP - valores por lag
acf_valuesMBP = acf(tabelaNascimentosMBP,plot=F)
acf_valuesMBP
pacf_valuesMBP = pacf(tabelaNascimentosMBP,plot=F)
pacf_valuesMBP

#Diferenciaçao - 1 diferenciação
```

```
diffMBP <- diff(tabelaNascimentosMBP,differences =1)
tabelaNascimentosMBPdiff = ts(diff(tabelaNascimentosMBP,differences =1))
qqnorm(tabelaNascimentosMBPdiff)
plot(tabelaNascimentosMBPdiff)

#FAC e FACP (Série diferenciada)
acf(diffMBP)
pacf(diffMBP)

#FAC e FACP - valores por lag (Série diferenciada)
acf_valuesdiffMBP = acf(diffBP,plot=F)
acf_valuesdiffMBP
pacf_valuesdiffMBP = pacf(diffBP,plot=F)
pacf_valuesdiffMBP

#Análise estacionaridade - 1 diferenciação
testdfMBP <- ur.df(diffMBP, type="drift")
summary(testdfMBP)
testppMBP <- ur.pp(diffMBP, model="constant",type="Z-tau", lags="short")
summary(testppMBP)
testkpssBP <-ur.kpss(diffBP,type="tau", lags="short")
summary(testkpssBP)

#Ajustar um modelo auto ARIMA
tabelaDadosmodeldiffMBP <- auto.arima(diffMBP,ic="aic",trace=TRUE)

#Estimar modelos ARIMA
arima(tabelaNascimentosMBP,order = c(2, 1, 2))
summary(arima(tabelaNascimentosMBP,order = c(2, 1, 2)))
BIC(arima(tabelaNascimentosMBP,order = c(2, 1, 2)))

#Melhor modelo: ARIMA (2,1,2)
arima(tabelaNascimentosMBP,order = c(2, 1, 2))
summary(arima(tabelaNascimentosMBP,order = c(2, 1, 2)))

#Juntar para a análise de resíduos
checkresiduals(arima(tabelaNascimentosMBP,order = c(2,1,2)))
tsdiag(arima(tabelaNascimentosMBP,order = c(2,1,2)))


#Significância dos parâmetros
confint(arima(tabelaNascimentosMBP,order = c(2,1,2)))


#Forecast NMBP
autoplot(forecast(arima(tabelaNascimentosMBP,order = c(2,1,2)),level=c(95),h=4)) +
scale_y_continuous(limits = c(0, 5000)) + labs(y = "N° de Nascimentos de Muito Baixo Peso", x = "Anos")
forecast(arima(tabelaNascimentosMBP,order = c(2,1,2)),level=c(90),h=4)
accuracy(forecast(arima(tabelaNascimentosMBP,order = c(2,1,2)),level=c(95),h=4))

#Forecast in sample
forMBP <- window(tabelaNascimentosMBP,start=1,end=26)
forecast(arima(forMBP,order=c(2,1,2)),level=c(90),h=4)
fMBP_in_sample <- forecast(arima(forMBP,order=c(2,1,2)),level=c(95),h=4)
accuracy(forecast(arima(forMBP,order=c(2,1,2)),level=c(95),h=4))
```

```
# Forecast MBP - Alisamento Exponencial
ses(forMBP,h=4,level = c(90, 95))
autoplot(ses(forMBP,h=4),ylim=c(0,5000))
aeMBP <- ets(forMBP, "ANN")
aeMBP
accuracy(ses(forMBP,h=4))

# Forecast MBP - Médias Móveis
rmMBP <- rollmean(tabelaNascimentosMBP,k=2)
rmMBP
accuracy(rollmean(tabelaNascimentosMBP,k=5))
autoplot.zoo(rollmean(tabelaNascimentosMBP,k=5),ylim=c(0,5000))
```